Technische Universität Berlin

Master Thesis

Leveraging Context Information in Spatio-Temporal Big Data Analytics

a Study in the Mobility Domain

Ricardo Ernesto Martinez Ramirez
EIT Innovation Matriculation #: 0376583

Supervisor: Prof. Dr. Volker Markl
Advisor: Dr. Holmer Hemsen

27/09/2018

Erklärung (Declaration of Academic Honesty)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

I hereby declare to have written this thesis on my own and without forbidden help of others, using only the listed resources.

Datum	Ricardo Ernesto Martinez Ramirez

Contents

List of Figures

List of Listings

List of Tables

List of Acronyms

Abstract

The main objective of this thesis is to evaluate the use of a Spatio-Temporal approach in order to provide contextual data in the mobility domain.

The Daystream project[1] aims to explore innovative techniques and tools to improve mobility experience, traffic safety, and quality of service. Daystream is funded by the Federal Ministry of Transport and Digital Infrastructure, *Bundesministerium für Verkehr und digitale Infrastruktur* (*BMVI*)[2]. This Master thesis is framed into the Daystream project, within a use case to improve mobility experience.

Spatio-Temporal is, in addition to other features of interest, the combination of spatial and temporal dimensions, which denotes the location and time.

Traffic engineering is a part of the Mobility field defining traffic parameters used in this thesis, such as Speed, Time Mean Speed, Space Mean Speed, Density, Time Headway, Distance Headway, among others. These traffic parameters are used to define Level of Service and the *Traffic Congestion* condition of a road.

The methodology goes through a process divided into two steps:

- *Traffic Congestion Areas Detection*: get road sensor data, calculate traffic parameters from that data, filter the concrete places of *Traffic Congestions* and detect areas where the density of sensors is high.

- *Levering Context Information*: use of *External Events* as the context information for *Traffic Congestion Areas*, by calculating relationships between *Traffic Congestion Areas* and *External Events*.

This thesis shows the use of a Spatio-Temporal approach in the mobility field to include *External Events* as the context of a *Traffic Congestion Area*. As the main contributions this thesis uses traffic parameters and formal definitions of *Traffic Congestion*, the main

[1] https://www.daystream-projekt.de/
[2] https://www.bmvi.de/EN/Home/home.html

focus is an analysis of an area rather than a single point or road segment, the changes over time are considered, and finally the use of data not related to roads.

The contextual data can be used to prediction traffic congestions impact or provide further information to take decisions on policies and strategies, and eventually to avoid or reduce such impact.

Zusammenfassung

Das Ziel dieser Arbeit ist es, den Nutzen von spatio-temporalen Ansätzen zu untersuchen, um kontextbezogene Daten in der Mobilitätsdomäne zur Verfügung zu stellen.

Das Daystream Projekt hat zum Ziel innovative Techniken und Tools zu untersuchen, um die Verkehrsqualität in Bezug auf Sicherheit und Zuverlässigkeit zu verbessern. Dieses Projekt wird vom Bundesministerium für Verkehr und digitale Infrasturktur (BMVI) gefördert. Die vorliegende Arbeit ist eingebettet in den Bereich der Mobilitätsverbesserung dieses Projekt.

Spatio-Temporal stellt, neben anderen nennenswerten Eigenschaften, die Kombination von Ort und Zeit dar.

Verkehrsingenieurwesen ist ein Fachbereich der Verkehrswissenschaften, welches unter anderem die folgenden Verkehrsparameter definiert: Geschwindigkeit, Durchschnittsgeschwindigkeit, Verkehrsdichte, zeitlicher und räumlicher Abstand. Diese Verkehrsparameter werden benutzt, um Qualitätsstufen und *Verkehrsstauung* der Straße zu bestimmen.

Die Methodik dieser Arbeit ist in zwei Schritte geteilt:

- Erkennung von staugefährdeten Bereichen: Sensordaten abrufen, Aufbereiten dieser Verkehrsdaten, Verkehrsparameter berechnen, um anschließend Verkehrsstauungen an konkreten Orten ausfindig machen und Bereiche mit hoher Verkehrsdichte erkennen zu können.

- Hinzufügen von kontextbezogenen Informationen: externe Ereignisse als kontextbezogene Informationen für Bereiche mit Verkehrsstauungen nutzen.

Diese Arbeit zeigt, wie ein *ST* Ansatz im Bereich Verkehrswissenschaften genutzt werden kann, um Bereiche mit Verkehrsstauungen mit zusätzlichen Informationen - externen Ereignissen - zu verbinden. Die kontextbezogenen Daten können dabei genutzt werden, um Vorhersagen über den Einfluss von Verkehrsstauung zu treffen. Des Weiteren können zusätzliche Informationen bereitgestellt werden, um Entscheidungen zur Vermeidung von Verkehrsstauungen treffen zu können.

1. Introduction

The main objective of this thesis is to evaluate the use of ST algorithms in order to provide contextual data in the mobility domain. It is expected to obtain enhanced traffic congestion information. In order to achieve such objective, data from road sensors and other data sources are described as ST events. Enhanced and aggregated ST data is being used in algorithms to detect where and when traffic congestions are happening. A further analysis is to measure the relationship with context data, such as events that might influence such traffic conditions.

The contextual data can be used to trigger the prediction of traffic congestion impact, or provide further information to take decisions on policies and strategies to avoid or reduce such impact. In the field of Geographical Information Science it is mentioned that "Everything is related to everything else, but near things are more related than distant things"[47], known as Tobler's first law of geography. This law can be extended to the Spatio-Temporal domain. In general, events happening closer in time and space are more likely to influence or be related to each other.

Most of these works focus on the actual algorithm to predict, i.e. either how the traffic will behave or what the impact in case of traffic congestion is, considering only traffic-related features, for example speed and quantity of vehicles in a road section, Pan et. al [41, 40, 51] extended this approach, by considering road accidents as the trigger of the impact-prediction algorithms. However, seasonality, accidents, and road characteristics are not the only influencing features in traffic behavior. The addition of contextual data might give new ways to model, describe, and eventually predict traffic conditions, as proposed in [50].

Therefore, the following questions will be discussed in the thesis:

- Which areas may be interesting spots of traffic congestions for further analysis?
- How close in time and space is "close enough" to determine if an event is likely to trigger traffic congestions?
- Which events are likely to trigger traffic congestions in a certain area?
- What are the factors for such process in a Big Data environment, considering an increase in volume, velocity, and variety?

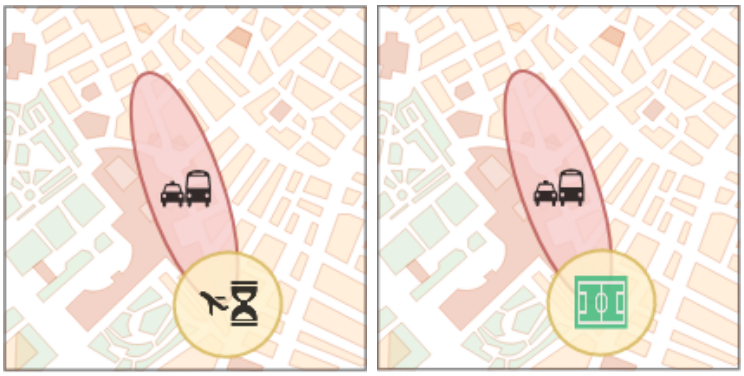
1.1. Motivation

Mobility is formally defined as "*The ability to move or be moved freely and easily*" [3] and affects humans all over the world. However, mobility can be affected by several factors, some of them directly related, such as accidents or reduced capacity on roads, and some others are not as obvious, because they happen outside the roads. The focus in this thesis is to link the context around places where mobility is affected. The term context is defined as "*The situation within which something exists or happens, and that can help explain it*" [4], in other words, everything that has an influence on some situation in particular. According to [9, 18, 10] transportation and mobility impact every single person, directly as a user, or indirectly as a consumer of products and services. Therefore, having a better quality on transportation and mobility should have a positive impact on the environment, society, and economics. In the recent years the use of information technologies is a key component of world-wide initiatives to increase security, efficiency, and to reduce the negative impact of mobility [2, 3, 4]. As part of those initiatives, the Daystream project[5] aims to explore innovative techniques and tools to improve mobility experience, traffic safety, and quality of service. Daystream is funded by the Federal Ministry of Transport and Digital Infrastructure, *Bundesministerium für Verkehr und digitale Infrastruktur (BMVI)*[6]. This thesis is framed into the Daystream project, within a use case to improve mobility experience.

In recent years digital tools are available to the mobility users, tools that help them to reach their destination in a shorter time, by suggesting the shortest path, by avoiding current congested areas, and, finally community-based traffic platforms inform about road accidents or road maintenance works. Google Maps, Here Maps, Waze, are some of the current well known platforms with such functionalities. In these digital tools the information is road related and it does not consider external factors. Therefore, a hypothesis is *can factors that explicitly are outside the roads, the context, also influence or trigger traffic congestions?* This thesis will explore the *ST* approach to provide a context to traffic congestions, which can be extended as a future work, for example, combined with other systems detecting events from social media, as in the case of Daystream project or as a platform to prevent users from traffic congestions caused by external factors.

One possible external factor is an unexpected event, for example a massive flight delays, where many people may travel in an unusual rate, it can be considered as part of the context of traffic congestions in the nearby area of the airport, as shown in Figure 1a.

[3] https://dictionary.cambridge.org/dictionary/english/mobility
[4] https://dictionary.cambridge.org/dictionary/english/context
[5] https://www.daystream-projekt.de/
[6] https://www.bmvi.de/EN/Home/home.html



(a) (b)

Figure 1: Contextual information of an area with traffic congestions.

Another example possible external factor is a planned event, for example a football match, where many people is expected to travel prior to the match, this is shown in figure 1b. In both cases, the red area represents a traffic congestion, and the yellow circle with a plane and an hourglass represent the airport location at the moment of the mass delays, while the yellow circle with a football field represents the location of the stadium prior the match.

One of the problems affecting users are traffic congestions [6], which has been the main focus of several related works in the mobility domain, included as a component in the description, modeling, and prediction of traffic behavior. Delays are a common problem caused by traffic congestinos, leading to some immediate effects on users, such as stress, and therefore, decreasing the quality of life. For consumers, as a non-direct users of the road network, traffic congestions have an impact by increasing the waiting time for products and services, with economical effects, energy and resource waste. Considering that traffic congestions can be triggered by different factors, the impact can be minimized by reducing the effect of such factors, for instance when a road maintenance work or a mass event is planned, alternative routes or means should be provided, in case of unexpected factors, *e.g.* a car accident or weather conditions. An accurate evaluation of the possible impact can help to take actions or change policies, which, at the end, should lead to a reduction of the general impact.

From the computer science perspective, recent works focused on traffic prediction, using techniques such as Artificial Neural Networks [7, 8, 9, 10], time series analysis [12, 13], or data mining [7, 8] on collected datasets from authorities or transport companies [11, 13], surveys [6], and real-time sensor data from roads [7, 8]. The results include conclusions to decide better mobility policies [11], determine the user experience and opinion to reveal possible hidden causes of traffic congestions [6], or to building up a component of a broader

Intelligent Transportation System (ITS) [7, 8]. Most of these works focus on predictions of car speed or quantity over a short road segment. Pan et. al [7, 8] extended this approach, by consider road accidents as the trigger of the impact prediction algorithms. However, seasonality, accidents, and road characteristics are not the only influencing features in traffic behavior. The addition of contextual data might give a different perspective to model, describe, and eventually predict traffic conditions, as proposed in [50], specifically with the so called Data Type 3.

An approach that holds a more realistic perspective is the *ST* approach, which considers 2 dimensions, (i) the spatial dimension, shorten as *spatio*, represented as geometric abstractions, e.g. the coordinates in the geographical projection defined in the *World Geodetic System 1984 (WGS84)*, a common system to identify places on earth by a given longitude and latitude; and (ii) the temporal dimension, usually represented in terms of instants or periods.

1.2. Objectives and Scope

The goal of this thesis is to evaluate the use of a *ST* approach, including algorithms and tools, to enrich mobility information, in order to describe road traffic events and their external context, for events happening nearby. This will provide a broader perspective of the possible hidden triggers of traffic congestions, in contrast to the only-road related features, such as car accidents.

The scope of this thesis is to use road sensors data from the mobility field, transform and aggregate it, in order to use *ST* algorithms and enrich the context of a traffic congestion with external factors *i.e.* events happening close in time and space. Because of the nature of the sensor data, the thesis consists of two steps, the first one preprocesses sensor data into traffic parameters and detects dense areas with traffic congestion reported by sensors, the second step detects matching events close in time and space, which adds information to the context.

1.3. Thesis Outline

The thesis is structured as follows. In Section 2 the basic terminology in the *ST* approach is presented, including definitions in the sub fields of time and space. Further in this section, a non-extensive, but fundamental, terminology related to the mobility field is introduced, including concepts used by authorities and government agencies in charge of mobility. In Section 3 a review on scientific works related to the use of computer

techniques applied on mobility is presented. A set of papers and studies is discussed in chronological order, including contributions, gaps or areas of improvement, and results. In Section 3 it is presented a summary of gaps from previous works, some of which are covered in this thesis. Section 4 describes the methodology followed in this thesis, a big picture of the process, which is detailed afterwards. Two steps are defined, the discovery of areas of interest and the levering of contextual information. Section 5 explains the experiments and the evaluation process of the methodology. Finally Section 6 contains the conclusions of the experiments and evaluations done of the propossed methodology.

2. Fundamentals

ST analysis and mobility are two vast topics. The aim of this section is to give a brief introduction to the fundamentals, such as spatial and temporal concepts used in this thesis and related work focused on the use of *ST* analysis and computer techniques for data analysis. In subsection 2.2.1, mobility terms and concepts of specific parameters in the traffic engineering field, and the definition of a traffic congestion are introduced.

2.1. Fundamentals of Spatio-Temporal Analysis

As mentioned in Section 1, Spatio-Temporal is, in addition to other features of interest, the combination of spatial and temporal dimensions. In this subsection both dimensions are defined as a way to represent observations in different fields. Afterwards, it is defined the corresponding mapping to computer instances to represent real world observations as objects. In a more general way, it is also introduced some operations over *ST* objects, used in algorithms for *ST* analysis.

2.1.1. Spatial Dimension

The spatial dimension is a representation of geometric-related components of an entity, for instance data within geography, anatomy, and neuroscience fields can benefit from the spatial extension in data. A close discipline is the image processing, however the main difference with spatial systems is that the later ones focus on representing entities in space with clearly defined location and extend [23]. Some authors [5, 1] include image processing and analysis as part of the spatial dimension, which involves additional referential data, for example the use of satellite images or fMRI video sequences.

Gülting and Schneider introduced three fundamental spatial abstractions, which are defined for single objects, *point*, *line,* and *region* [23, 22]. These abstractions are called by Gülting *realms*, an analogy to enumerations in programming languages, in such way that it is a finite structure and all of them can be defined in terms of points and line segments. A point represents the geometric aspect of an object, only location in space, as the extent is irrelevant. A line is a link between two points, a line gives the notion of moving or conextions in space. A region is the abstraction of an entity with extent.

Spatially related collections of objects are partitions and networks. Partitions are set of adjacent regions. While a network can be seen as a graph, consisting of a set of points and a set of lines, describing nodes and edges respectively. In practice, formats

define other collections, such as a set of points, set of lines, or set of regions, also called polygons, GeoJSON is an example of such formats [19] or PostGIS, as a spatial extension of PostgreSQL database system [21].

Operations over objects with spatial attributes proposed in [23, 22] and defined in [21] and other tools, for example the geo-analytics suit ArcGIS[7], include but are not limited to:

- **intersection** $line \times line \rightarrow points$ returns a set of points that both lines share.
- **intersection** $line \times region \rightarrow line$ returns a line that represents the shared portion of the line and the region.
- **intersection** $region \times region \rightarrow region$ returns a region that represents the area of the first region that is also part of the second region.
- **minus** $region \times region \rightarrow region$ returns a region that represents the area of the first region that does not intersect with the second region.
- **contour** $region \rightarrow line$ returns a line representing the exterior of a region.
- **length** $line \rightarrow real$ returns a number representing the length of the line in a determined units system, such as meters.
- **area** $region \rightarrow real$ returns a number representing the area of the region in a determined units system, such as squared meters.
- **distance** $realm \rightarrow realm \rightarrow real$ returns a number representing the cartesian minimum distance between the realms in a determined unit system, such as meters.
- **inside** $realm \times region \rightarrow bool$ returns a true or false value whether the realm is completely inside the region.
- **adjacent** $region \times region \rightarrow bool$ returns a true or false value whether the regions have at least one point in common, but their interiors do not intersect.

This operations are used in algorithms in time and space, in the defined format of ST-Data Type within the systemms discussed later in this chapter, however an example is the *K Nearest Neighbors (KNN)* algorithm which returns the k-closest objects, generally in euclidean distance, used for classification or regression.

[7] https://www.arcgis.com/index.html

2.1.2. Temporal Dimension

The time dimension, or *temporal* dimension, also extends the possibilities to analize and explore data. Generally time is perceived as a one dimension space [23], and can be considered as bounded or unbounded [23, 2], a bounded model assumes an origin and end of time; discrete or continuous, discrete time assumes each time reference as an atomic time interval, while continuous time assumes that any time reference corresponds to a point in time; absolute or relative, also known as anchored or unachored, respectively, where anchored corresponds only to a one point in time in particular, while unachored or relative can be representative of any time in point, "three wheeks" for instance. Time can be expressed in different ways according to [23, 20, 33]:

- *instant* also known as *chronon*, is a point on the time line in a continuous model, it can be represented as a combination of *date* and *time* or as a *timestamp*.
- *period* is the time between two instants, an anchored interval on the time line, demarcated by a start time and an end time.
- *interval* is a directed, unachored duration of time. It acts as a bounding element, for example *3 weeks*.
- *date* is a particular day from a year, for instance January 1st 2018.
- *time* is a particular second within a range of 24 hours, for instance 14:05:03.
- *timestamp* is a particular fraction of a second, usually in microseconds, of a particular day. A common implementation is the *UNIX Epoch time*, defined as a *Coordinated Universal Time (UTC)*, a value that approximates the number of seconds that have elapsed since the Epoch, fixed on 1 January 1970.

In the time model explained in *Temporal Databases: Research and Practice*[20], implemented by many temporal *Database Management Systems (DBMSs)*, two concepts are fundamental for the *ST* approach, *valid time* and *transaction time*. The *valid time* of a fact is the time when the fact is true in reality [20], it is called *event time* in the data flow approach [2]. The *transaction time*, or *processing time* in the data flow approach, refers to the time when a change is recorded in the system, or the time interval during which a particular state of the database exists [23].

In contrast with spatial attributes, operations over objects with temporal attributes are more implementation dependent, according to [38], for more than 25 years the SQL standard could not allow an easy definition, manipulation and querying of temporal databases, and this situation extends to many programming languages and libraries. As a reference

of potential operations over objects with temporal attributes, a list of operations mentioned in the so called Alle's interval algebra [3], analog to ROSE algebra in the time dimension, and the standard SQL:2011 [33], the latest based on and compared to the first, is as follows and shown in Figure 2:

- **equals** $XequalsY \rightarrow bool$ returns a true value if X and Y periods are equal, therefore start and end time in X and Y are the same. False Otherwise.
- **contains** $XcontainsY \rightarrow bool$ returns a true value if, intituively, every time point in period Y is also in X. False Otherwise.
- **overlaps** $XoverlapsY \rightarrow bool$ returns a true value if, intituively, at least one time point in period Y is also in X. False Otherwise.
- **precedes** $XprecedesY \rightarrow bool$ returns a true value if, intituively, the last time point in period X is lower than the first time point in X. False Otherwise.
- **succeeds** $XsucceedsY \rightarrow bool$ returns a true value if, intituively, the first time point in period X is higher than the last time point in X. False Otherwise.
- **immediately precedes/succeeds** $Ximmediatelyprecedes/succeedsY \rightarrow bool$ returns a true or false value following *precedes* and *succeeds* rules for the specific case when last and first time points of the corresponding periods are consecutive.

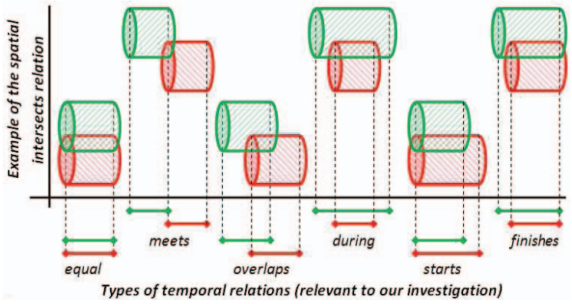


Figure 2: Time operations.

Additional to the period-related operations, some implementations allow to use time-related functions and operators including, but not limited to:

- **add interval** Adds an interval to an instant. For example, *'2018-01-01T00:00:00'*[8] **ADD** *'1 hour'* $\rightarrow$ *'2018-01-01T01:00:00'*.

[8]Format ISO 8601: YYYY-MM-DD'T'hh:mm:ss, where YYYY refers to the year in four digits; MM refers to the month in two digits; DD refers to day in two digits, 00 through 31; hh refers to hours in

- **substract interval** Substracts an interval to a instant. For instance, fot the instant *'2018-01-01T00:00:00'* **SUBSTRACT** *'1 hour'* → *'2017-12-31T23:00:00'*.
- **current date** Returns the current time in the system, also known as SISTEM_TIME.
- **truncate** Returns the part of the instant related to a time or date field, such as hour or day. For instance, fot the instant *'2017-12-31T23:00:00'* **TRUNCATE** *'month'* → *'2017-12-01T00:00:00'*.
- **extract** Returns a numeric value of the instant related to a time or date field, such as hour or day. For instance, fot the instant *'2017-12-31T23:00:00'* **EXTRACT** *'month'* → *'12'*.

As mentioned before, this operations and functions are not standarized in all the *DBMS*, programming languages, or libraries, therefore some operations can be a chained combination of functions and operation of the available within the system.

2.1.3. Spatio-Temporal Approach

In the previous sections, space and time dimensions are already individually defined. But the cornerstone of *ST* systems are the *Spatio-Temporal Data Types (STDTs)*, objects with both dimensions, with the development of new operations and algorithms, enhancing the analysis over data. In that respect, it will also increase the amount of data to be processed, and the use of more complex algorithms, bringing new challenges, as [1] refers, modern geospatial big data (including the temporal dimension) are spatial data sets exceeding capacity of current computing systems. One of the most important issues to solve within *ST* analysis is how we exploit the large amount of spatial data generated nowadays.

In *ST* systems, such as *Geographic Information System (GIS)*, the basic abstractions are Moving Objects, a representation of an entity with time-dependent position; entities with extent are characterized as Moving Regions. Traditionally, geospatial data can be categorized into three forms: raster data, vector data, and graph data. The raster data include geoimages. The vector data consists of points, lines, and polygons. The graph data mainly appears in the form of road networks [1].

More specifically, data in the Spatio-Temporal approach is described in four common categories of *STDT* [5]:

24h format, 00 through 23; mm refers to minutes, 00 through 59; ss refers to seconds, 00 through 59. https://www.w3.org/TR/NOTE-datetime.

1. **Event data**, which comprises of discrete events occurring at point locations and times, *e.g.* incidences of crime events in the city, an alert triggered by a fixed sensor which measurement crosses a threshold. As a representation of an instantaneous fact, an object including the ST attributes or features (l_i, t_i), it denotes the location and time, additionally to the non-ST features or variables. For certain real-world applications, a point and an instant are not enough to describe an event, therefore in [5] is suggested that other geometries and time representations can contribute to a better description, for instance a forest fire event can be represented as a polygon of the affected area and it can be associated to a time period, with a start and end instant or a start instant and an interval.

2. **Trajectory data**, where trajectories of moving objects are represented, *e.g.* the patrol route of a police surveillance car.

3. **Point reference data**, where a continuous field is being measured at moving reference sites *e.g.* measurements of surface temperature collected using weather balloons, as balloons are moving while measuring weather conditions, the representation of the field, in this case the geographic portion of land is fixed or continuous.

4. **Raster data**, where observations of an ST field is being collected at fixed cells in an ST grid *e.g.*, fMRI scans of brain activity. Similar to the point reference data, with the difference that the measures are done within a fixed grid.

According to [5], a *Spatio-Temporal Data Instance (STDI)* is the basic unit to be processed by a computer algorithm, containing the types of data that will be used by the algorithm. an $STDI$ can represent different $STDTs$. The choice of the right approach for constructing $STDI$ instances from a $STDT$ type depends on the nature of the question being investigated, and the available ST methods that can be used. Some example represenations of ST data are presented in [23] are as follow:

- **Events in space and time**, *(point, instant)*. A car crash on the road can be described as an event which happened on a specific point at a specific instant. It is considered that the duration is not relevant.

- **Locations valid for a certain period of time**, *(point, period)*. A maintenance work on the road can be described as a location which is affected or with a different status along the period of such maintenance work. Another example can be a car crash impact on the road, if the extent of the impact is not relevant, but duration is.

- **Set of location events**, *sequence of (point, instant).* A set of sensors along the roads detecting a traffic congestion at a specific instant, at rush hour for instance. Another example can be the set of locations where there were

- **Stepwise constant locations**, *sequence of (point, period).* Locations of shops, stadiums, airports, which are valid over years, until they are closed or moved to another location. Depending on the time frame of analysis, period can vary from fraction of seconds up to years or decades. Another example can be the status of traffic conditions over time reported by sensors along the roads.

- **Region events in space and time**, *(region, instant).* Weather conditions in a region at a specific instant.

- **Regions valid for some period of time**, *(region, period).* The car crash impact along a road or the area of within the traffic congestions for certain period, in this case both extents are important.

- **Set of region events**, *sequence of (region, instant).* All regions with certain weather conditions at a specific instant, for instance all regions with heavy rain during the rush hour at evening.

- **Stepwise constant regions**, *sequence of (region, period).* A country shape that changes over time, the evolution traffic levels by neighborhood or area in a grid.

One of the main aspects to be considered in data analysis is the similarity, in other words, how close data is along a dimension, a feature, or an attribute. Similarity, or dissimilarity, is used in algorithms such as clustering, classification, pattern discovery, among others. Similarity can be defined within ST using at least space and time dimensions, as a basic predicate defining spatial analysis states, the so called Tobler's first law of geography [47]: "everything is related to everything else, but near things are more related than distant things." This law can be extended to the ST domain, in general, events happening closer in space and time have a higher probability to influence or be related to each other. A similarity measure differs from an event data, trajectory data, or any other representation of a real world observation and the expected analysis results. [5] mentions, among others, *Point Similarity* and *Trajectory Similarity* as ways to define similarity form the ST point of view.

The term *Point Similarity* considers two points close, or similar, if they lie within the same ST neighborhood, using a fixed distance threshold in space and time, *e.g.* within a 1 km radius and 1 hour. Other concepts can be applied to define a space and time neighborhood, such as tumbling or sliding windows, mentioned in [2]. This similarity

approach is used in *KNN* and therefore in several algorithms, such as *Spatio-Temporal DBSCAN (ST-DBSCAN)* [6].

In the case of *Trajectory Similarity*, similarity is often measured in terms of the co-location frequency, which is the number of times two moving entities appear spatially and temporally close to each other. This technique was used in [29], discused in Section 3, where different vehicle navigation system probes where aggregated to determine the traffic conditions at certain roads.

With *STDT* already defined as the way of representing real world observations from the *ST* approach, the next step is to mention some of the problems and methods, including some of the algorithms developed. From the perspective of data analysis and data mining, [5] classifies previous literature into six categories: clustering, predictive learning, change detection, frequent pattern mining, anomaly detection, and relationship mining.

- **Clustering**, refers to the process of grouping of instances in a data set that share similar feature values, those might include spatial and temporal similarities. In *ST*, clustering techniques can be applied to points with two different objectives, first to identify the so called hot-spots, highly dense *ST* points, for instance areas where there are several traffic congestions, and can be subject of a deeper study. The second objective in *ST* point clustering is to detect clusters with similar non-*ST* features, for instance traffic accidents. Examples of the use of algorithms for trajectory clustering are [48, 29]. In a similar way, clustering raster data has been done over climate data, climate models were obtained by the use of clustering. According to [5] clustering raster data is non-trivial and often requires domain expertise.

- **Predictive learning**, the objective of predi**Event data**, which comprises of discrete events occurring at point locations and times, *e.g.* incidences of crime events in the city, an alert triggered by a fixed sensor which measurement crosses a threshold. As a representation of an instantaneous fact, an object including the *ST* attributes or features (l_i, t_i), it denotes the location and time, additionally to the non-ST features or variables.ctive learning is to map from input features to the output variables using a representative training set. For instance, predictive learning techniques have been used to predict the future location of a moving object or group of objects, given the history of visited locations. Techniques in predictive learning have been used to predict the response in a certain location and time, using observations at other locations, for instance the impact of traffic congestions in similar areas, as mentioned in [51].

- **Change detection**, the main objective is to identify the time point when a system significantly changed its behavior, compared to the past. Change detection has been

studied in time series data, discovering time intervals with homogeneous properties, such as descriptive statistic measurements, or change of state.

- **Frequent pattern mining**, is the process of discovering patterns that occur frequently over multiple instances, one of the referent examples is the items in a market-basket transactions, however in ST applications the problems can be formulated within the spatial and temporal components. Two approaches are the co-occurrence in ST points and the sequential pattern in ST points. The first approach measures statistical patterns of attraction or repulsion among pairs of ST points, in other words, how different types of ST events are related in space and time. The second approach considers a sequence of ST events, with a triggering type of event, for example, a car crash on a highway can trigger traffic congestions and some other impacts.

- **Anomaly detection**, also known as outlier detection, refers to techniques to find instances which are remarkably different from the majority, mainly focused to discover interesting but rare phenomena. Anomaly detection can be viewed from two different approaches, one is the time series or as a spatial outliers, usually defined as points with non-ST features different from those in the close neighborhood. One technique used to detect anomalies is clustering, but instead of focusing on the elements forming groups, the focus is on those which did not fit in any cluster.

With this overview of the ST approach, through spatial and temporal definitions and operations, and finally ST data types, operations, and algorithms, in the Subsection 2.2 terms and concepts related to mobility are defined, in order to give the context of the use case in this thesis.

2.2. Fundamentals of Mobility

Through this section, going from general to particular, concepts of mobility, traffic engineering, and traffic congestions are introduced, giving a better understanding to the use case, the previous work, and the data process, explained in Section 3.

Mobility, by the simplest definition refered in Section 1, involves the ability to move or be moved, it does not specify the object which changes position, or the mean to do so, however in this thesis mobility refers to the specific movement of human beings and goods by vehicles, from the point of view of an interaction with the context and the impact that it have. Impact can be measured in different ways, in the transport engineering field there are several metrics related to the moving entities, such as travel time or delays, another related to the distribution and availability of the road networks, in this thesis

the metric will be the traffic congestions, which derives in increased travel times, public transport delays, and therefore in a lower quality of life. This subsection covers concepts used in transportation, followed by traffic flow parameters definition, including formulas to calculate direct a derived parameters, and, finally, describes the conditions and thresholds for levels of service, as well as traffic congestions, the main focus of the use case in this thesis.

2.2.1. Transportation and Traffic Engineering Fundamentals

Transportation is a wide area of studies, it includes transportation planning, transportation network design, roads design, traffic engineering, among other topics with relevance to the human life and the development of the societies. Transportation systems are as complex as the diversity of elements and the interactions between them, the approaches are also diverse, some of the mentioned in [37] are:

- **Modality**, depending on the modes of transport; air, land, water, for both passengers and freight. It can also involve different means of transportation, *e.g.* car, bicycle, train, bus, plane.
- **Sector**, depending on the viewpoint, public or private.
- **Problem, or focal issue**, *i.e.* national or international policies, planning regional system, regulations.
- **Planning range**, depending the scope of the study, short, medium, or long term.
- **Passenger transport**, focused specifically on the human mobility and the means involved.
- **Freight transport**, focused on routing goods. Highly correlated to logistics.
- **Demand**, focused on the demand patterns and how to solve related problems.
- **Technology**, focused on the technology used in transportation systems, *i.e.* fuel-efficient buses, high-speed trains, self-driving cars.
- **Operational policy**, focused on the policies affecting transportation, *i.e.* incentive for car-pooling, bus fares.

- **Values of the public**, as mentioned in [37], the transportation systems should take care of the target groups, *i.e.* pedestrians, drivers, people in a city or specific area.

[37] indicates two main principles in a transportation system analysis, (i) the total transportation system must be viewed as a single multi-modal system; and (ii) considerations of transportation systems cannot be separated from considerations of social, economic, and political system of the region. The second principle supports the idea behind the thesis, as mentioned in Section 1.

Traffic flow theory has been done formally for more than 50 years [27], a clear understanding of how traffic flow operates is essential to model, measure, and, eventually, predict and prevent traffic congestions. Traffic flow can be described with different models, categorized depending on characteristics such as level of detail, operationalization, scale of application, among other more specialized. A basic categorization is done from the observations level of detail, divided into two main categories on regard of the data collection strategy, *microscopic* and *macroscopic*, however some authors consider a third one [27], *mesoscopic*:

1. **Microscopic**, describes the time and space behavior of each entity, *e.g.* individual vehicles and drivers. An example of this type of data collection is the GPS probes of each driver and car, as well as detailed interactions, such as lane-change for individuals. This approach is taken as the way of data collection in [29].

2. **Mesoscopic**, describes the behavior of individuals but rather using probabilistic measurements, *i.e.* aggregated values of GPS trajectories. Traffic is represented in small groups of traffic entities, which may include for instance, lane-change rate. This approach is also used in [29] to aggregate and analise data in the report.

3. **Macroscopic**, describes the traffic in higher level of aggregation, it does not consider any type of individual interaction. The main characteristics in a macroscopic approach are flow-rate, density, and velocity. Due to the available data, this thesis will take this approach to define measurements.

Despite the category, fundamental parameters, relations, and derived parameters in traffic flow are common definitions in the transportation engineering field, many of them will be used in this thesis as part of the features, briefly will be defined as [37] summarized:

Speed *Speed or velocity is the rate of motion in distance per unit of time. It is given by the mathematical expression:.*

$$v = \frac{d}{t}$$

Spot Speed *Is the instantaneous speed of a vehicle at a specified location. It can be measured using different techniques, including radar speedometer or pressure contact tubes, among others.*

Time Mean Speed *Is the average speed of all vehicles passing a point on a highway over some specified period. It is an aggregation from the spot speed. It can be calculated as follows:*

$$v_t = \frac{\sum_{i=1}^{n} q_i v_i}{\sum_{i=1}^{n} q_i}$$

Space Mean Speed *Is the average speed of all the vehicles occupying a given section of a highway over some specified time. It can be calculated as follows:*

$$v_s = \frac{\sum_{i=1}^{n} q_i}{\sum_{i=1}^{n} \frac{v_i}{q_i}}$$

Flow *Flow or volume is the amount of vehicles passing a point on a road or a given lane or direction during a specific time interval. The flow q is usually expressed as the vehicles per hour using the mathematical expression:*

$$q = \frac{n_t}{t}$$

Types of Volume Measurements *Other common volume measurement types among in the transportation engineering field are: Average Annual Daily Traffic (AADT), Average Annual Weekday Traffic (AAWT), Average Daily Traffic (ADT), Average Weekday Traffic (AWT).*

Density *Density is defined as the number of vehicles occupying a given length of road or lane. The density k is usually expressed in vehicles per km. Density in a given lane is expressed as:*

$$k = \frac{n_x}{x}$$

Time Headway *Time headway is defined as the time difference between two successive vehicles when they cross a given point. As a derived parameter, the average time headway can be calculated as the inverse of flow, as follows:*

$$h_{av} = \frac{1}{q}$$

Distance Headway *Distance headway is defined as the distance between two successive vehicles at a given time. Average distance headway can also be calculated as the inverse of density, as follows:*

$$S_{av} = \frac{1}{k}$$

Capacity *Capacity is defined as the maximum number of vehicles (usually passenger or similar vehicles) per unit time a road can afford. It is usually determined by physical features of the road, environmental conditions, traffic rules. It is usually obtained through field observations and it is independent of the density. It is often measured in vehicles per hour.*

volume-to-capacity (v/c) ratio *Volume-to-Capacity is defined as the ratio of the flow rate to capacity for a transportation facility.*

$$v/c = x = \frac{q}{c}$$

There are other parameters and relations which are also mentioned in the traffic engineering literature but not used in this thesis, such parameters include, but are not limited to travel time, capacity of the road, journey speed, running speed, urban traffic, or signalized intersection parameters. Signalized traffic flows happen when there is a controlling element, *i.e.* traffic lights, in this thesis only non-signalized, uninterrupted highway traffic flows are considered, *e.g.* motorway *(Autobahn, A)*, 1st class road *(Bundesstraße, B)*, 2nd class road *(Landes- oder Staatsstraße, L or S)*, 3rd class road *(Kreisstraße, K)* [17].

Traffic flow parameters hold a relation, known as the fundamental equation of traffic flow, as follows [37]:

$$q = kv_s$$

Where q is flow in vehicles per hour, k density in vehicles per km, and v the mean space speed in km per hour.

From the fundamental parameters, three relations have been established, in the so called fundamental diagrams of traffic flow, in combination, they give a notion of general traffic behavior characteristics. Figure 3 shows the fundamental diagrams, taken from [37]. This relations are:

- *Flow-Density.* Some of the ideal characteristics are: when density is zero, flow will be zero, as there are no vehicles on the road; when the number of vehicles increases, density will increase; if there are too many vehicles, so they can not move, it is refered as maximum density or jam density; there is a density between zero and

jam density, when flow is maximum. This relation ideally is represented with a parablolic curve.

- *Speed-Density.* Similar to Flow-Density, speed will be maximum when density is minimum, this is refered as free flow, it means that if there are no vehicles, there is no limitation for one vehicle to drive at maximum speed. Opposite situation occurs when density reaches jam density, cars will not move, so speed will be zero. This relation ideally is represented with a line.

- *Speed-Flow* Ideally, when the flow is zero, it is either because there are no vehicles or there are too many that they can not move. At maximum flow, speed is between zero and free flow speed, and that speed can give a threshold for congestions.

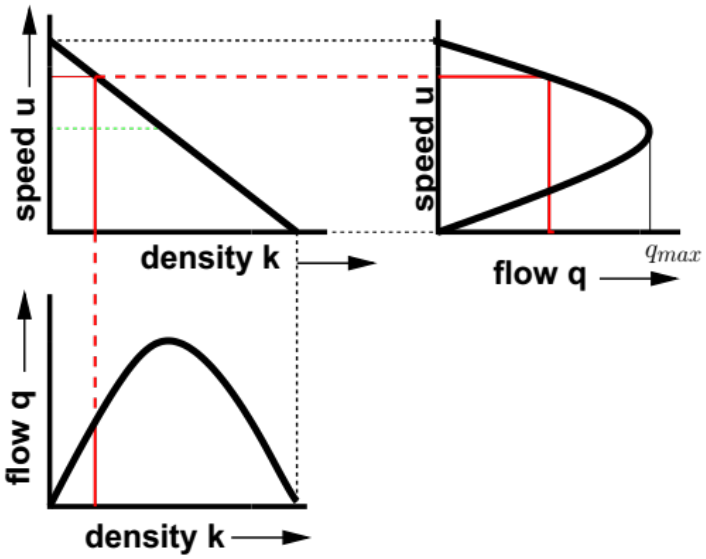


Figure 3: Fundamental diagrams, corresponding to each fundamental relation.

These fundamental relations are the starting point for many more theories around traffic engineering, more specifically in topics related to traffic behavior and road design, including traffic congestions, described in the following subsection.

2.2.2. Traffic Congestions

Traffic congestions are a problem for road transport, according to [29] the cost of road congestion in Europe is estimated to be equivalent to 1% of GDP, therefore the reduction

of traffic congestions is a main priority for many infrastructure, traffic management, and transport authorities. The better information regarding traffic congestions the better perspective to solve problematics affecting people's life, therefore the use of the traffic parameters is essential to determine the traffic conditions, including traffic congestions.

A traffic congestion happens whenever the *Level of Service (LOS)* is not satisfactory, *LOS* is a quality measure of how the traffic demand is in relation the road capacity. *LOS* were defined initially in the *Highway Capacity Manual (HCM)*, developed by the Transportation Board of USA, a range from A to F, where A represents the best operating conditions, and F represents a condition for which the flows are unstable and vehicle delays are high [7]. *LOS* and many other procedures from *HCM* are used as the basis for manuals by many authorities around the world, including the *Handbuch für die Bemessung von Straßenverkehrsanlagen (HBS)* in Germany [35]. *HBS* was written under the German roads conditions and traffic rules to accurately measure the traffic, *i.e.* the lack of speed limits in some highway segments.

HBS describes the methods to determine *Measure of Effectiveness (MOE)*, capacity, and *LOS* in each type of highway facility. Table 1[35], Table 2[34], and Table 3[34] will be used as a simplified process to calculate *LOS* in this thesis, where the travel speed index $I_{VF,N} = \frac{V_{F,N}}{V_{AS,N}}$ is a metric mentioned in [34], and defines an alternative way of measuring the quality of the traffic, based on the relationship between the freeway corridor passenger car speed and the target passenger car speed. This is done due to the lack of additional information regarding roads, such as type as shown in *HBS*, gradient of the segment, and category. Further work can be done to improve an make a more accurate *LOS* estimation according to the formal procedure indicated in *HBS* based on design capacities and the volume-to-capacity parameter. Table 3 English explanation taken and adapted from [37]. German explanation taken from [36].

Table 2 is used as a guideline to determine *LOS*, where each road category is taken from the types of road in [17] and used to know the $V_{AS,N}$ value; $V_{AS,N}$ is assumed to be v_s (space mean speed); due to lack of information, it is assumed no difference in all the sub-categories, as suggested in [34].

Additional labels are considered for the F *LOS*, based on the Swiss Traffic Authority [16], in which terms Slow Traffic (Stockender Verkehr) and Traffic Jam (Stau) are defined as follows:

- **Slow Traffic**, is the status in a main road, when at least for one minute the average speed is lower than 10 Km/h.
- **Traffic Jam**, is the status in a main road, when at least for one minute the average speed is lower than 30 Km/h.

Facility	**Performance Measure**	**Parameter**
Freeway segments	volume-to-capacity	x
Freeway diverge, merge, and weaving segments	volume-to-capacity	x
Segments of rural roads	density	k
Highway diverge, merge, and weaving segments	density	k
Signalized intersections	delay	t_w
Unsignalized intersections	delay	t_w
Segments of major urban streets	density	k
Bicycle facilities	turbulance rate	S
Pedestrian facilities	density	k
Access to parking facilities	delay	t_D

Table 1: Performance measures to determine *LOS* in the *HBS*.

Category	$V_{AS}[Km/h]$	**Types of road [17]**
AS 0/I	90	Autobahn, A
AS II	80	Bundesstraße, B
Stadtautobahnen(AS 0/I and AS II)	70	Landes- oder Staatsstraße, L or S

Table 2: Categories of freeway corridors and target passenger car speed.

Level of Service	**Travel Speed Index** $I_{VF}[-]$	**Traffic State**	**Traffic State in Germany (German text)**
A	≥ 1.25	"Free Flow Zone."	"Verkehrsqualität Sehr Gut."
B	≥ 1.20	"Reasonably Free Flow Zone."	"Verkehrsqualität Gut."
C	≥ 1.10	"Maneuverability Restricted Within The Traffic Flow."	"Verkehrsqualität Befriedigend."
D	≥ 1.00	"Maneuverability Noticeably Restricted Within The Traffic Flow."	"Verkehrsqualität Ausreichend."
E	≥ 0.85	"Stream Operation At Capacity."	"Verkehrsqualität Mangelhaft."

F	< 0.85	"Unstable Traffic Flow."	"Verkehrsqualität Unzureichend."

Table 3: Thresholds of the travel speed index that define the *LOS* in basic freeway segments.

The following steps are used to define *LOS*, and therefore traffic congestions in a specific point:

1. Retrieve measurements from sensors.
2. Compute measurements into traffic parameters.
3. Map the traffic parameters in a *LOS*.
4. Filter by Level of Service F.
5. Determine if it is *Unstable Traffic Flow*, *Slow Traffic*, or *Traffic Jam*.

After introducing the mobility field, the parameters used in traffic engineering, as a description of the mobility on roads, and finally the definition of what a traffic congestion is in terms of those parameters, both fundamentals in *ST* and mobility are the basis of the language used in the following sections.

3. Related Work

In this section it is presented a selection of papers and documents of related work to mobility from a computer science perspective, however other specific to mobility are presented as referent. As an introduction, Table 4 contains a chronologically overview of the research questions or motivations, the computer technique used, in case it was mentioned, the data set description in high level, and the conclusion or outcome of each of the selected documents. Afterwards each document will be further discussed, specially conclusions, contributions, and finally a list of weaknesses and gaps identified that can be improved, either in the contributions of this thesis or as proposed future work.

Pub. Year	**Research Question**	**Compute Technique**	**Data Set Description**	**Conclusion or Outcome**
2006	Empiric description of traffic behavior [30, 32, 31], enhancing mathematical models.	Not explicit computer technique	Individual car probes, not explicitly described	Better understanding of the traffic behavior from empiric observations.
2007	Can an aggregation approach to short-term traffic flow prediction have better results compared to an individual approach? [46]	Time Series (Moving Average, Exponential Smoothing, ARIMA) and Neural Networks (For data aggregation purposes).	Historical Dataset of hourly measured traffic flow (quantity) on one specific point of a road.	Validation of a Data Aggregation approach for a better short-term prediction on a historical dataset compared to an individual approach. Tested with 3-hours ahead prediction. It showed it is more efficient to do the aggregation, instead of a single approach, despite the more complex architecture.

2012	Will more grained road traffic data provide better quantitative estimations of congestion levels? [29]	Trajectory Clustering, Time Series and Statistics (Not explicit)	Historical probes of in-vehicle navigation systems aggregated into a speed profile within a 5 minutes period.	Better quality mapping and monitoring of real congestions. Congestion indicators were used in transport network models.
2013	Forecast impact of traffic incidents on road networks of South California. [40, 41, 51]	Time Series, Neural Networks	Historical collection and real time stream of traffic flow and speed sensor network in LA County. Live stream of incident reports from crowdsourcing and agencies.	Prediction of an incident impact over a route, as a key component in a consumer navigation system.
2014	Measure quality of transport within Europe from a user perspective. [14]	Statistics.	Face-to-face surveys.	General information about transportation in Europe.
2015	Compare different models for real-time short-term traffic predictions. [15]	Neural Networks, different training methods	Historical GPS location and speed obtained from Navigation Systems on cars.	Compare the accuracy of methods for online applications in short (5-15 minutes) and medium (30-60 minutes) terms.
2017	Use of spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks [52]	Artificial Neural Networks	Historical Dataset of 2-minutes collected Traffic Speed of 3 Months	Better accuracy and stability in Short and Long term predictions over SVM, not considering any other factor.

Table 4: Related work summary.

3.1. Discussion of Related Work

The approach presented in [46] is to use moving average, exponential smoothing, and *Autoregression Integrated Moving Average (ARIMA)* individually as the basis for prediction over three different historical data sets of the traffic flow in a specific point, weekly, daily, and hourly time series. With the prediction results, an artificial neural network is used to compute an aggregation as the final step of the prediction. The main conclusion is that the use of an aggregated approach provide more accurate results. A contribution is that gives a background to calculate more accurate predictions by using different tools in combination. Despite the contributions and positive results, there are weaknesses and gaps, such as the limited use of traffic parameters, only traffic flow was considered; there are not considerations for scaling to more sensors; the approach is faulty if an external factor influence the traffic flow, it is mentioned that if an accident happens, the pattern will be lost.

[29] was meant to give an overview of traffic congestions impact and support policy making for authorities in Spain. A mesoscopic approach was taken, with GPS probes and an undeclared trajectory clustering technique was used to create speed profiles, and with statistical methods an assessment was done. The main conclusion is that the use of the methodology allows to have a better assessment and comparison of traffic congestions across countries and areas. The most important contributions are a more accurate use of available data and algorithms to create profiles along the roads, using specific traffic engineering parameters; different techniques were used to create a richer map and assessment, including the trajectory clustering techniques, a *ST* technique. Regarding weaknesses and gaps, [29] focuses only on an overview and assessment of traffic congestions, not considering any other trigger than the seasonality of traffic flow and speed; the study does not clarify which algorithms were used and does not provide detailed information of the data set.

As part of a project [40, 41, 51] present a progression of different elements towards a system which aims to help reducing traffic congestions and improving the user experiences. Those elements are discussed as follows.

[40] introduces the use of real-time data, including the public transportation agency, the highway patrol, specially traffic detectors on the roads. The use of time series techniques to create models is the basis to predict traffic congestions, in particular *ARIMA*, *Historical Average Model (HAM)* and an enhanced version of *ARIMA*, proposed by them, the first goal was to evaluate how the algorithms prediction accuracy is with different horizons of prediction, it means short or long term predictions. The second goal was to evaluate the how the algorithms adapt to sudden changes, *e.g.* in case of an incident happens, used

due to the availability of a descriptive incident data set, introducing also the correlation between both events, traffic congestions and road incidents.

[41] extends the prediction of a road incident in the ST dimensions, in particular the exact length of time of the evolution of congested spatial span. One of the core concepts presented in this paper is the propagation behavior, as the evolution of the distance from the incident location over time. The impact prediction is done in a multi-step process, initially with incident attributes, such as number of lanes affected, fatality, highway, and direction, etc. Secondly, additional data is considered, in particular traffic density, under the assumption that more crowded areas have higher impact after an incident, considering similar incident attributes. Finally, in a third step, it is suggested to add other type of data to get a more accurate prediction, such as weather or other sources of data not available. The prediction improvement was done by an initial behavior similarity in the initial timestamps.

[51] is one of the latest papers related to this project, it considers the evaluation of real time data with historical data that better match in conditions, the so called *context*. The hypothesis behind it is that predictions are more efficiently estimated if the context space is adaptively partitioned to choose the best base predictor in each situation. One of the drawbacks mentioned in the paper is the exhaustive and impractical training stage for each traffic situation, where the complexity relies on the numerous situations or contexts, depending on the features considered in the context space. Therefore the main task is to select the context more relevant to the traffic prediction, reducing the implementation complexity and providing guidelines for traffic policy making. The context information can include but it is not limited to location context, such as road or area type; time context, whether on weekday or weekend, daytime or night, rush hours; incident context, what type of incident, number of affected lanes; and finally, any other context that is available in a data set. The conclusion of this paper is that context information improves the choice of predictors. The contribution is the confirmation of the hypothesis that a context is an important element on prediction tasks. It is remarkable the approach and achievements covered by the overall project, however the main weakness is that it only relies on the road context information, despite it is mentioned that it is desirable to include other data sources, such as weather, temporary events.

As part of the authority generated studies [14] presents a survey report, where the main objective is to get information from citizens, including habits and perceptions of transportation. The survey was done over the 28 European Union members in 2014. It present general statistics which can be useful to have a better understanding from the user point of view. Contributions are more valuable in terms of a higher level authority organization, however it can provide a new perspective to integrate context information, if data is available, such as the perception of quality, *i.e.* citizens in Germany indicated road

maintenance and amount of freight transported by road as the most serious problems after road congestions. The importance of this relies on the possible insights that can be discovered both ways, from transport users and policies.

The main focus in [15] is the evaluation of *Neural Networks (NNs)* and *Bayesian Networks (BNs)* for different short (5-15 minutes) and medium-terms (30-60 minutes) predictions. Additionally to the evaluation of both methods and temporal horizons, it is proposed to consider upstream and downstream traffic flow, which means the traffic flow before and after a reference point, the paper focuses in the flow previous and after one intersection, as well as both directions, foreward and backward. The main conclusion is that the use of different methods can benefit the prediction in different conditions, supporting ideas shown in previously discussed papers. The main contribution, in terms of the conditions set in this thesis, is the consideration of a more complex interaction, using data available for both foreward and backward traffic flow.

Inspired by image processing techniques using *NN*, [52] presents a novel representation of a road network and traffic behavior prediction approach. This approach considers a 2-Dimensional grid, similar to the grids used in images and videos, each grid box represents a spatial region with an average speed, so that roads, intersections, ramps, and curves are represented within the grid, the goal is to predict the future values of each grid box. Spatial relationships are captured by a *Deep Convolutional Neural Networks (DCNNs)*, treating the grid as an image, where each pixel is a traffic state. For the temporal features *LTTM Neural Networks (LTTM-NNs)* are used, because of the performance shown for long term learning. The architecture *ST Recurrent Convolutional Network (SRCN)* uses *DCNNs* and *LTTM-NNs* in different layers, so that the *ST* features are processed. The results show that using this approach can be more efficient than the time series approach. Some contributions of this paper are the approach that does not rely on the previous traffic flow models, avoiding strong assumptions, leading to a more simple construction of the prediction, tested previously in video and image processing. Also the use of a grid gives a simplified way of processing interconnections and other elements on the road, such as ramps, without the complexity of modeling explicitly each one. The main weaknesses and gaps are towards the limited inclusion of additional features, such as weather or external events, so that the predictions can be done also with different contexts.

From the traffic engineering perspective, [30, 31, 32] is a set of studies based on empirical features of traffic breakdown, a proposed transition between free flow to congested traffic. The aim is the establishment of the bases for any traffic and trasnportation theory used for control and optimization in traffic networks. It represents an alternate methodology to previous theories, including the undetrstanding of highway capacity, used in *HCM* and *HBS*. The main element is the set of transitions from Free Flow to Syncronized Flow and finally to Wide Moving Jam, and vice versa. The inclusion of this work is

meant to be as a referent from traffic engineering field, and as a contrast to the work done from the computer science, as this last approach has focused efforts in predicting behavior, specifically prediction of speed or traffic flow under the same set of rules and assumptions. One of the main criticisms of Kemer's work on three phase traffic is the fact that observations were done in a specific highway in Germany, where, for instance, in some segments there is no speed limit, discussions are on-going [32, 45] to determine whether it is valid as a general theory under other conditions.

3.2. Weaknesses and Gaps in Related Work

As a summary of the related work, the weaknesses and gaps are used as motivation for this thesis and further work, the following list depicts aspects that ideally a traffic related system should consider.

- In the related work it is limited the use of traffic parameters, and it is either focused on speed or quantity of vehicles on a road. The use of more traffic parameters, such as space mean speed and density, can enhance the context of traffic congestions. That can also help further studies in the traffic engineering field, however for other applications it can cause saturation of information, for instance in an end-user routing mobile application. Traffic Parameters are defined in Section 2.2.1, including the formulas to calculate them.

- The related work focused either in a single sensor or a segment of the same road and the same direction. The use of as many sensors in an area as possible, whether upstream, downstream, or different directions, has not been explored before. Considering Tobbler's 1st law of geography, also closer sections or road elements can influence in the traffic congestions.

- Only recent related work is using spatial and temporal dimensions, following that direction, the extensive use of them, including a common *STDT*, should be considered as part of the scalability. That means, in case the this thesis shows positive results, there should be an established process to integrate new sources.

- Most of the related work focuses in the prediction algorithm and do not consider scalability in a distributed approach, therefore the use of algorithms suitable for distributed processing should be considered. As mentioned in [1], one of the main challenges is to exploit the bigger amount of data, and distributed systems have shown good results in different use cases.

- Only few additional factors have been included in the related work, specially in [40, 41, 51], therefore it should be consider the inclusion of external factors, more over not road-related, as the main contextual information.

- As mentioned in [1], real-time processing and interactive analytics are a challenge to solve within *ST* systems, most of the previous related work is not considering any real-time processing, this approach should be considered.

- Previous works focused only in a single approach, ideally a system should allow change or compare algorithms or frameworks. The evolution of knowledge and approaches is continuous, usually it requires further studies and comparisons to generalize or extend them, for example the suggested approach by Kemer can be compared with other traffic engineering theories in a faster way.

- Use of open data, when possible. It will allow further research on the field by organizations and individuals interested in the topic.

As mentioned, this thesis is not meant to cover and develop all the gaps and weaknesses listed above, however the focus is to consider areas instead of single sensors or road sections, and non-road-related external factors. In Section 4 the methodology is presented, using the background from Section 2 and motivations from Section 1. Also in the methodology section, the specific goals of each step are mentioned, what the expected results, and the challenges in such system.

4. Methodology

In this section a more detailed explanation on how this thesis was conducted is given. The following research question is used as a guideline: *is a specific type of events in an area likely to be part of the context of traffic congestions?*. At first it is presented the data sources, followed by an overview on the process, steps, expected outcomes. Further in this section, it is detailed how the data is transformed along the process and the rationales taken. Each step of the process is explained, with the corresponding algorithms, parameters, and expected outputs in separated subsections. Finally, a subsection is focused on the considerations to use the process at scale, from a big data perspective.

4.1. Data Sources

Each data source used in this thesis is described in terms of the format, available fields, and the corresponding *STDIs* transformations. As explained in Section 2.2 traffic congestions happen when the free flow is affected, based on the definitions given in Sections 2.1.1 and 2.2.1 instances used in the subprocesses are explained in the following subsections.

4.1.1. Hessen Mobil QnV, Traffic Congestion Event

Hessen Mobil - Straßen- und Verkehrsmanagement, shortened as "Hessen Mobil" from now and on, is the federal state of Hessen government agency in charge of administration, construction, maintenance, and monitoring of all roads and highways within the federal state of Hessen. Hessen Mobil provides information of the quantity *(q)* and speed *(v)* collected periodically from around 2500 sensors over highways and roads. The data is publicly available at the MCloud[9] platform or provided by the *Mobility Data Marketplace (MDM)*[10], in the form of an XML file following DATEX II [11] standard, a Europe-wide *ITS* standard. The XML file provides a reference field the sensor where the data was collected, the measurement of quantity or speed, and finally instant when the data was collected, event time in Data flow model terms [2], Listing 7, in Section A.1 is an extract of the XML file, that will be referred as *QnV XML* file, containing the following elements, values, and attributes:

- **measurementSiteReference** element has the attribute **id** to identify a single sensor.

[9] https://www.mcloud.de/
[10] https://www.mdm-portal.de/
[11] https://www.datex2.eu/

- **measurementTimeDefault** element has a text value, the instant when the sensor emitted the measutement. It is expressed in ISO 8601: YYYY-MM-DD'T'hh:mm:ss+ZZZ format[12].

- **measuredValue** element has the attribute **index** used in combination with the Location XML file to obtain detailed data for that specific measurement.

- **basicData** element has the attribute **xsi:type** used to define whether the measurement is a q (TrafficFlow) or a v (TrafficSpeed) value.

- **vehicleFlowRate** and **speed** elements contain numeric values for quantity and speed, respectively. Quantity is the *Flow* and speed is the *Spot Speed*, described in Section 2.2.1.

The *QnV XML* file does need the complementary spatial information, as well as the description of the measurement site, for instance the road, the lane, or the type of vehicle, contained in the measurement site table file, that will be referred as the *Location XML* file, provided as well by Hessen Mobil, and Listing 8 in Section A.1 is sample extract of the file. Location XML file contains the following elements, values, and attributes:

- **measurementSiteReference** element has the attribute **id** to identify a single sensor, matching with the same element and attribute in QnV XML file.

- **measurementSiteIdentification** element has a text value, with the name of the road or highway code, as defined in [17], an additional identification is present, however it is ignored in this thesis.

- **measurementSpecificCharacteristics** element has the attribute **index** matches the attribute **index** of the **measuredValue** element in the QnV XML file, used to identify each measurement characteristics.

- **period** element has a numeric value, is a representation in seconds of the interval of measurements from the sensor. For example, if the value is 60, the interval measured by the sensor is 1 minute, in case of the speed, is the spot speed in the previous minute, and in case of quantity, is the amount of cars counted in the last minute multiplied by 60, to be consistent in the units of q in vehicles per hour.

- **specificLane** element has a value indicating which lane is measured, where lane1 is the most left lane, and the number increases towards the right lanes.

[12]https://www.w3.org/TR/NOTE-datetime

- **specificMeasurementValueType** element has a matching value with the **basicData** element in the QnV XML file.
- **vehicleType** element has a value of the type of vehicle measured by the sensor, for example a car, refers to a passenger car, as specified in *HBS*.
- **measurementSiteLocation** is an element containing more elements and values related to the characteristics of the road link.
- **alertCPoint** is a wrapping element containing other elements with information from *Traffic Message Channel (TMC)*, *Location Code List (LCL)*, and *Event Code List (ECL)* [13], associated to the definitions in [17]. This information is only available in sensors on the road, excluding the ones on the exits.
- **alertCDirectionCoded** element has a value for a codad direction, either positive or negative, related to the link description in [17].
- **specificLocation** element value refers to the *LCL* table and coding.
- **offsetDistance** element value is the distance in meters from the **specificLocation** in the direction indicated by **alertCDirectionCoded**.
- **pointByCoordinates** element contains two more elements, **latitude** and **longitude**, with values in *WGS84*.

Considering every single measurement as an individual record, called *QnV Record*, for each fetched QnV XML file, around 16600 records are received per minute. This number of *QnV Records* may vary from the collected data due to missing records, failure in the communications, or a deprecated sensor.

In this thesis the main focus is the data generated by the sensors and, following guidelines from Section 2.2, it will be preprocessed before the *ST* algorithms ingest them, this will give a better quality in the data, as *Traffic Parameters* are included. This preprocess is explained in detail in Section 4.3. Additional data is used in this thesis, as part of the contextual information of a *Traffic Congestion*, the sources and structure of the additional sources are explained in Section 4.1.2.

[13] https://www.bast.de/BASt_2017/DE/Verkehrstechnik/Fachthemen/v2-LCL/location-code-list.html

4.1.2. External events

As mentioned in [50], interconnecting contextual data for transport-relevant events has not been used largely in current mobility systems, this data includes, but is not limited to, weather conditions, events such as concerts, football games, and local sights. The main challenge relies on the variety of sources and formats, mainly available through different platforms and, sometimes, only provided in a human readable format, which has to be transformed into a structured data. In this Section the additional sources considered in this thesis are explained and in Section 4.6 is explained how it is transformed into an *STDI* to be used in the algorithms along the process.

According to [12] there are two types of events, depending on how the event is generated: (i) spontaneous and (ii) planned. In this thesis, it is meant to cover both as part of the *External Events*, therefore two additional sources of events are suggested.

In Section 4.1.1 was explained how *QnV Records* are obtained as part of a real data set, which is used in some of the experiments in the Daystream Project, at the moment this thesis is printed it is an in-progress project. This situation leads to consider at first labeled test data, generated with the purpose of going through the general process and confirm that it can be used with other real data. This test data will be located in an airport within the selected area and labeled as "Flight Delays", so that it has a meaningful description. This data set will comprise the spontaneous events.

In order to have a "Flight Delays" test dataset to test the process to include contextual data into *Traffic Congestions*, some of the samples will be explicitly created to match *Traffic Congestion Areas of Interest* detected *a priori*.

The second type of events, used to evaluate a planned event, are extracted from the public available data of football teams in the area of Hessen, it will be selected a stadium and a team within the area under the criteria listed in Section 4.5. The label used to identify this events will be "Football Match"

The following features and attributes are expected from both event type data sets:

- **Type of Event**. The label selected for each type of event, either "Flight Delays" for the spontaneous events or "Football Match" for the planned ones.
- **Location**. Specific coordinates of the location of the selected airport or stadium.
- **Start Time**. Date and time when the event started.
- **End Time**. Date and time when the event finished.

- **Duration**. In case "End Time" is not available, an estimated duration will be provided and "End Time" can be derived from "Start Time" and "Duration".

In Section 4.6 is explained how these additional sources are transformed into an *STDI*, which additional assumptions and considerations are taken for each event type.

After introducing the data sources, both *QnV Records* and the additional sources representing *External Events*, the overview of the proposed methodology in this thesis is explained.

4.2. Overview of the Process

The methodology divides the process into two steps, see Figure 4 (i) detecting areas of interest, and (ii) levering context information to the selected areas of interest. The description of each step is as follows:

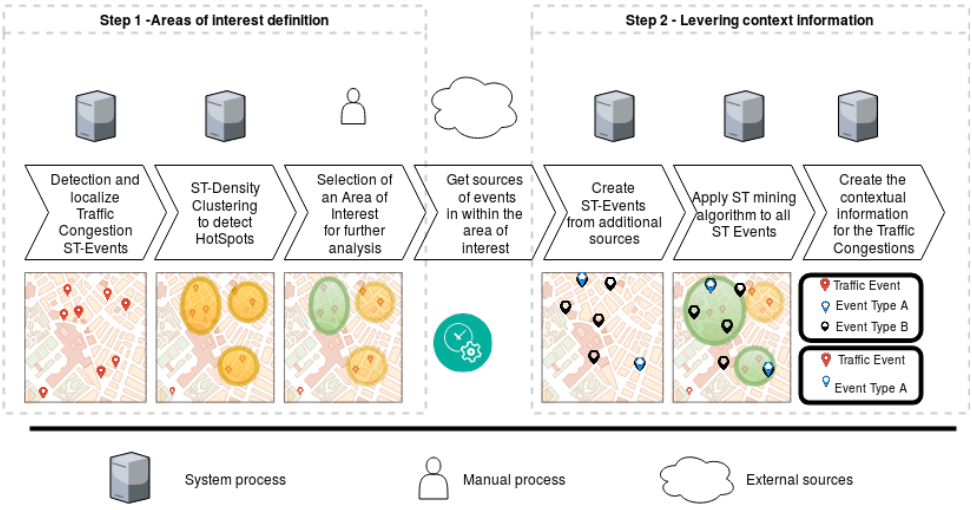


Figure 4: Linear sequence of subprocess.

Detecting Areas of interest. In contrast to previous works mentioned in Section 3, the main focus of this thesis are the areas where traffic congestions are denser, also known as hotspots. Instead of considering a single point on the road, or a road section linked by sensors. The detection of these areas is calculated by a density clustering algorithm from individual traffic congestion events data.

Levering context information. Context information is considered as other type of events in the spatial and temporal vicinity of the areas of interest, moreover the ones not related to the road. The selection of the contextual information is highly dependent upon the availability of open data. With both, areas of interest and nearby events, a *ST* mining technique is used to identify frequent patterns to evaluate. In this step, it is also possible to evaluate different notion of locality of external events, this allows to discard events too far in time and space to the areas where the traffic congestions are.

This process is implemented in a system that ingests the raw data produced by the road sensors, transforms it into a *STDI*, and process it with the selected algorithms to generate the areas of interest. Manual intervention is suggested to delimit the search of external event sources. The 2-Step process is depicted in Figure 4 as a linear sequence of subprocesses, which are described in more detail. In the bottom of each subprocess a visualization is shown as a graphical explanation of how the process would be.

Subprocesses are described as follows:

- **ST Traffic Congestion Events Detection and Localisation.** Based on the definitions given by *HBS* and [34] mentioned in Section 2.2 the traffic data is pre-processed to create a traffic congestion instance, the use of traffic parameters is essential for data quality.

- **ST Density Clustering.** ST Traffic Congestion Events are clustered by density in both dimensions, space and time, to spot areas with high concentration of traffic, regardless of the traffic features, such as road, type of road, and direction.

- **Manual selection of an Area of Interest.** This selection is done manually in order to delimit the search sources of external events. However the basis to integrate further sources will be established and make an automated process.

- **Creation of ST Events for other additional sources.** This subprocess will allow ingesting data from other sources, wrapping them with an *STDI*.

- **Apply ST Mining algorithm.** This subprocess extracts the external events in the space and time vicinity of a traffic congestion area, this information will be used to create the context of the traffic congestion.

- **Contextual information for traffic congestions.** The context of a traffic congestion should be descriptive, not meant to provide conclusions on such relations. A possible output is a human readable format, i.e. "Traffic congestions in the area

close to Place *"Name"* (with characteristics A, B, and C) have appeared while D, E, and F events were happening".

4.3. Preprocessing of data

As an initial processing phase, as shown in Figure 5, measurement and time-related fields from QnV XML and location description fields from Location XML files are joint to create a *Traffic Record*, as an initial *STDI* and contribution of this thesis. A Traffic Record is an abstraction of the situation reported by a sensor, aggregating speed and quantity measurements from different lanes and vehicle types, so that it contains more descriptive information of a point in a road within a single record. Compared to the up to 12 individual records for a specific instant sent by a sensor, a Traffic Record is easier to process in further steps, as the granularity of the data is at specific location in a road despite the number of lanes. This information is used to map the traffic condition into one of the 5 *LOS*, defined in Section 2.2.2. The main reason to create a Traffic Record relies on the fact that it contains the initial detection and localisation of a traffic congestion in time and space.

As the main focus of study in this thesis, a *Traffic Congestion Event* will be defined as the *Traffic Record* reporting an unstable traffic flow, defined as *LOS* **F** in Table 3, especially the ones falling in the category of *Slow Traffic* and *Traffic Jam*.

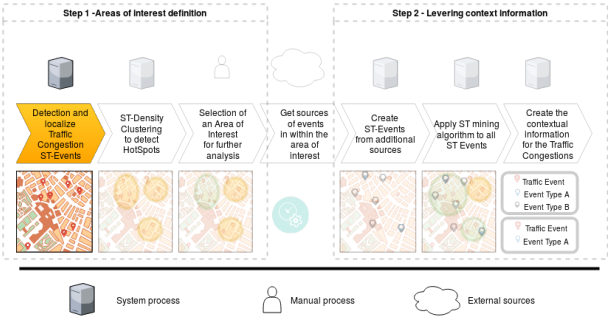


Figure 5: Creation of a Traffic Record and Traffic Congestion Detection and Localisation.

A Traffic Record contains time, space, and traffic parameter fields, calculated as mentioned in Section 2.2.1 and described in detail as follows:

- **sensor id** is the identification for each sensor, extracted from the element **measurementSiteReference** and attribute **id** in both QnV XML and Location XML files.

- **date time** is an instant extracted from the **measurementTimeDefault** element, as the main temporal field.
- **timestamp** is a calculated field from the *UTC* **date time** in milliseconds.
- **location** is the geometry representation of the sensor obtained from **pointByCoordinates**.
- **period** used to calculate the *valid time*, it is the value of the interval in seconds for which the measurement was done in the sensor.
- **traffic flow (q)** is the aggregation of the passenger vehicles for all the lanes at a specific sensor.
- **time mean speed (vt)** is calculated using the **speed** value from all the lanes, as the average speed.
- **space mean speed (vs)** is calculated using the **speed** value from all the lanes as mentioned in Section 2.2.1.
- **headway (h) and (s)** headway derived from q and v in meters and seconds, respectively.
- **density (k)** is calculated as a derived value from q and v in passenger cars per kilometer in all the lanes.
- **passenger vehicle ratio** is calculated as the percentage of passenger vehicles from the total vehicles, including lorrys.
- **lane quantity** obtained from the count of **specificLane** elements.
- **level of service** calculated as mentioned in Section 2.2.1, using the *Travel Speed Index* [34].
- **short description in English and German** is a human readable text, from Table 3. For *LOS* F, it is added the different state of *Slow Traffic* or *Traffic Jam*, as mentioned in Section 2.2.1.

The creation of a Traffic Record is done in two steps, the enrichment of a QnV Record with information from the Location XML file and the aggregation of *QnV enriched records* into traffic parameters, as seen in Figure 6. The enrichment of a QnV Record is an intermediate process which expands attributes regarding vehicle type, lane, road, and other to the measurement recorded by the sensors. This enrichment provides enough

information to interpret the values and derive the missing ones, for example sensors from Hessen Mobil do not provide directly traffic flow measurements for passenger cars, instead sensors provide traffic flow for any vehicle and for lorrys, therefore the passenger cars traffic flow measurement can be obtained by subtracting the flow for lorrys from the any vehicle one. This process gives all the data needed to apply the formulas explained in Section 2.2.1 to obtain the traffic parameters.

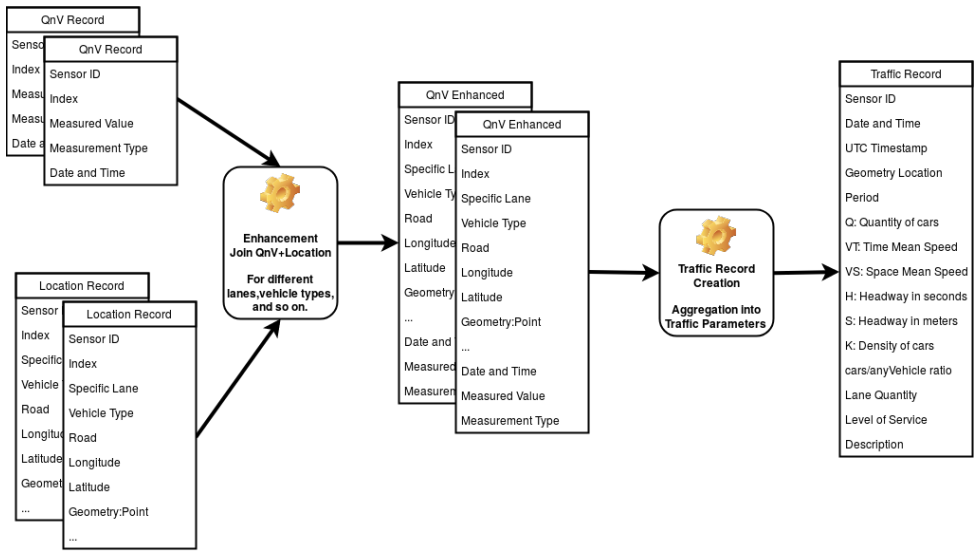


Figure 6: Creation of Traffic Record from QnV Record and Location Record.

Listing 1 contains a pseudo code of the transformation from the QnV XML and Location XML file into Traffic Records, for example. This is the first subprocess in the step of Areas of interest definition, as it is shown in Figure 5, with the title and the figure highlighted in that specific subprocess. As mentioned at the end of Section 3.2, the use of more traffic parameters is a gap covered in this thesis.

```
1  fetch QnV XML file(s)
2  parse QnV XML file into QnV Record
3  parse Location XML file into QnV Location
4  join QnV Record and QnV Location where (id and index match)
5    group by sensor_id and date_time
6    reduce group
7      set sensor_id, date_time, timestamp, road, geometry, LCL data, period
8      set spot_speed and flow for each lane and vehicle_type
9      set number_lanes
10     //Calculation of Traffic parameters
11     //The most important part in the Traffic Record creation
12     if data contains vehicle_type = car
```

```
13          //Missing values are derived
14          //as LOS is defined for cars and not lorrys
15          //some measurements are derived,
16          //such as car[flow] = anyVehicle[flow] - lorry[flow]
17          assign missing_values
18          calculate flow totals by vehighlitedhicle_type
19          calculate harmonic qi_vi //used in space_mean_speed
20          //Calculate and set Time Mean Speed
21          vt = avg(car[speed])
22          //Calculate and set Traffic Flow
23          q = car[flow] //all lanes
24          //Calculate and set Space Mean Speed
25          vs = q/harmonic qi_vi
26          //Calculate and set Headway in seconds
27          h = 1.0/(q/number_lanes)
28          //Calculate and set Headway in meters
29          s = vs * h
30          //Calculate and set Density in all the lanes covered by the sensor
31          k = 1.0/s * number_lanes
32          //Calculate Travel Speed Index with VS
33          travel_speed_index.calculate(vs,road_type)
34          //Map and set LOS and LOS_highliteddescription
35          los.map(travel_speed_index)
36      filter LOS=F
37      filter Slow Traffic and Traffic Jam
```

Listing 1: Traffic Record transformation pseudo code.

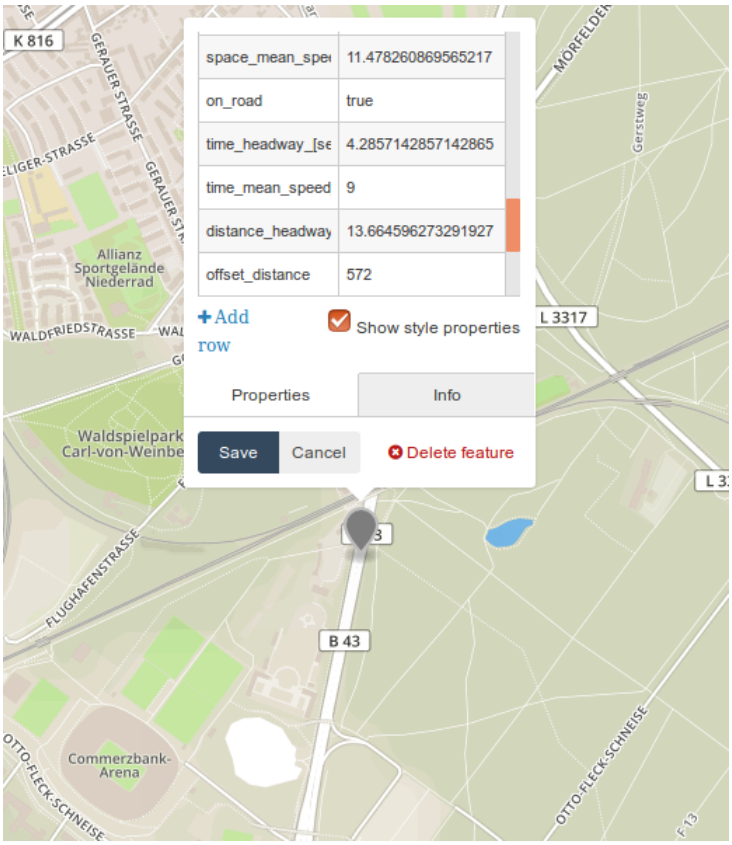


Figure 7: Map with one Traffic Record.

The transformation from *QnV Records* into *Traffic Records* is done for every sensor, Listing 9 is an example of a Traffic Record formatted as a GeoJSON object. Figure 7 is a graphical representation of a sensor, including some of the traffic parameters.

The preprocessing of data creates *STDIs* with time and space as the main *ST* features, and traffic parameters as non-*ST* features, used in the following step, explained in Section 4.4, which focus is to discover areas of congestions. As mentioned in Listing 1, the *Traffic Records* are filtered, only Slow Traffic and Traffic Jam are passed as a traffic congestion state representation.

4.4. Finding Traffic Congestion Areas of Interest: ST-Clustering

This section describes the subprocess called *ST-Density Clustering to Detect Hotspots*, as shown in Figure 8. It is also explained what a *hotspot* or *Area of Interest* is. From the

previous process, explained in Section 4.3, it is explained how *Traffic Records* are used to detect *areas of interest* and how changes in traffic congested areas are tracked.

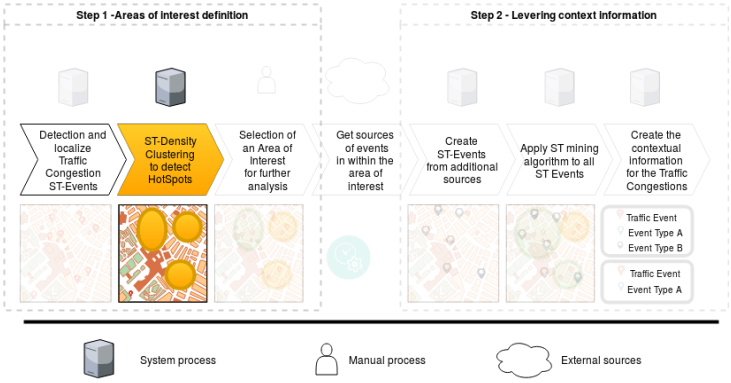


Figure 8: ST-Density Clustering to detect Areas of Interest (Hotspots).

In contrast to previous works and as mentioned in Section 4.2, one of the contributions of this thesis is to examine an *Area of Interest*, instead of just a single road spot or segment. Such task consists on determining where the concentration of individual traffic congestion events is dense, no matter the road, direction, or if it is an upstream or downstream flow. The following assumptions and requirements are made in the definition of an *Area of Interest*:

1. The number of areas of interest is *a priori* unknown. As traffic congestions can happen in different places by different reasons, one of the main reasons for this task is to discover where and how many *areas of interest* are there.

2. The shape of areas of interest is arbitrary, since the distribution of sensors varies in each road segment and in junctions, as well as the distance along the road, therefore the shape a cluster of sensors reporting traffic congestion is not fixed.

3. Areas of interest evolve over time, in shape and number. From an *ST* perspective, entities change over time, in case of traffic congestions, new ones appear, traffic congestions areas can increase or decrease, and finally get dissolved, so that the traffic conditions are back to free flow. Therefore, cluster *STDI* should keep track on those changes.

As explained in [26] *clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Usually, distance functions are used to measure similarity.* In this thesis, an *Area of Interest* is represented by a cluster and

the objects, or elements grouped into that cluster, are the Traffic Congestion Events, which were explained in Section 4.3. Clustering algorithms are focused on the creation of the clusters as a static dataset, however the elements can change over time, therefore an algorithm to manage the non-static behavior is needed.

In particular, density based clustering algorithms, *DBSCAN* [13, 44] or *OPTICS* [4] for example, are suitable to identify clusters with arbitrary shape and do not require number of clusters as an input parameter, in contrast to *K-Means* for example. *DBSCAN* is used in this thesis as the clustering algorithm, because of a more straight implementation compared to *OPTICS*.

The density-based model introduced by *DBSCAN* in [13] include the following definitions:

- **density** is defined as a minimum threshold of elements in the vicinity of a particular element. It is denoted as *MinPts*.
- **radius** is defined as the distance of the vicinity. In geometric data sets it is usually a euclidean distance, however it can be any other distance metric. It is denoted as *eps*.
- **core element** is such element in a data set that within its vicinity there are at least *MinPts* elements.
- **density reachable**, an element is density reachable if it is part of the vicinity of a *core point*.
- **border element** is such element in a data set that is density reachable, but has not enough elements in its own radius.
- **noise element** is such element in a data set with not enough elements in its vicinity to be considered as *core* and is not *density reachable* by other *core points*.

DBSCAN algorithm works over a set of elements, processing one element at the time, it first identify if the element is a *core element*, otherwise it is considered as *noise*. If the element was labeled as *core*, the algorithm explores the neighbors, to determine whether they are considered as *core* or *border elements*. Figure 9 depicts the *DBSCAN* model, where red points, for instance A, are *core elements*; yellow points are *border elements*, B and C are considered as so; and finally the blue point is considered as a *noise*, (point N) as it does not have enough elements in the vicinity and it is not reachable by any *core point*.

Listing 2 is the pseudocode of the *DBSCAN* algorithm as defined in [44].

```
Input: DB: Database
Input: eps: Radius
Input: minPts: Density threshold
Input: dist: Distance function
Data: label: Point labels, initially undefined
  foreach point p in database DB do // Iterate over every point
    if label(p) is not undefined then continue // Skip processed points
    Neighbors N = RangeQuery(DB, dist,p,eps) // Find initial neighbors
    if |N| < minPts then // Non-core points are noise
      label(p) = Noise
      continue
    c = next cluster label // Start a new cluster
    label(p) = c
    Seed set S = N \ {p} // Expand neighborhood
    foreach q in S do
      if label(q) = Noise then label(q) = c
      if label(q) is not undefined then continue
      Neighbors N = RangeQuery(DB, dist,q,eps)
      label(q) = c
      if |N | < minPts then continue // Core-point check
      S = S union N
```

Listing 2: DBSCAN algorithm.

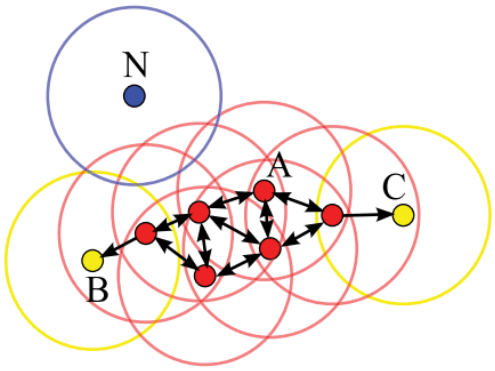


Figure 9: *DBSCAN* cluster model.

In order to keep focus on the traffic congested areas, *DBSCAN* algorithm ingests filtered *Traffic Records*. The only criteria is to filter traffic jams and slow traffic reports, in this way the number of elements and processing time tend to be lower than processing the entire data set of sensors.

As explained above, *DBSCAN* parameters are *eps* and *MinPts*, changes in those parameters result in different number and shape of clusters. In Figures 10, 11, and 12 an example

of how *DBSCAN* calculates the clusters from Traffic Records, core elements are in red, border elements are in yellow, and noise elements are in grey.

The change of *eps* may influence the label of some elements, either it can be considered as core or border elements, therefore the shape of the bounding polygon of the cluster can change, this is discussed with the following example.

As an example of how *DBSCAN* works over a Traffic Records data set, Figure 10 shows two clusters (A and B) calculated with parameters $eps = 2000$ meters, $MinPts = 2$; for this specific data set, distance between elements of each cluster is short enough to make all of them as core elements, however the space between clusters is big enough to differentiate them.

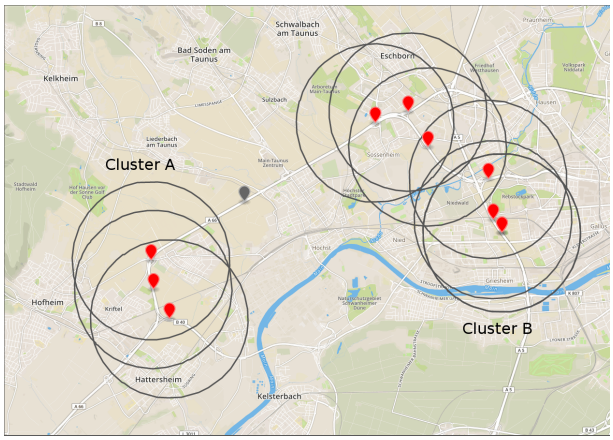


Figure 10: *DBSCAN* cluster model over Traffic Records with traffic congestion, parameters $eps = 2000$ meters, $MinPts = 2$. Core points are in red color and noise point in gray.

To clarify how *DBSCAN* parameters can influence the form of the clusters, in Figure 11 the parameter of *MinPts* is changed, now each point needs 3 neighbors instead of 2, from Figure 10 cluster A is not longer valid, because cluster A had only 3 elements, that means each element had two neighbors and then each one can be considered as core element. Now with *MinPts*=3 any of these points is has enough neighbors, therefore no cluster is detected. Cluster B now has two border elements that before were core elements, because in their vicinity there are not enough neighbors, however as they are still within the vicinity of other core elements the yellow elements remain in cluster B.

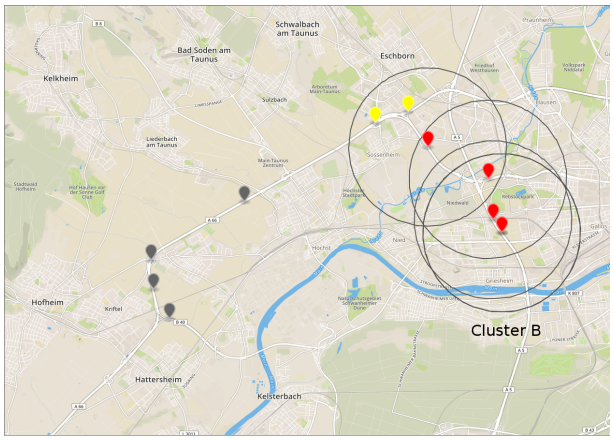


Figure 11: *DBSCAN* cluster model over Traffic Records with traffic congestion, parameters $eps = 2000$ meters, $MinPts = 3$. Core points are in red color, border points are in yellow, and noise in gray.

Finally, a clustering algorithm output can include the elements or the bounding area, in this thesis it is considered the resulting area as the main output, including the radius of the core elements, so that it can be used as the traffic congestion area to be enriched with external information. Figure 12 shows the resulting bounding polygons of the *DBSCAN* example with parameters $eps = 2000$ meters, $MinPts = 3$, explained in Figure 11. Radius from border points is ignored.

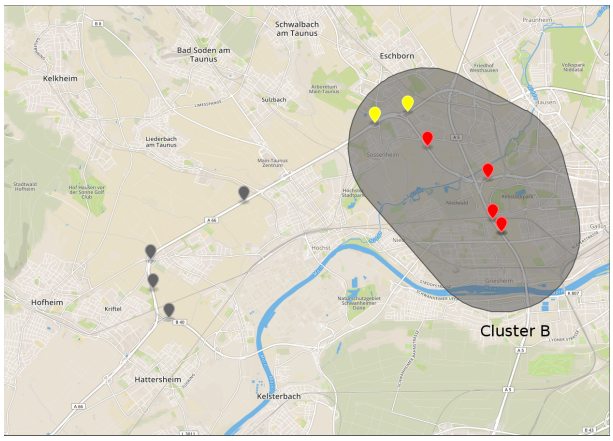
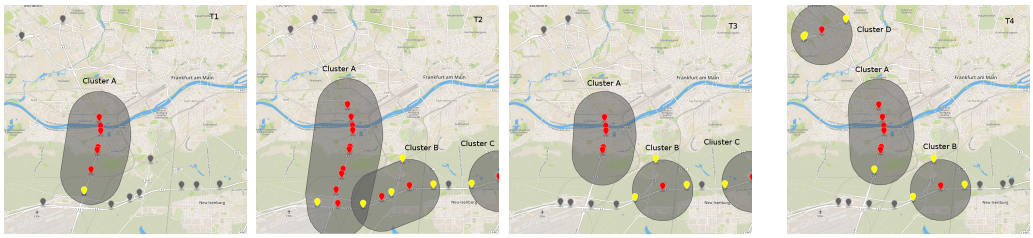


Figure 12: *DBSCAN* cluster model over Traffic Records with traffic congestion, parameters $eps = 2000$ meters, $MinPts = 3$. Core points are in red color, border points are in yellow, and noise in gray.



(a) t=1 (b) t=2 (c) t=3 in (d) t=4

Figure 13: Evolution of clusters within 4 consecutive minutes.

The optimal selection of parameters for the algorithm is mentioned in [44] as the result of different experiments, where heuristics point towards a *MinPts* parameter set to as twice the dimentionality for data sets with low levels of duplicates and noise, not too large, low dimensionality. It is mentioned that the radius *eps* is generally harder to set, in an ideal scenario an expert in the use case domain should give the notion of what a *neighbor* is.

The management of evolving areas of interest over time is not considered initially in the *DBSCAN* algorithm, however it established the basis for other algorithms with further functionalities, for instance *ST-DBSCAN*[6] which considers additional features in a time frame, moreover to identify distinct adjacent clusters; another algortithm based on *DBSCAN* is *DenStream* [8], which manages the evolution on a data stream, using parameters of validity of the clusters, which means the algorithm identifies when a cluster is still valid; among other approaches, in [28] a simple algorithm is presented to define moving clusters in intervals by calculating the similarity of elements of calculated clusters in consecutive instants.

Snapshots of four consecutive instants, shown in Figure 13, confirm that a cluster changes over time, from a human understanding the cluster tagged as Cluster A is present in all the four consecutive instants, starting in Figure 13a as the starting instant. Cluster B and C are detected in the second and third minute, shown in Figure 13b and 13c, respectively. Finally, Cluster C is no longer detected but Cluster D is detected in the last minute, shown in Figure 13d. This progression over time shows the elements with the same colouring code for core, border and noise elements of a cluster, and the bounding area of the cluster. The aim of the management of evolving areas of interest is to provide an identifier to the evolving clusters over time, due to the fact that *DBSCAN* process the elements in an arbitrary order, the cluster identifier is not always the same.

In this thesis the evolution of the clusters represents how the traffic congestion changes over time. *DBSCAN* algorithm detects clusters for a single instant, and it is non deterministic in the identification of each detected cluster, therefore a further process is needed to track such evolution. The main purpose of tracking changes in clusters is in the de-

scriptions of specific *Traffic Congestion Areas of Interest* instances, for example what is the longest period of a traffic congestion and to refer to specific instances and its context for further analysis, but it should not influence the mining algorithm result as a general, as discussed in Section 4.7.

As mentioned before, there are algorithms to compute the evolution of a cluster, in this thesis MC1 algorithm [28] is used, it allows to detect and identify the clusters which are still valid in two consecutive instants, it means if a specific cluster exists and if its elements have not changed dramatically.

MC1 algorithm determines if a cluster is valid based on the elements within the cluster, if in two consecutive instants the individual elements have certain degree of similarity, therefore new elements or removed elements are allowed. MC1 algorithm is not position dependent, so that if a cluster moves in space but elements remain similar, the cluster will be tracked as one. This similarity between clusters is calculated by a integrity function, in [28] a Jaccard index is suggested to check similarity in, in which counts of intersecting and union elements are used to calculate similarity, such as $\frac{|c_i \cap c_{i+1}|}{|c_i \cup c_{i+1}|} \geq \theta$, where $0 \leq \theta \leq 1$. If the value of integrity is 1, it means that the cluster has exactly the same elements from the previous instant, a value of 0 means that none of the elements belongs to any cluster, therefore the cluster was not detected any more.

Listing 3 shows the pseudo code of MC1 algorithm, as presented in [28]. In short words, for each instant within a timeslice the clusters calculated by a clustering algorithm, in the case of this thesis *DBSCAN*, the MC1 algorithm checks the integrity for two consecutive instants. Integrity is defined as the ratio of the intersection of elements in both clusters and the union of the elements of both clusters.

```
1  G := null; // set of current clusters
2  for i:=1 to n // for each timestamp
3    for each current moving cluster g in G
4      g.extended := false
5    Gnext := null; // next set of current clusters
6    // retrieve timeslice clusters at Si
7    L := DBSCAN(Si, eps, MinP ts);
8    for each timeslice cluster c in L
9      assigned := false;
10     for each current moving cluster g in G
11       if g(c) is a vali doing a recapd moving cluster then
12         g.extended := true;
13         Gnext := Gnext union g(c);
14         assigned := true;
15       if (not assigned) then
16         Gnext := Gnext union c;
17     for each current moving cluster g in G
18       if (not g.extende doing a recapd) then
```

```
19          output g;
20    G := Gnext;
```

Listing 3: MC1 algorithm.

In Listing 3, line 11 checks if the integrity function is greater than the threshold to consider that g evolved into c and will be included in the Gnext set of valid clusters in line 13. In line 16 falls into the case of a new cluster which is integrated as moving cluster. Eventually when a cluster is no longer valid, it is not considered in the next iteration and is taken out as shown in line 19. Finally in line 20 the set of valid clusters is ready to be processed in the following iteration.

The representation of an *Area of Interest* should consider the aspects mentioned in this Section, the arbitrary areas surrounding the traffic congestion and the valid time of the traffic congestion. The proposed *STDI* to represent an *Area of Interest* is not traffic specific and is defined as follows:

- **cluster id** an identification number, unique for each cluster along the valid time. This identifier is updated by the MC1 algorithm, keeping track of the evolution of each cluster.
- **cluster type** a string identifying the cluster nature, in this case it is set to *Traffic-Congestion*
- **start date and time** ISO format of the start date and time when a cluster was detected for the first time.
- **end date and time** ISO format of the end date and time when a cluster was tracked for the last time by the MC1 algorithm.
- **geometry** a polygon representing the bounding area of the cluster. It is created from the core elements detected by *DBSCAN* and the radius *eps* as offset, and smoothed between two circles with a common tangent.
- **list of elements within the cluster** as a reference of the identifiers of each element within the cluster at a specific instant.
- **description** is a humalgortithman readable text of the detected situation.
- **timestamp** is the end instant in UTC timestamp for machine processing purposes.
- **non-ST features** are such features providing further information, but not used in the clustering algorithm. These features are in most of the cases human readable

and can be used as part of the final description of the contextual information of the *Area of Interest*. For example, it can be possible to include road name, whether the sensor is located on the road or in an upstream or downstream exit.

- **measurements** are numeric values that are aggregated from all the elements of the cluster. In this thesis the *Traffic Parameters* are used as measurements of the defined cluster. Aggregations can be descriptive statistics, such as mean, max and min values, standard deviation, among others.

Listing 4 is a truncated example of a GeoJSON representation of a cluster *STDI*.

```
{
  "type": "FeatureCollection",
  "features": [
    {no dense enough
      "geometry": {
        "coordinates": [bounding are
          [
            [
              8.6629,
              50.2191
            ],
            ...
            // Truncated output
            ...
            [
              8.6629,
              50.2191
            ]
          ]
        ],
        "type": "Polygon"
      },
      "type": "Feature",
      "properties": {
        "date_time_start": "2016-11-30T07:48+01:00",
        "Road": [
          "A5",
          "B455"
        ],
        //This description is a cumulative human readable description.
        "description": "Cluster evolved. Area with 9 sensors reporting
        traffic congestion. Roads: [A5, B455]. OnRoad: [OnMainRoad].
        HeadwayTime average in cluster:6.39760460916938.
        Speed average in cluster:17.91330183173179.
        HeadwayDistance average in cluster:25.433696247541306.
        Quantity average in cluster:2366.6666666666665.
        Density average in cluster:128.79666912707697. ",
```

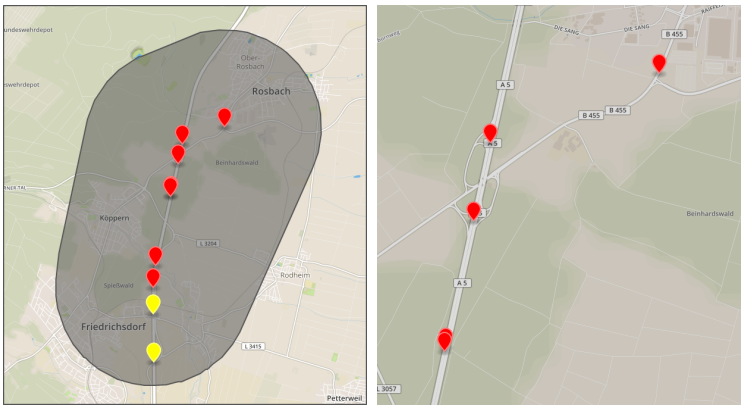
```
        "date_time_end": "2016-11-30T07:50:00+01:00",
        "cluster_type": "TrafficCongestion",
        "avgDensity": 128.79666912707697,
        "avgSpeed": 17.91330183173179,
        "avgQuantity": 2366.6666666666665,
        "cluster_id": 1480488540000001,
        "avgHeadwayTime": 6.39760460916938,
        "avgHeadwayDistance": 25.433696247541306,
        "cluster_elements": [
          "R2000908",
          "R2010598",
          "R2007859",
          "R2007877",
          "R2007853",
          "R2007871",
          "R2008265",
          "R2007843",
          "R2007865"
        ],
        "OnRoad": [
          "OnMainRoad"
        ],
        "timestamp": 1480488600000
      }
    }
  ]
}
```


Listing 4: GeoJSON Representation of a Cluster STDI.

The cluster described in Listing 4 is shown in Figure 14a, where the red elements represent the core points and the yellow ones represent the border, the gray area is the bounding polygon of the cluster, including the offset at *eps* distance. As part of the contribution of this thesis, the use of *Traffic Parameters* as part of a wider description of the cluster and therefore an extended context of the *Area of Interest*.

Additional information is used from the non-*ST* features, such as the name of the roads. In Listing 4 it is defined that nine sensors are part of the cluster, distributed along the roads A5 and B455, Figure 14b shows a zoom in of the area, where four sensors are along road A5 and one sensor is on road B455, within the defined distance *eps*. As an example of the measurements, the average of each *Traffic Parameter* is calculated. Listing 4 is in GeoJSON format for two main reasons, it is human readable and it is ready to display using map visualization services, such as GeoJSON web site[14].

[14]http://geojson.io



(a) Cluster overview. (b) Cluster zoom in.

Figure 14: Cluster with sensors in different roads.

Up to this point in the process, it is possible to detect *Areas of Interest* from the filtered *Traffic Records* with a state of *Traffic Congestion*, according to the traffic engineering field, a state reached when certain thresholds have been reached by specific *Traffic Parameters*. As areas are not limited by sensors being in the same road, *Areas of Interest* can be wider compared to what previous works have considered, therefore the context can include more elements. After the detection of *Areas of Interest*, the criteria used in this thesis to select an area for further study is explained in Section 4.5.

4.5. Selection of an Area of Interest for Further Analysis

This subsection describes the manual subprocess of selecting an area for further analysis, moreover the criteria used to select the geographical area within Hessen Federal State, from where the data is collected. Automation for the selection of an *Area of Interest* is out of the scope of this thesis, guidelines will be discussed in Section 6.2.

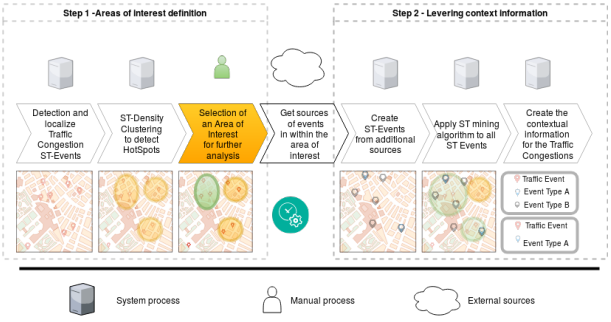


Figure 15: Selection of an Area of Interest for Further Analysis.

The criteria followed to select the *Area of Interest* for further analysis is as follows:

1. The area should concentrate sensors along the roads higher than the average.

2. Within the area there should be at least one important landmark, defined in this thesis as a facility, building, public or private well defined location, with a reference of people flow rate of at least 1000 persons at the time for a single purpose. This threshold has been taken from the definition of *mass gathering*[15].

3. Within the area there should be at least one motorway *(Autobahn, A)*, one 1st class road *(Bundesstraße, B)*, and one 2nd class road *(Landes- oder Staatsstraße, L or S)* defined in [17] where at least one sensor is located.

4. The area can be composed of more than one cell of the geohash grid with high number of sensors. Cells must be adjacent to each other.

The following assumptions are made as part of this thesis:

- The area of study is limited to the German Federal State of Hessen.

- Location of sensors is assumed to be exact, therefore deviations are ignored and corrupted data is excluded, such as sensors located outside Hessen.

- In order to simplify the area selection, a grid-based approach is taken, using geohash grid implemented in PostGIS[21]. Geohash length 5 is used[16].

These short guidelines are meant to provide the basis of the experiments and evaluations.

[15]https://en.wikipedia.org/wiki/Mass_gathering
[16]https://www.movable-type.co.uk/scripts/geohash.html

4.6. Creation of ST Events from Additional Sources

This section describes the subprocess in which the additional sources of events are transformed into a *STDIs*, which will represent the context of a *Traffic Congestion*. This subprocess is meant to prepare the additional data sources presented in Section 4.1.2 in a format that can be consumed by the mining algorithm presented in Section 4.7.

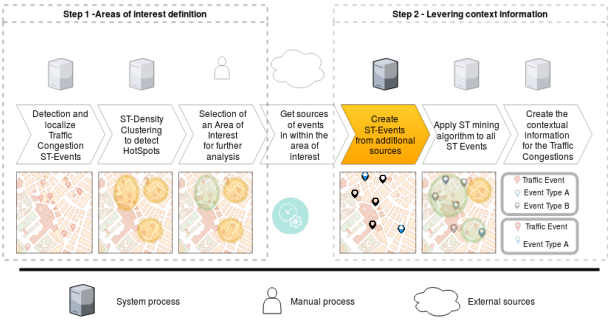


Figure 16: Creation of ST Events from Additional Sources.

In this thesis an *External Event* is such *STDI* representing something happening in specific time and space, it must not be related to road conditions, for example car accidents or road maintenance works are excluded for the purpose of this thesis. Matching with the *STDI* described for an *Area of Interest*, the *ST* features are expected to be the point coordinates representing the place where the event happened and a radius offset; a period represents the time duration of the event; a type of event, so that it can be representative for the context of *Traffic Congestion Area of Interest*.

As in the mining algorithm it is expected an overlap over time, the time dimension will consider also an offset for each event, on the premise that an external event official time is not always representative of the influence over a traffic congestion. This is shown in Figure 17.

(a) Before match. (b) During match. (c) After match.

Figure 17: Traffic Congestion influenced by External Event.

In a planned event, a football match as an example, it is assumed that people plan to arrive before the kick-off, which usually is the official time shown in advertisements, tickets, and also in TV transmissions. Figure 17a represents the time before the match starts, when people are going towards the stadium, in that case, it is expected a traffic congestion. It is shown in Figure 17b, ideally, when the match is going on, people are already at the stadium, therefore it is not expected a traffic congestion nearby the stadium related to the people attending the match. Finally, when the match is over, people will leave and there might be traffic congestions, this is shown in Figure 17c.

In a spontaneous event, for example if all the flights in a specific airport get delayed during a long period, the concentration of affected passengers can cause a breakdown in roads nearby. In this example it is assumed that no previous information is available, nor the quantity of people affected or the duration of delays, therefore no additional time offset is required. However, the space offset is required to relate the traffic congestions in the vicinity of the airport.

The representation of an *External Event* comprises a time-wise offset, which means additional interval before the start time and after the end time of the event, so that it overlaps with *Traffic Congestion Areas of Interest*. The representation of an *External Event* also incorporate a space-wise offset as an additional radius from the center where the event happens. In this thesis the variation of time and space offsets, or distances, is explored in Section 5.1 in order to test the closeness of the *External Events* and the *Traffic Congestion Areas of Interest*.

In this Section the *STDIs* creation was defined and will be used as part of the context of *Areas of Interest*, by applying a mining algorithm to detect relationships, explained in Section 4.7.

4.7. Levering Context Information on a Traffic Congestion Area: ST-Data Mining

After defining *Traffic Congestion Areas of Interest*, the second step considered in this thesis is levering context information, see Figure 18. It is defined as the process to include additional information from different sources to a *Traffic Congestion Area of Interest*, defined in Section 4.4 and narrowed by the guidelines in Section 4.5. This process aims to combine different types of events that occur close in space and time as a pattern. From a data analysis perspective, mining algorithms discover frequently appearing itemsets in transactions, defining a type of relationship among data. Frequent itemsets algorithms produce subsets of items sharing the same transaction, in *ST* those transactions can be defined, for instance, as spatial intersections and temporal overlaps operations. Figure 18 shows the last two subprocesses, the *ST* mining algorithm, where both traffic congested areas and other type of events relationships are established, and the eventual creation of the contextual information for the traffic congestions, which is the ultimate output.

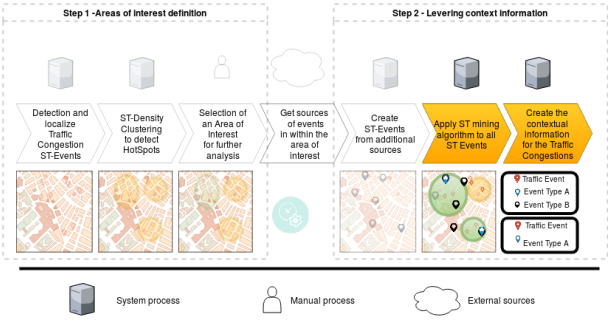
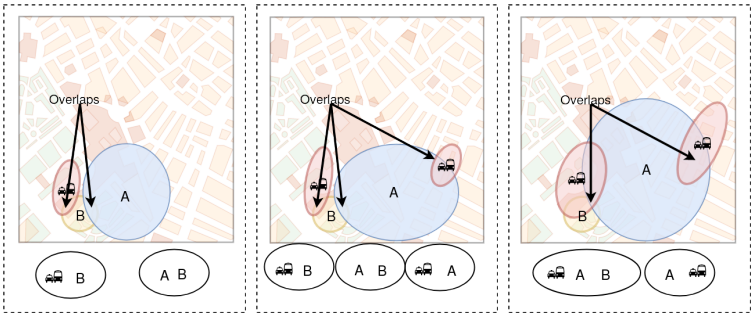


Figure 18: Mining the context information.

The principle of detecting a relationship between different type of events is shown in Figure 19. Each color represents a different type of event, "traffic congestion" in red, "A Event Type" in blue, and "B Event Type" in yellow. The ellipse represents the area where the event is happening, including the space offset mentioned in 4.6. When an ellipse intersects other, that can be interpreted as a match between two different types of events, therefore a relationship degree can be calculated.

Figure 19a is the first instant to be evaluated, the "Event Type A" intersects the "Event Type B" and the "traffic congestion" intersects only "Event Type B", however there is no relation between the "Event Type A" and the "traffic congestion". The relationships are shown in the bottom as tuples of event types.



(a) t=1 (b) t=2 (c) t=3

Figure 19: Principle of event relationship, ilustrated as polygon intersection.

In Figure 19b the second instant is evaluated, the "Event Type A" became larger in space and another "traffic congestion" was detected intersecting on the right hand side. Compared to the previous instant there is an additional relationship detected, "traffic congestion" is also related to "Event Type A".

Finally in Figure 19c a third instant is evaluated, the "Event Type" and "traffic congestion" on the left hand side became larger and now both intersect as well, that is detected as a new relationship between the three even types.

At each instant all the detected relationships degrees are evaluated and with a threshold it is possible to filter the ones appearing the most and with a larger intersection. That is interpreted as the context of a *Traffic Congestion Area of Interest*.

This principle was introduced in [43] as an ST mining algorithm, it finds and measures the so called co-occurring patterns among different type of events. In this algorithm, also in this thesis, each event is represented by a polygon changing over time, a co-ocurring pattern are those polygons overlaping in time and intersecting in space [43] evaluates the relationship between different type of events by a similarity measurement at each instant and a participation index, which indicates a specific relationship is over the entire interval of evaluation.

Let $E = e_1, ..., e_m$ be a set of ST event types and $I = i_1, ..., i_n$ a set of instances of those event types, such that $M \ll N$, definitions introduced in this algorithm are as follow:

- A ST co-occurrence is a subset of ST event types that occur in both space and time.
- A *size(k)* co-occurrence is denoted as $SE = e_1, .., e_k$ where $SE \subseteq E$, $SE \neq \emptyset$.
- A pattern instance is a single tuple of instances from I of event types specified in a *size(k)* co-occurrence pattern, the instances overlap in time and intersect in space.

- A collection of pattern instances is the subset of tuples containing instances from I from the entire data set.
- *cce* is an indicator of the strength of the ST relationship. In [43] an ST overlap is used, which is the Jaccard index of the intersection of the polygons. It is calculated by the ST intersection I_v over the ST union U_v. Figure 20 shows the ST intersection and union of green and red event types. The ST intersection and union are explored in a concrete example in further paragraphs.
- ST intersection I_v is defined as the volume resulting from the intersection of trajectories of all the instances of ST event types in a pattern instance. Figure 20 shows the ST intersection of green and red event types.
- ST union U_v is defined as the volume resulting from the union of trajectories of all the instances of ST event types in a pattern instance. Figure 20 shows the ST union of green and red event types.
- p is the prevalence measure, in [43] calculated as the prevalence, an average of the *cce*.

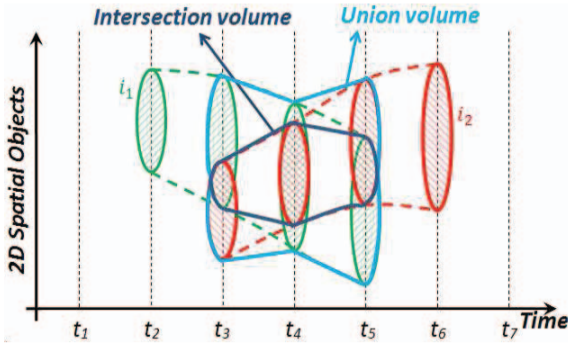


Figure 20: ST union and ST intersection example.

Listing 5 summarizes the co-occurrence algorithm proposed in [43], in which the user defined parameters are cce_{min} and p_{max} thresholds. Initialization of the algorithm is co-occurrence size $k = 1$, the initial candidate co-occurence patterns are the different event types $C_1 = E$, prevalent co-occurrence patterns will also derive from the initial $size - (1)$ event types, finally T_1 is the set of instances of co-occurrence pattern size 1, it uses the additional parameter t_s to represent the incremental time steps. After initialization, the iterative process goes from line 3 to 9, until there are no more patterns to be mined.

Step 4 generates the $size(k+1)$ ST co-occurring candidates, considering only the prevalent patterns, those which satisfied the participation threshold in the previous iteration. In

step 5 the table instances for $size(k+1)$ candidates and calculates the ST intersection and union across all time slots.

Step 6 selects from the table instances for $size(k+1)$ those in which the participation index is lower than the user defined threshold, to create the prevalent ST co-occurring patterns, that will be used in the following iteration. Steps 7 and 8 are the final result expected from the algorithm and the increment of the k value, respectively.

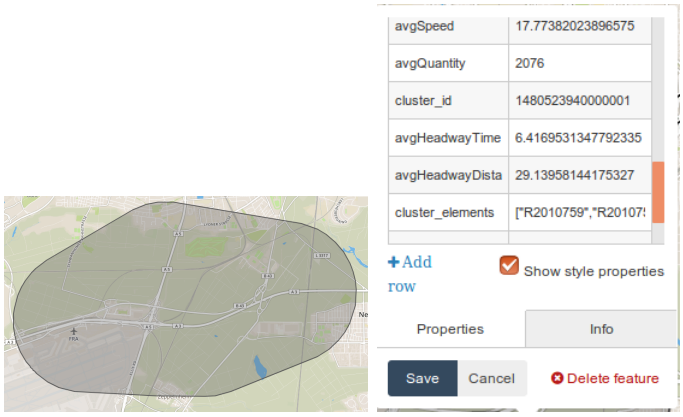
The prevalent ST co-occurrence pattern table, P_{final}, from step 10, represents all the relationships between event types which are strong enough and will be interpreted as the context of a *Traffic Congestion Area of Interest* which is the ultimate goal of the process and this thesis.

```
1  k=1, C1=E, P1 = E, Pfinal = empty;
2  T(1) = gen loc(C1, I, ts);
3  while (Pk is not empty) {
4    C(k+1) = gen candidate coocc(Pk);
5    T(k+1) = gen tab ins coocc(C(k+1), cceth);
6    P(k+1) = pre prune coocc(C(k+1), pith);
7    Pfinal = Pfinal union P(k+1);
8    k = k + 1;
9  }
10 return Pfinal;
```

Listing 5: Co-occurrence pseudo algorithm.

In this thesis, as explained in Section 4.1, the types of event are enumerated as follows: (i) *TrafficCongestion*, (ii) *FlightDelays*, and (iii) *FootballMatch*. Therefore $k_{max} = 3$ and co-occurrence candidates are as follows:

$k = 1$ (*TrafficCongestion*) Polygon calculated by the clustering algorithm explained in Section 4.4, it represents a *Traffic Congestion Area of Interest*.



(a) Area of *Traffic Congestion* (b) Some traffic parameters.

Figure 21: Event type *TrafficCongestion*.

$k = 1$ (*FlightDelays*) Polygon of a simulated external event, with center in an airport. It represents a period of time when in a specific airport flight delays where reported. Figure 22 shows Frankfurt Airport polygon, using a space offset of 5 Km.

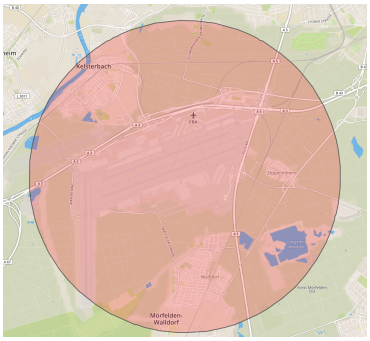


Figure 22: Event type *TrafficCongestion*.

$k = 1$ (*FootballMatch*) Polygon of a simulated external event, with center in a stadium. It represents a period of time of a football match. Figure 23 shows Commerzbank-Arena polygon, using a space offset of 5 Km.

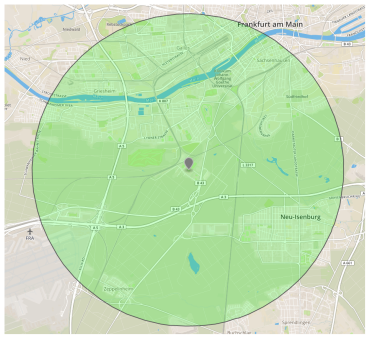


Figure 23: Event type *TrafficCongestion*.

$k = 2$ (*TrafficCongestion*,*FlightDelays*) Co-occurrence pattern candidate of a *Traffic Congestion* happening at the same time of *Flight Delays* at the airport. Therefore the context of the traffic congestion is the flight delays. Figure 24 shows this co-occurrence pattern.

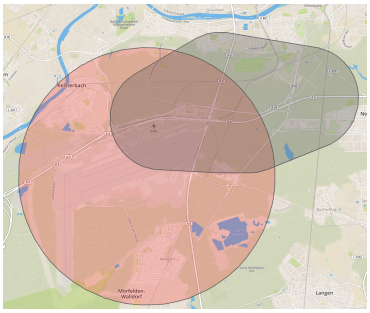


Figure 24: Co-occurrence pattern *TrafficCongestion* and *FlightDelays*.

$k = 2$ (*TrafficCongestion*,*FootballMatch*) Co-occurrence pattern candidate of a *Traffic Congestion* happening close to the stadium, assuming it can happen before or after a *Football Match*, due to the people attending the event. Therefore the context of the traffic congestion is the football match. Figure 25 shows this co-occurrence pattern.

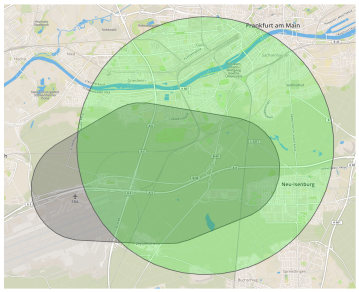


Figure 25: Co-occurrence pattern *TrafficCongestion* and *FootballMatch*.

$k = 3$ (*TrafficCongestion*,*FlightDelays*,*FootballMatch*) Co-occurrence pattern candidate of a *Traffic Congestion* happening close to the stadium and the airport, assuming it can happen before or after a *Football Match* and during *Flight Delays*. Therefore the context of the traffic congestion is the football match and flight delays. Figure 26 shows this co-occurrence pattern.

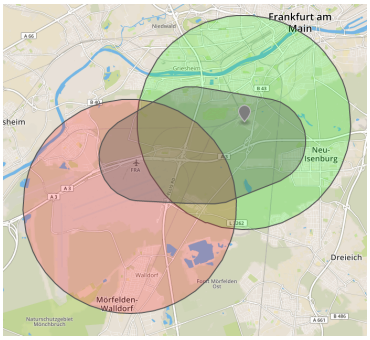


Figure 26: Co-occurrence pattern *TrafficCongestion*, *FlightDelays* and *FootballMatch*.

In the example from Figure 26, using both *External Event Types*, for a single instant the *Co-occurrence Pattern Candidates* are evaluated to get the *cce*. The results are presented in Table 5. In Table 5 *TrafficCongestion*, *FlightDelays* and *FootballMatch* are referred as e_1,e_2, and e_3, respectively.

Co-occurrence Pattern	**Union Area** U_a	**Intersection Area** I_a	$cce = \frac{I_a}{U_a}$
e_1, e_2	0.01221	0.00279	0.22861
e_1, e_3	0.01078	0.00423	**0.39287**

e_1, e_2, e_3	0.01752	0.00187	0.10681

Table 5: Co-occurrence example for *Traffic Congestion*, *Flight Delays* and *Football Match* events.

Table 5 shows an evaluation of $size(2)$ and $size(3)$ *Co-occurrence Pattern Candidates*, excluding the candidate of $size(2)$ with no *TrafficCongestion* event type.

Figure 26 shows that the *Traffic Congestion Area* is located on the right hand side of the *FlightDelays* airport and, visually, there is more *Intersection Area* with the *FootballMatch* stadium location. This is confirmed by the computation of *Intersection Area*, *Union Area*, and *cce* for each *Co-occurrence Pattern Candidate*.

The interpretation of Table 5 is as follows. The specific *Traffic Congestion Area* has a stronger relationship with a *Football Match* than with *Flight Delays* at the airport, and stronger than the *cce* of both *External Events*. If the highest value of *cce* is selected as a näive approach, the *Contextual Information of the Traffic Congestion is a Football Match*.

To clarify, Union and Intersection areas in Table 5 are unit-less as part of the *WGS84* projection, however, *cce* is a proportion and remains the same as if union and intersection areas would be transformed into area units, such as square meters or kilometers. All the numeric values are truncated in the 5th decimal.

As either the parameter of the space offset of an *External Event* or the *eps* parameter in a *Traffic Congestion* clustering algorithm increase, the *cce* of an *ST* overlap will increase, therefore the *ST* relationship between types of events is stronger. This change of parameters can reflect what Tobler's first law of geography states, *"Everything is related to everything else, but near things are more related than distant things"*. In Section 5 changes in this parameters are explored and further discussed in Section 6.

This Section is the final part of the process followed from to explore the use of the *ST* approach, including space and time dimensions, algorithms, and transformations for the *STDTs*. Section 4.8 presents the process from a *Big Data* perspective, which assumptions are done, type of frameworks that will be preferred, and, finally, the architecture of the system.

4.8. Big Data Perspective: The Scalability Issues

Big data is a widely accepted concept, coined by Roger Miugalas in 2005 from O'Reilly Media, to describe a set of data that is almost impossible to process using traditional business intelligence tools [49, 39], due to the size and complexity. According to Gartner, a well-known research and advisory firm in the Information Technology (IT) industry among other areas, *big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation*[17].

It has been already expressed in [1] that due to the complexity to abstract geographical problems, in addition to the temporal dimension and size of data sets, *ST* problems can be considered as *Big Data*, despite the lack of specific thresholds for volume, velocity and variety.

As described in previous sections, this thesis meets the three aspects of *Big Data*.Variety is covered by the process explained in Section 4.1, specifically on the subprocess of *ST Data Mining* in Section 4.7, in which different sources of data are processed. The potential to include further data sources remarks the aspect of variety of data.

[17]https://www.gartner.com/it-glossary/big-data/

Regarding volume, the opinion of Pietsch in [42], referring the explanation given in [24] for the quantity of samples to be processed and have a result that represents or at least approximates what was needed in a digital image processing algorithm, it is not clear what is the threshold in samples needed to be considered big data, nor the number of features, parameters, or conditions. It is clear the conclusion, the more amount of samples, the better the algorithms can get to a more accurate model.

Velocity is seen as the rate of processed elements and is probably the least clear with the data sources used in this thesis. As explained in Section 4.1.1, the ingestion rate of the *QnV Records* is about 16600 per minute, creating around 2500 *Traffic Records* per minute, only from sensors located in highways within the Federal State of Hessen. Compared to what popular world-wide social network and internet platforms, used in some *Big Data Frameworks* examples[18], the rate is not as high, as shown in Table 6. Scaling the sources for *Traffic Congestions* or area covered will increase the rate and therefore processing speed gets important.

Table 6[19] considers as an entire message one atomic information piece, to be consumed or produced within the platform or service, size of the messages are ignored.

Internet Service	**Description**	**Messages rate**
QnV Records	QnV XML parsed into QnV Record	16,600/Minute
Traffic Records Records	Processed QnV Record according to Listing 1	2,500/Minute
Twitter	Tweet, a JSON document posted by a registered user	8,167/Second
Instagram	Instagram Photo, multimedia and text posted by a registered user	865/Second
YouTube	Video viewed by any used with access to the platform	75,223/Second

Table 6: Estimated popular Internet services message rate, comparison with QnV Records and Traffic Records. Information obtained in *Internet Live Stats*

Scaling a system can be done in different ways, however there are principles that can help to organize and to scale up using some platforms, frameworks, and principles. To list

[18]https://hortonworks.com/tutorial/sentiment-analysis-with-apache-spark/
[19]http://www.internetlivestats.com/one-second/ Retrieved on 27th August 2018.

the principles, models and architectures towards scaling the amount of sources and the number of samples as follows:

- **Distributed and parallel tasks** This principle meant to reduce the overhead and contention in CPU and memory access by splitting the amount of data to be processed among different and independent sets of resources. One architecture that covers this principle is the so called **Master-Worker**, used in the model of **Map-Reduce**. Figure 27 shows the **Master-Worker** architecture and the **Map-Reduce** model. A *Master-Node* is a device or process controlling and coordinating the distribution of the tasks among one or more other devices or processes called *Worker- or Slave-Nodes*. Each *Worker- or Slave-Node* has the responsibility to finish the assigned task, usually a subset of operations or a subset of data. In **Map-Reduce** model, *Worker-Nodes* parallelize task of *Map* or transform each datum, and *Reduce* or gather a single result. In the *Big Data* perspective, this principle copes with data volume increase.

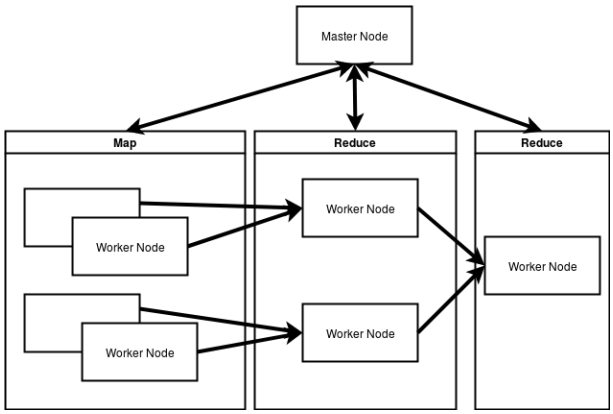


Figure 27: Master-Worker architecture and Map-Reduce model.

- **Independent sub-systems** This principle meant to reduce the complexity of the process done in each node, it can be done in different levels, from a capability to an entire set of transformations. In this thesis the level of independence is each step in the process explained in Section 4.2, so that there is a set of nodes for each algorithm, one for the clustering and one for the mining. This is shown in Figure 28, as a complementary information of the Figure 27. In the *Big Data* perspective, this principle copes with data volume the increase in combination with the distribution and parallelization of tasks.

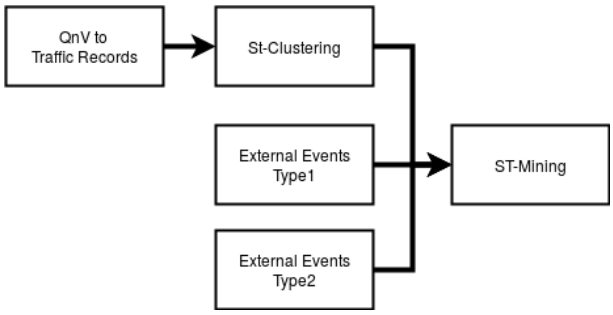


Figure 28: Sub-systems architecture.

- **Fast and flexible communication channel for multiple producers and consumers** This principle meant to provide a flexible communications channel between sub-systems, in terms that information can be consumed by different types of nodes, for example the same data is consumed by different algorithms, or producDistributeed by different types of nodes into the same type of nodes, for example new event data sources are pushed into the same co-occurrence algorithm. The architecture that covers this principle is **Publisher/Subscriber**, as shown in Figure 29. In the *Big Data* perspective, this principle copes with data volume and variety increase, and avoiding message delays.

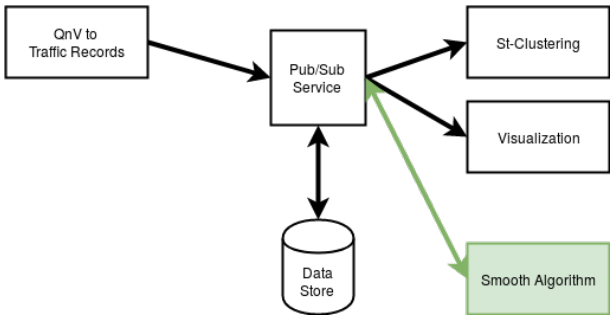
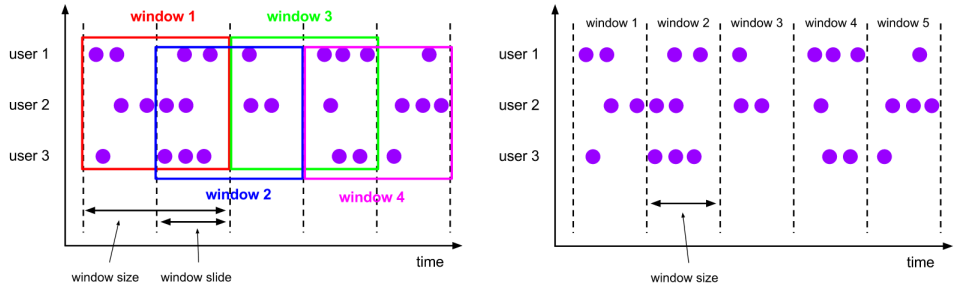


Figure 29: Publisher/Subscriber architecture.

- **Real-time processing** This principle meant to provide faster results, ideally looking for the shortest delay from ingestion time. The **Data Flow** [2] model covers this principle, instead of doing the computation for an entire data set, data is split over time in windows and processed incrementally. Figure 30 shows the two main window types to process data. Figure 30 was taken from Flink Window Description



(a) Thumbling Window. (b) Sliding Window.

Figure 30: Data Flow Window model.

Web Page[20], because Flink is the main Stream Processing Framework used in the implementation.

Following the principles listed before, Figure 31 shows the design of the entire system that implements the process presented in Section 4.2.

[20]https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/operators/windows.html

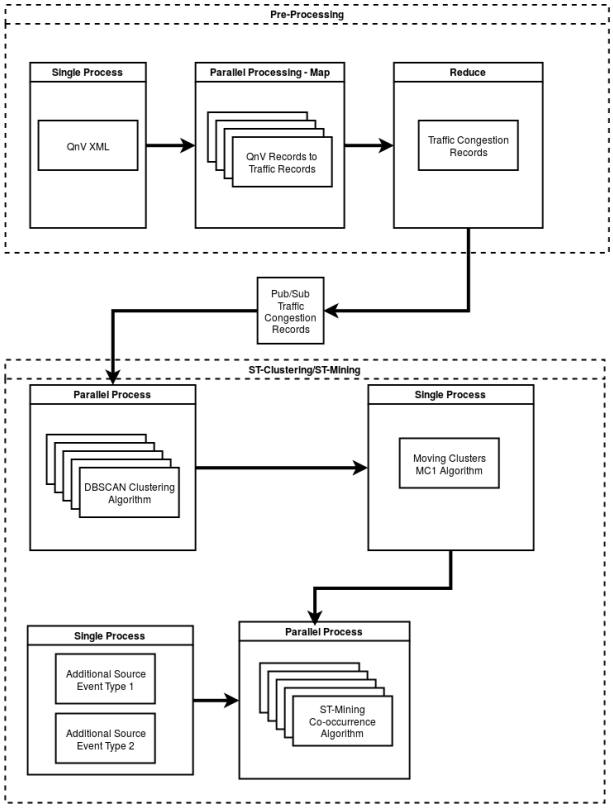


Figure 31: System Architecture.

Giving priority to the functionality of the system, the reference design of the system does not fully cover the *Big Data* perspective principles, this is discussed in Section 6.2, as part of the future work.

5. Evaluation

The proposed experiments and evaluations are divided into two parts, following the two steps shown in Figure 32. The first experiment corresponds to the detection of *Traffic Congestion Areas*. The second experiment focusses on the evaluation of the process of *Levering Context Information*. Each experiment will focus in one of the wrapping steps, rather than individual subprocesses.

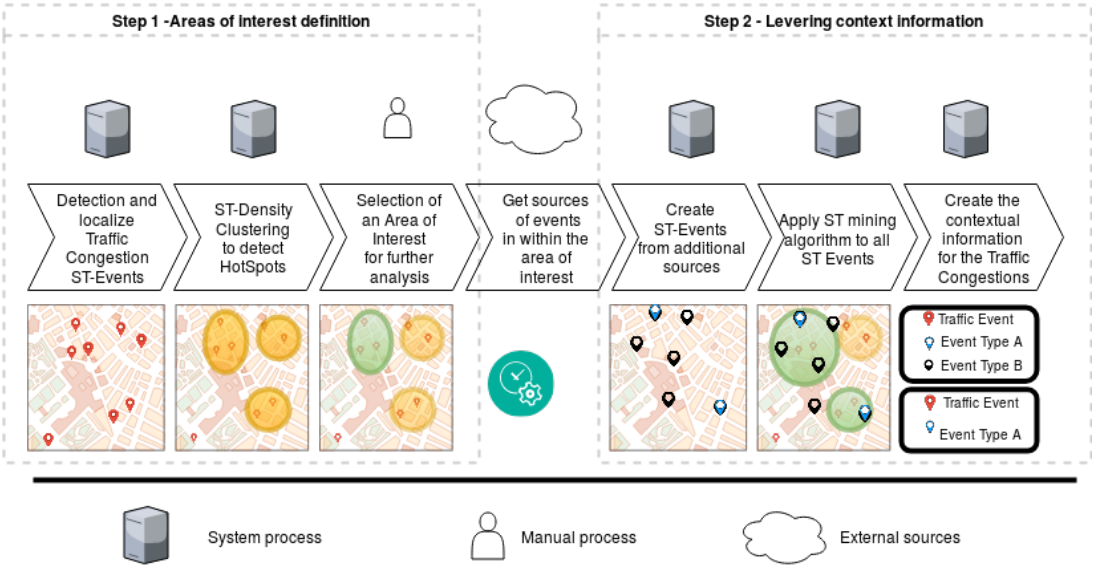


Figure 32: Process overview.

In Section 5.1 the experiment of detecting *Traffic Congestion Areas* is explained and discussed. The main objective is to determine parameters of the main clustering algorithm and use those parameters in the next evaluation. In Section 5.2 the experiment of *Levering Context Information* is described and discussed. The main objective is to use the generated *Traffic Congestion Areas* and provide a context with the mining algortithm, evaluating distance and time closeness.

5.1. Evaluation of Detection of Traffic Congestion Areas

In this section the evaluation of the *Traffic Congestion Areas Detection*, which is mapped into the first step of the general process, explained in Section 4.2 and detailed in Section 4.4. Figure 33 shows the subprocesses to be evaluated in this section.

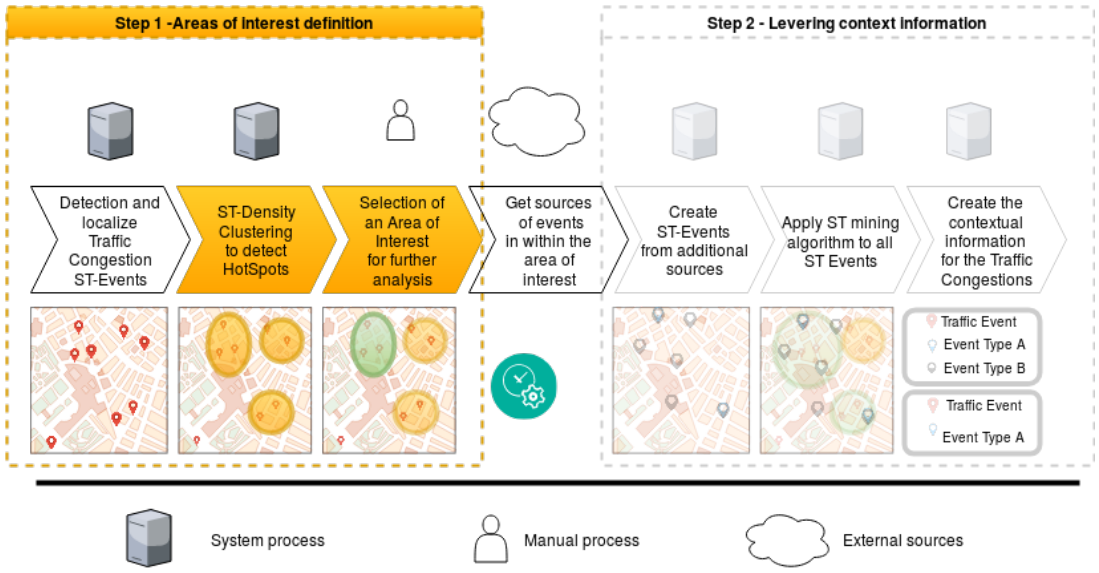


Figure 33: Evaluation of *Traffic Congestion Areas* detection.

The proposed experiment focused on the *Traffic Congestion Areas of Interest* detection is suggested as qualitative, using different values of the parameters *eps* and *MinPts* of the the *ST* clustering algorithm over the instant of maximum number of *Traffic Congestion Events*, one with the average number of *Traffic Congestion Events*, and one instant with 95-percentile number of *Traffic Congestion Events*. The detected areas are evaluated according to the following criteria:

1. *Traffic Congestion Events* within cluster or *Traffic Congestion Area of Interest* can be located on the main road, upstream or downstream exits, any direction, and in different roads. The use of areas instead of single points or sections of a road is the main contribution in this thesis compared to previous works, as discussed in Section 3.

2. For the maximum number of sensors reporting *Traffic Congestion* at one instant within the evaluation period, there must be more than one *Traffic Congestion Area*. This is recommended in [44] as an initial evaluation for the parameters.

3. For the average and the 95-percentile number of sensors reporting *Traffic Congestion* at one instant within the evaluation period, there must be detected at least one *Traffic Congestion Areas*. This is an extension of the previous criteria, suggested in this thesis.

The evaluation is done at first with $MinPts = 4$, suggested in [44] as the general rule for the minimum neighbors parameter, further testing is done decreasing the parameter $MinPts = 3$ and $MinPts = 2$, as the lower number of neighbors, the higher the probability to detect a cluster, due to the fact that it requieres less neighbors to label one element as core.

The evaluation is done at first with $eps = 1500$ [meters], this two distances have been selected according to Table 7, where the average coeficient of variation is lower. Further testing is done increasing the parameter $eps = 2500$ [meters], as the wider distance to detect neighbors, the more probability to detect a cluster.

This criteria is established from a non expert point of view on traffic engineering, but rather as minimal features to test the system and the methodology proposed in this thesis. The criteria is applied for each test group, in which *eps* and *MinPts* parameters are the same, so that the parameters are accepted or discarded for all the test group. If any of the elements in the test group do not pass the criteria, the parameters are discarded.

This experimentation is not done as quantitative, in the sense of measuring the *quality* of the clustering algorithm as suggested in [25, 11], due to the following particular conditions of the sensor data set:

- The cluster algorithm is applied to a subset of filtered locations, therefore it is not clear whether it is a data set with clustering tendency. This implies that (i) all the points that naturally belong to the same cluster will eventually be attached to it by the algorithm, (ii) no additional data set points (i.e., outliers or points of another cluster) will be attached to the cluster.

- The density distribution varies over time. Some subsets are of one element.

- The nature of the use case requires expert advise from the traffic engineering field, as suggested in [44].

Table 7 is calculated with the distance between a pair of sensors from the sensors defined in *QnV Location* file, considering if both are in the same road, both are on the main road or at least one is in an exit, and if both are in the same direction. The coefficient of variation is defined as a measure of dispersion of the data set and is calculated $c_v = \frac{stddev(\sigma)}{mean(\mu)}$. The lower the c_v the less variability on the set, as the values tend to be more similar. In this thesis 1500 and 2500 meters have been used as the *eps* test values, as those are the two lowest c_v for sensors in the same road, both on main road, and both in same direction, assuming that a traffic congestion can be better detected under those conditions. The reader should focus on the last column, as it contains the coefficient of variation, the lower the better, it means the data set does not vary too much.

Same Road	**Both on Main Road**	**Same Direction**	**Number of Pairs**	**Average Distance**	**Std Dev Distance**	c_v
5000 m						
t	t	t	7106	2652.84	1835235.96	0.51
10000 m						
t	t	t	13760	4975.29	7712610.70	0.55
2500 m						
t	t	t	3278	1397.09	424712.59	0.46
1500 m						
t	t	t	1808	900.03	155760.80	0.43

Table 7: Coefficient of Variation for different distances between two sensors.

For this experiment, it is used a data set collected between November 25th 2016 and December 9th 2016, two weeks period. In rough numbers, the data set comprises 22.000 *QnV XML* files, eparsed into 368.520.000 *QnV Records*, which are transformed into 49.680.000 *Traffic Records*, and finally 71.214 filtered *Traffic Congestion Events*. The chart in Figure 34 shows the count of *Traffic Congestion Events* each instant when there was detected at least one.

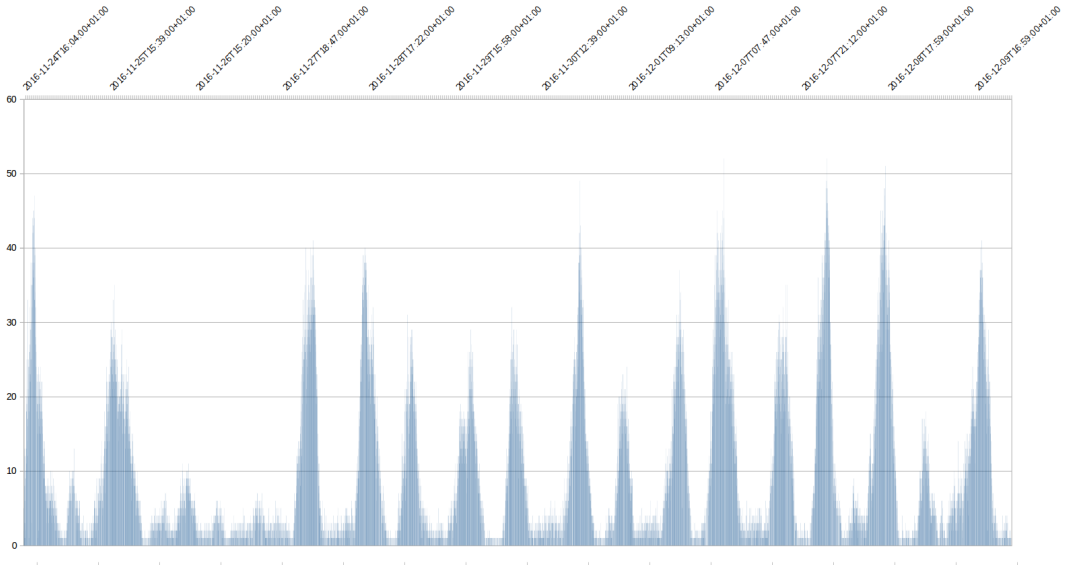


Figure 34: Traffic Congestion Events count over time.

From the processed data set, the following results are obtained and the test group of subsets is defined by the indicated instants:

- Maximum number of *Traffic Congestion Events*: 52, detected on 2016-12-07T08:45:00+01:00.
- Average number of *Traffic Congestion Events*: 9, detected on 178 instants. Instant 2016-12-07T06:37:00+01:00 will be evaluated.
- Standard Deviation of *Traffic Congestion Events*: 10, therefore the 95-percentile is 39 elements. 95-percentile was detected on 31 instants. Instant 2016-12-07T07:35:00+01:00 will be evaluated.

Table 8 shows the results of the change of parameters for the clustering algorithm on the test group, as suggested previously, including total number of clusters detected.

Subset	***eps*** **mts.**	***MinPts***	***Subset size***	***Num Clusters***	***Clustered elements***	***Criteria Acceptance***
Max.	1500	4	52	1	6	No
Avg.	1500	4	9	0	0	No
95-Perc.	1500	4	39	0	0	No
Max.	2500	4	52	1	10	No
Avg.	2500	4	9	1	7	Yes

95-Perc.	2500	4	39	0	0	No
Max.	1500	3	52	1	7	No
Avg.	1500	3	9	1	5	Yes
95-Perc.	1500	3	39	1	4	Yes
Max.	2500	3	52	2	15	Yes
Avg.	2500	3	9	1	7	Yes
95-Perc.	2500	3	39	2	8	Yes
Max.	1500	2	52	4	19	Yes
Avg.	1500	2	9	1	7	Yes
95-Perc.	1500	2	39	2	7	Yes
Max.	2500	2	52	3	22	Yes
Avg.	2500	2	9	1	7	Yes
95-Perc.	2500	2	39	5	18	Yes

Table 8: Evaluation of different *eps* and *MinPts* parameters in the test group.

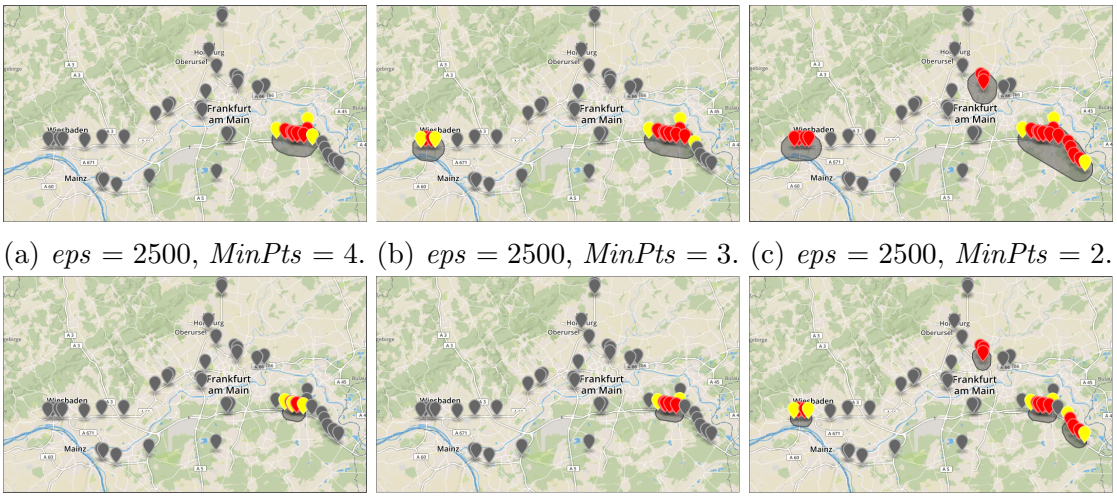
From the Table 8 results, a qualitative evaluation is done in order to select the *eps* and *MinPts* parameters and proceed with the experiment and evaluation of the *Levering of Context Information* for the detected *Traffic Congestion Areas*. The evaluation discussed by each subset with different parameters, so that the difference is more clear.

For *Traffic Congestion Area* with 52 *Traffic Congestion Events*, detected on 2016-12-07T08:45:00+01:00, identify as Max. in Table 8 resulting clusters are shown in Figure 35, which is a partial screenshot, zoomed in the area where the clusters were detected.

As the maximum number of elements to be clustered and using the parameter of *MinPts*=4, suggested in [44] as the initial value, it was expected to have more than one cluster, however as shown in Figures 35d and 35e, it seems there might be more areas to be considered as *Traffic Congestion Areas*.

As expected by decreasing the *MinPts* to 3, at a distance of 2500 for the *eps* parameter there are two clusters, as shown in Figure 35b, however in Figure 35e with a *eps* = 1500 only one cluster was detected.

Finally, with the *MinPts* = 2, as the minimum value explored in this thesis, results are as shown in Figure 35c for *eps* = 2500 and in Figure 35f for *eps* = 1500. Both results are more of what is expected for the maximum number of *Traffic Congestion Events* with three and four clusters, respectively. The main difference is at the bottom-right, where a cluster is devided into two for the shorter *eps* value.



(a) *eps* = 2500, *MinPts* = 4. (b) *eps* = 2500, *MinPts* = 3. (c) *eps* = 2500, *MinPts* = 2.

(d) *eps* = 1500, *MinPts* = 4. (e) *eps* = 1500, *MinPts* = 3. (f) *eps* = 1500, *MinPts* = 2.

Figure 35: Change of *eps* and *MinPts* parameters for maximum number of *Traffic Congestion Events*.

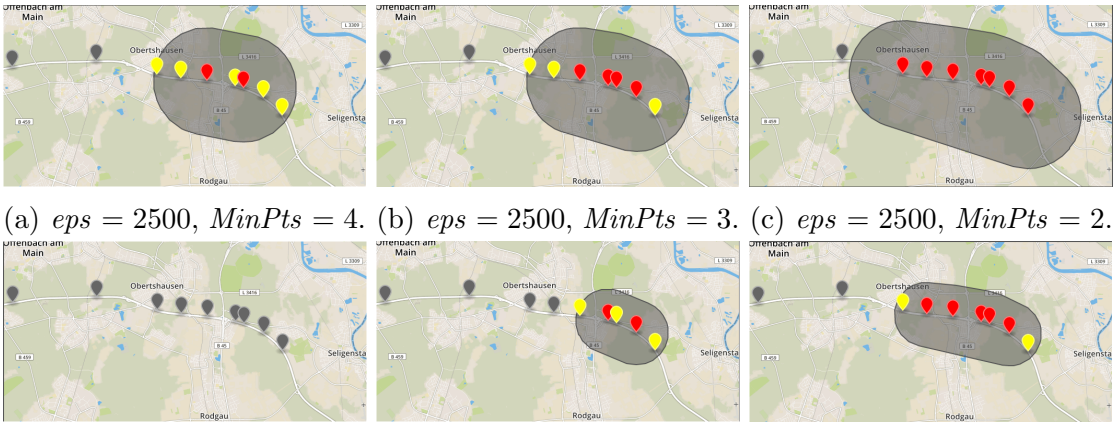
Comclusion is that *MinPts* = 4 is fully discarded, as suggested in [44], avoiding that one single cluster contains all the clustered elementsas there is only one cluster in some combination with *eps*. For better selection on parameters, the Average and 95-Percentile subsets are discussed.

For *Traffic Congestion Area* with 9 *Traffic Congestion Events*, detected on 2016-12-07T06:37:00+01:00, identify as Avg. in Table 8, resulting clusters are shown in Figure 36, which is a partial screenshot, zoomed in in the area where the clusters were detected.

As the average number of *Traffic Congestion Events* to be clustered is expected at least one cluster or *Traffic Congestion Area*, however there was a risk at picking the sample, as 9 sensors represent only the 0.3 % of the total number of sensors. As shown in Figure 36, for the selected instant all the sensors are *nearby*, therefore is a good candidate to do the evaluation of the *eps* and *MinPts* variation.

For *MinPts* = 4, it is not detected any cluster with *eps* = 1500 as shown in Figure 36d, however with *eps* = 2500 one cluster is detected with two Core elements in red and five Border elements in yellow, shown in Figure 36a.

For *MinPts* = 3 in both distances a cluster is detected, in case of *eps* = 2500 includes 7 elements, while *eps* = 1500 only 5 elements are part of the cluster, shown in Figures 36e and 36b, respectively.



(a) *eps* = 2500, *MinPts* = 4. (b) *eps* = 2500, *MinPts* = 3. (c) *eps* = 2500, *MinPts* = 2.

(d) *eps* = 1500, *MinPts* = 4. (e) *eps* = 1500, *MinPts* = 3. (f) *eps* = 1500, *MinPts* = 2.

Figure 36: Change of *eps* and *MinPts* parameters for average number of *Traffic Congestion Events*.

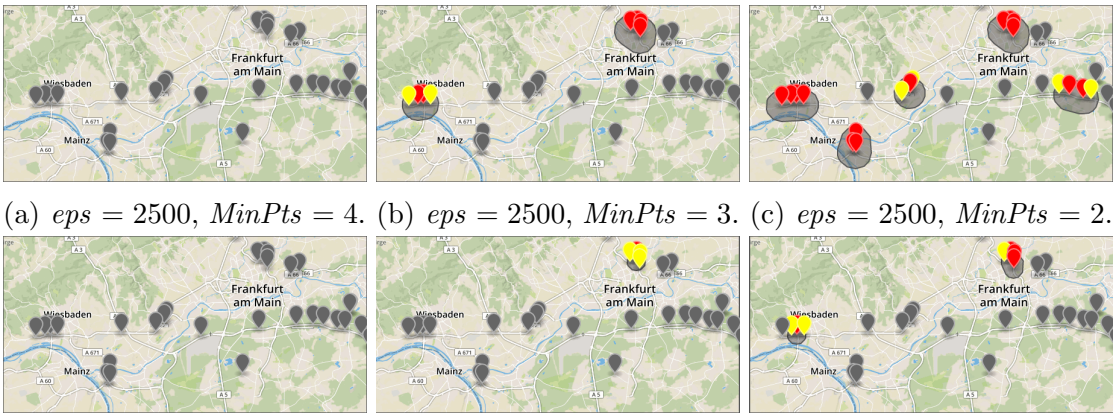
Finally with *MinPts* = 2 one cluster is detected for both distances, with seven elements, exactly the same elements, but with the difference of having two border elements with *eps* = 1500, as shown in Figure 36f, while with *eps* = 2500 all elements are labeled as core, shown in Figure 36c.

The fact to discuss in this subset evaluation is the number of elements detected for *MinPts* = 2 and *eps* = 2500 and 1500, *MinPts* = 3 and 4 and *eps* = 1500, in all those cases the elements is the same, the only difference is the label of some elements. In this case *MinPts* = 2 would be prefered as the parameter of selection, because *MinPts* 3 and 4 consider more elements as border. However, the last evaluation subset is discussed to define the final parameters used to evaluate *Levering Context Information* in Section 5.2.

For *Traffic Congestion Area* with 39 *Traffic Congestion Events*, detected on 2016-12-07T07:35:00+01:00, identify as 95-Perc. in Table 8 and resulting clusters are shown in Figure 37, which is a partial screenshot, zoomed in in the area where the clusters were detected.

For this subset the number of clusters detected is expected to be similar to the Max. subset, as the number of *Traffic Congestion Events* to be clustered represents 95% of the maximum.

For both distances with *MinPts* = 4 no cluster were detected, as seen in Figure 37d and 37a, it confirms the previous evaluations, and *MinPts* is completely excluded as valid parameter for the evaluation of the *Levering Context Information* in Section 5.2.



(a) $eps = 2500$, $MinPts = 4$. (b) $eps = 2500$, $MinPts = 3$. (c) $eps = 2500$, $MinPts = 2$.

(d) $eps = 1500$, $MinPts = 4$. (e) $eps = 1500$, $MinPts = 3$. (f) $eps = 1500$, $MinPts = 2$.

Figure 37: Change of *eps* and *MinPts* parameters for 95-Percentil number of *Traffic Congestion Events*.

In the case of the evaluation of $MinPts = 3$, in combination with $eps = 1500$ only one cluster was detected, as shown in Figure 37e, however only 4 out of the 39 elements were included into that cluster. For the same value of $MinPts = 3$ in combination with $eps = 2500$ two clusters were detected, as shown in Figure 37b, in this case the number of elements within a cluster increased to eight, represents 20% of the 39 elements, compared to the Max. subset with the same parameters the clustered elements are 28% of the 52 total elements.

Finally with $MinPts = 2$, the lowest value tested in this thesis, in combination with $eps = 1500$ only two clusters were detected, seven out of the 39 elements were labeled as part of a cluster, shown in Figure 37f. While with $eps = 2500$, five clusters were detected, shown in Figure 37c, and 18 elements were labeled into one of them, it represents 46% of the total elements evaluated, higher than the 42% of the elements within a cluster for the same parameters evaluating the Max. subset.

From the tested values for *MinPts* and *eps* parameters and according to the criteria specified before, $MinPts = 4$ is excluded as candidate value for further evaluation, none of the criteria was positive in the selected subsets. For $MinPts = 3$, both subsets Max and 95-Perc did not pass the criteria using $eps = 1500$, with $eps = 2500$ the criteria matched. $MinPts = 2$ gave the results as expected in both distances for *eps*, therefore it is selected as the parameter to use in following section. In case of *eps*, the differentiator is the number of elements labeled as part of a cluster, which was higher in $eps = 2500$ for the Max. and 95-Perc. subsets. Therefore $MinPts = 2$ and $eps = 2500$ are used as values to evaluate *Levering Context Information* in Section 5.2.

Following the process explained in Section 4.2 and shown in Figure 33, the *Traffic Congestion Areas* detection is followed by the selection of an area of interest and further analysis of *Levering Context Information*. For such selection, guidelines mentioned in Section 4.5 mention that the area should include more than average number of sensors, additionally, landmarks where high consentration of peopoe exist.

Figure 38 shows the grid based in Geohash length 5 where at least one sensor is located, according to *QnV Location* file. Each grid cell is approximately 4.89 km x 4.89 km.

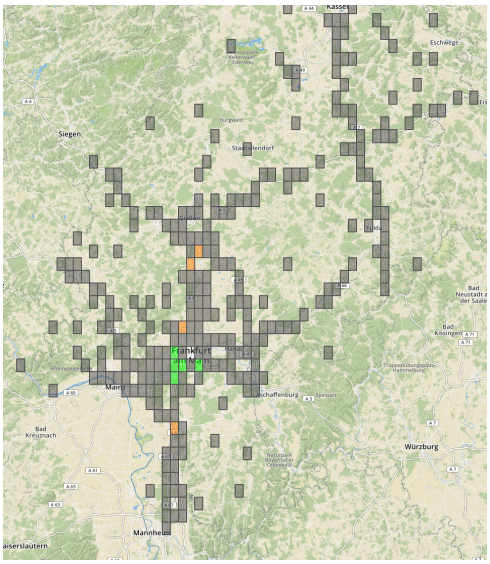


Figure 38: Sensor location in Hessen Federal State by geohash grid cell.

Guidelines in Section 4.5 require that the selected area has a concentration of sensors higher than the average, which for the *QnV Location* file is 7.5 sensors per grid-cell. Table 9 contains the top 10 grid-cells and the number of sensors within, from which the area should be selected. Each Geohash is in length 5 and can be displayed at *Movable Type Scripts* page [21].

Geohash	**Count of Sensors**
u0yj5	87
u0yhe	60
u0yhu	54
u0yhy	54

[21] https://www.movable-type.co.uk/scripts/geohash.html

u0yhg	53
u0yjs	34
u0ynv	32
u0ypn	31
u0y5e	31
u0yhd	30

Table 9: Top 10 sensor count by Geohash Grid-Cell.

Figure 39 is a closer view of the area where the first 5 grid-cellsfrom Table 9, colored in green are the Top 5 highest sensor concentration per grid-cells.

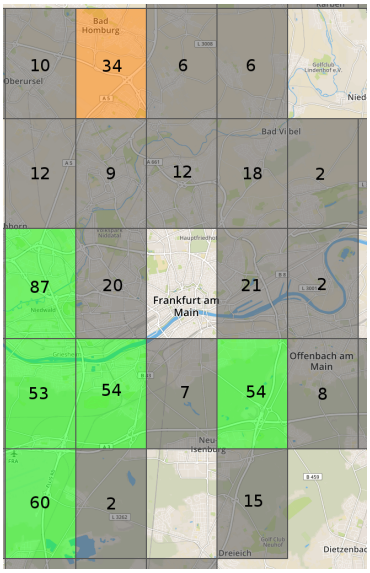


Figure 39: Sensor location in Frankfurt am Main Area by geohash grid cell.

Within the area shown in Figure 39 and following guidelines in Section 4.5, two main landmarks are Frankfurt Airport and Commerzbank Arena, where concentration of people can be over 1000. Frankfurt Airport had 60.8 Million passengers in 2016, with an average of 9254 passengers in an hour, as time of operation is from 5:00 to 23:00[22]. Commerzbank Arena capacity is 50,300 spectators for a Football Match[23].

[22]https://www.frankfurt-airport.com/en/travel/transfer.detail.suffix.html/article/b2b/airlines_tourism/airlines/facts-and-figures.html

[23]https://www.commerzbank-arena.de/english/the-arena/facts-and-figures

Finally, roads wirthin the area shown in Figure 39 are A5, A3, B43, B44, L3317, L3117, among others, matching with the requierement from Section 4.5.

Specifically, the area selected is defined by the bounding box with the top-left corner in *WGS84* coordinates at Longitude: 8.525390625, Latitude: 50.1416015625 and bottom-right corner coordinates at Longitude: 8.701171875, Latitude: 50.009765625. This bounding box is considering adjacent grid cells that do not contain sensors, this gives additional space to visualize and to evaluate the distance in the following section. Figure 40 shows the map and the area for further analysis. The black border represents the bounding box, the red marker represents Frankfurt Airport and the green marker represents the Commerzbank Arena.

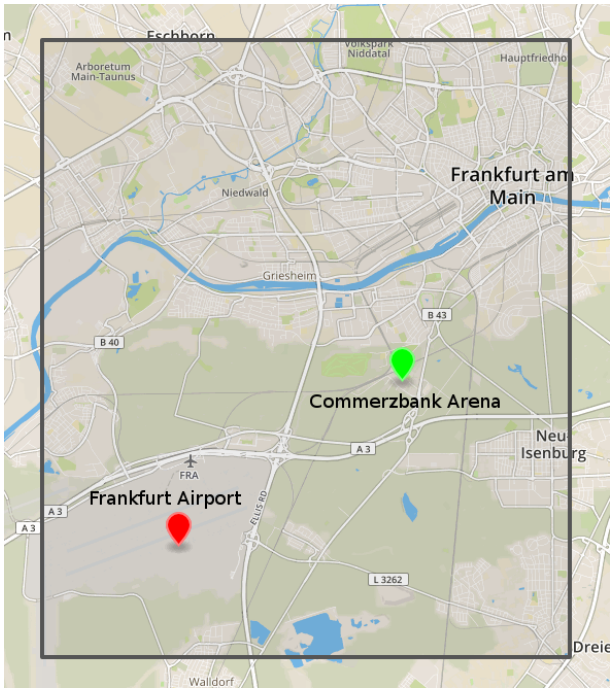


Figure 40: Area for further analysis. Frankfurt Airport and Commerzbank Arena.

With the *Traffic Congestion Area of Interest* and the values for *eps* and *MinPts* parameters defined in this section, the evaluation of *Levering Context Information* is described and discussed in Section 5.2.

5.2. Evaluation of Levering Context Information

In this section the evaluation of the *Levering Context Information*, which is mapped into the second step of the general process, explained in Section 4.2 and detailed in Section 4.7. Figure 41 shows the subprocesses to be evaluated in this section.

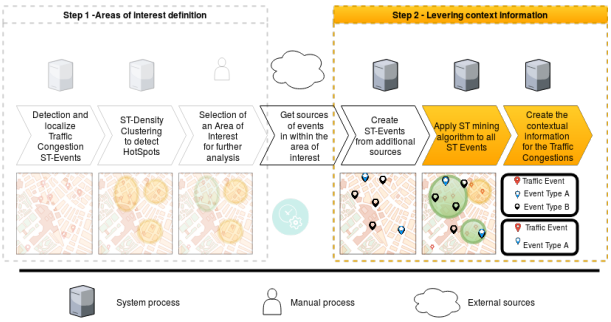


Figure 41: Evaluation of *Levering Context Information*.

The proposed experiment to evaluate the *Levering Context Information* is divided into two, the first part is done with the generated data, simlutating *Flight Delay Events* by reverse engineering, in order to check the *ST* mining algorithm, the selection of the time periods is done so that there is a *Traffic Congestion Area* close to the Frankfurt Airport, as defined one of the landmarks in Section 4.5, and a second period at the same time of a *Football Match Event*, so that the $size(3)$ of the mining algorithm is tested.

The selection of the first period of the generated *Flight Delay Events* is done by a filter of the *Traffic Congestion Events* detected within a ratio of 5000 meters from the coordinates of Frankfurt Airport.

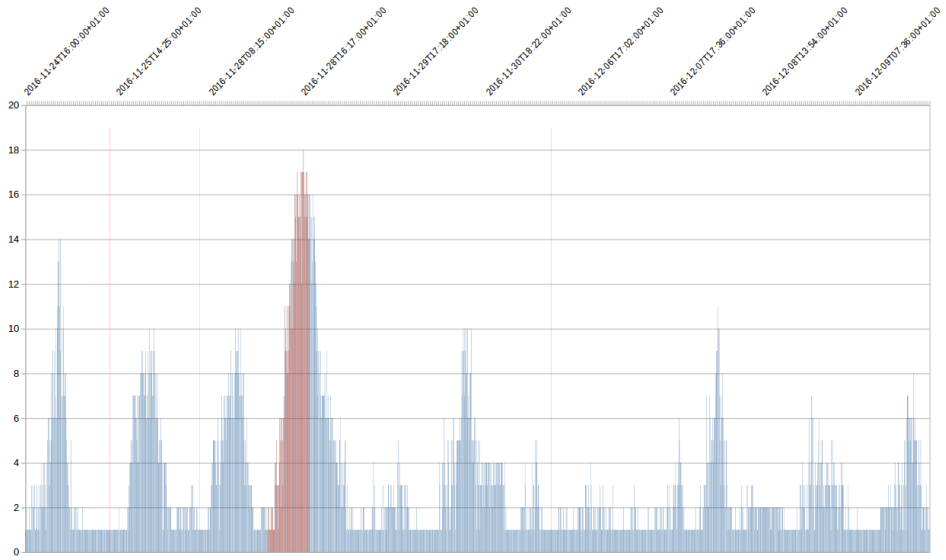


Figure 42: *Traffic Congestion Events* within 5000 meters from Frankfurt Airport.

Figure 42 shows over time the amount of *Traffic Congestion Events* within 5000 meters from Frankfurt Airport, the area in red represents the selected time for the first *Flight Delay Event*, comprises from 2016-11-28T15:00:00+01:00 until 2016-11-28T16:30:00+01:00.

For the *Football Match Events*, it is considered an offset time before and after the official time of the match, as mentioned in Section 4.6. For the evaluation data set, which comprises from November 25th 2016 until December 9th 2016, there are only two *Footbal Match Events* listed in WeltFussbal.de[24], as follows:

- Eintracht Frankfurt - Borussia Dortmund on November 27th 2016, at 15:30 local time.
- Eintracht Frankfurt - 1899 Hoffenheim on December 9th 2016, at 20:30 local time.

With this information the second *Flight Delay Event* is set on November 27th 2016 at 14:00, with a duration of 1 hour.

Table 10 shows the four *External Events* to test the mining algorithm, ordered chronologicaly. *Football Match Events* are listed with the official timing, the offset will be part of the experiments.

[24]https://www.weltfussball.de/alle_spiele/bundesliga-2016-2017/

Event Type	Start Date Time	End Date Time
Flight Delay e_2	2016-11-27T14:00:00+01:00	2016-11-27T15:00:00+01:00
Footbal Match e_3	2016-11-27T15:30:00+01:00	2016-11-27T17:30:00+01:00
Flight Delay e_2	2016-11-28T15:00:00+01:00	2016-11-28T16:30:00+01:00
Footbal Match e_3	2016-12-09T20:30:00+01:00	2016-12-09T22:30:00+01:00

Table 10: *External Event* instances.

For the test of the mining algorithm the ratio of the *External Events* will vary in 5000.00, 10000.00 and 15000.00 meters. The maximum *cce* will be considered and the prevalence measure p, as the average of *cce*. For simplicity the event types are shorten as follows, *Traffic Congestion Area* -¿ e_1, *Flight Delay* -¿ e_2, and *Footbal Match* -¿ e_3. In *Football Match* it is also evaluated the offset of 1 hour, 2 hours, and 3 hours, both, before and after the official times, it is identified as $e_{3,o=X}$, where the X can be 1, 2, or 3, respectively. Results are calculated with all the data set and the test *External Events*.

Co-occurrence candidate	***External Event* radius in meters**	cce_{max}	**Instant of** cce_{max}	$p = cce_{avg}$	**Count Overlaps (minutes)**
e_1e_2	5000	0.6804	2016-11-28T15:44:00+01:00	0.4202	67
e_1e_3	5000	0	*No ST overlap*	0	0
$e_1e_2e_3$	5000	0	*No ST overlap*	0	0
$e_1e_{3,o=1}$	5000	0.3748	2016-12-09T19:31:00+01:00	0.3748	1
$e_1e_{3,o=2}$	5000	0.3748	2016-12-09T19:31:00+01:00	0.2266	43
$e_1e_{3,o=3}$	5000	0.3748	2016-12-09T19:31:00+01:00	0.1931	100
e_1e_2	10000	0.3595	2016-11-28T16:15:00+01:00	0.2171	67
e_1e_3	10000	0	*No ST overlap*	0	0
$e_1e_2e_3$	10000	0	*No ST overlap*	0	0
$e_1e_{3,o=1}$	10000	0.0937	2016-12-09T19:31:00+01:00	0.0937	1
$e_1e_{3,o=2}$	10000	0.3195	2016-12-09T18:37:00+01:00	0.2117	45
$e_1e_{3,o=3}$	10000	0.3195	2016-12-09T18:37:00+01:00	0.2284	109
e_1e_2	15000	0.3654	2016-11-28T16:29:00+01:00	0.1460	67
e_1e_3	15000	0	*No ST overlap*	0	0

$e_1e_2e_3$	15000	0	*No ST overlap*	0	0
$e_1e_{3,o=1}$	15000	0.0416	2016-12-09T19:31:00+01:00	0.0416	1
$e_1e_{3,o=2}$	15000	0.2389	2016-12-09T18:37:00+01:00	0.1443	46
$e_1e_{3,o=3}$	15000	0.4297	2016-12-09T17:43:00+01:00	0.2147	112

Table 11: Co-occurrence *cce* evaluation.

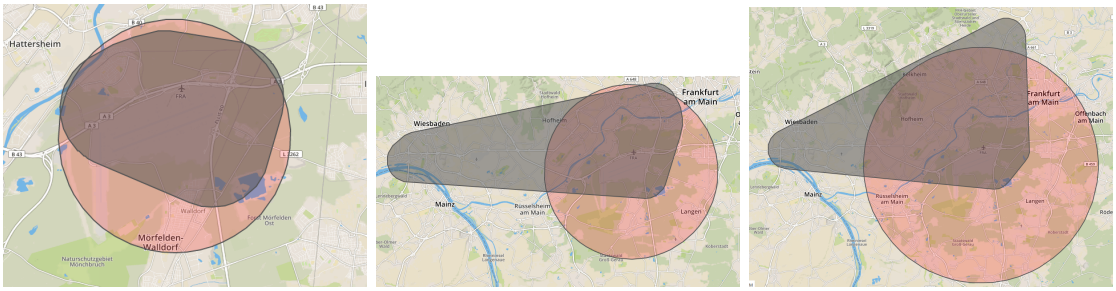
Table 11 contains the results of the experiments for *Levering Context Information* mining algorithm, for three different scenarios, each varies the distance offset for the *External Events*, 5000, 10000, and 15000 meters, this represents the evaluation of the closeness in distance. To test the closeness in time, it is used the *Football Match Event* type, as explained in Section **??**, for the planned events it is added a time offset, before and after, 1, 2 and 3 hours are tested in combination with the variation in distance.

In the following paragraphs it is explained the results of the mining algorithm, this is done for each co-occurence pattern candidate, comparing the different time and space offsets and its results in cce_{max} and cce_{avg}. As a reminder, cce_{max} is the strongest relationship in the sample data set, it is not meant to establish a threshold, but rather explore the use of this approach. For each cce_{max} a visualization is created, as the most representative instant of the context of the *Traffic Congestion Area*, the ultimate goal of the thesis.

For the generated data set of *Flight Delay Events*, one matching *Traffic Congestion Events* and the second one matching a real *Football Match Event*, so that it is possible to test $size(3)$ co-occurrence pattern candidates. First the matching *Flight Delay Event* is evaluated at 5000, 10000, and 15000 meters, as shown in Table 11, results of the cce_{max} are shown in Figure 43.

In Figure 43a the radius of the *Flight Delay Event* is set to 5000 meters, the cce_{max} is a *Traffic Congestion Area* almost contained entirely by the *Flight Delay Event*, that is reflected in the high value of cce_{max}, compared to the two following radius, in which the *Traffic Congestion Area* only intersects partially, as shown in Figure 43b and 43c.

The strong cce_{max} in 5000 meters, compared to 10000 and 150000, is an indicative that there is a higher relationship between the two types of events, in other words, *Traffic Congestion Areas* shown in Figures 43b and 43c can be related to other type of events, if that other type of events would have a bigger intersecting area than the *Flight Delays Event* used in this test.



(a) *External Event* radius = 5000. (b) *External Event* radius = 10000. (c) *External Event* radius = 15000.

Figure 43: *cce Traffic Congestion Area* and *Flight Delays*.

In Figure 44 it is shown the effect of increasing the radius of an *External Event*, which intially may give the notion of *"matching"* more *Traffic Congestion Areas*, however *cce* is calculated in this case with a Jaccard Index, as a proportion of the intersection area over the union area. This is clear for the highest cce_{max} calculated in the test data set, in which the area is increased, so that the same *Traffic Congestion Area* relationship with the *Flight Delays Event* is weaker, therefore is discarded as the strongest relationship.

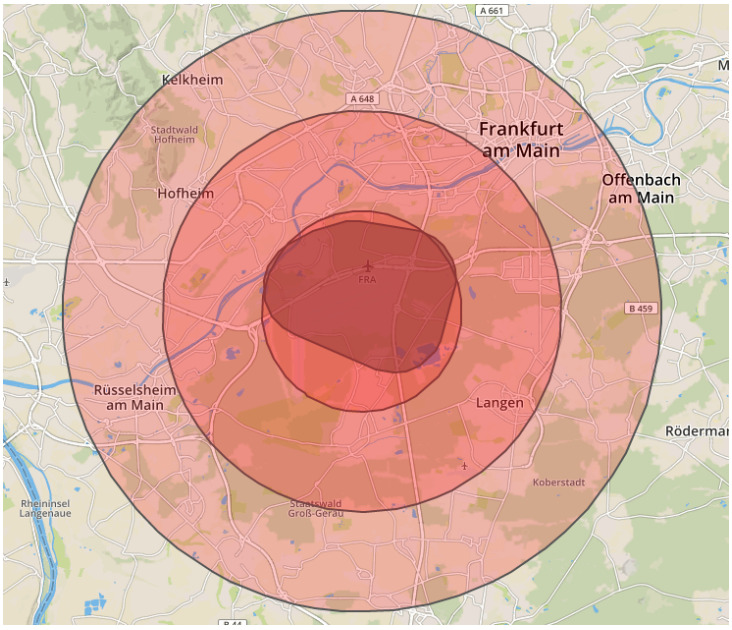
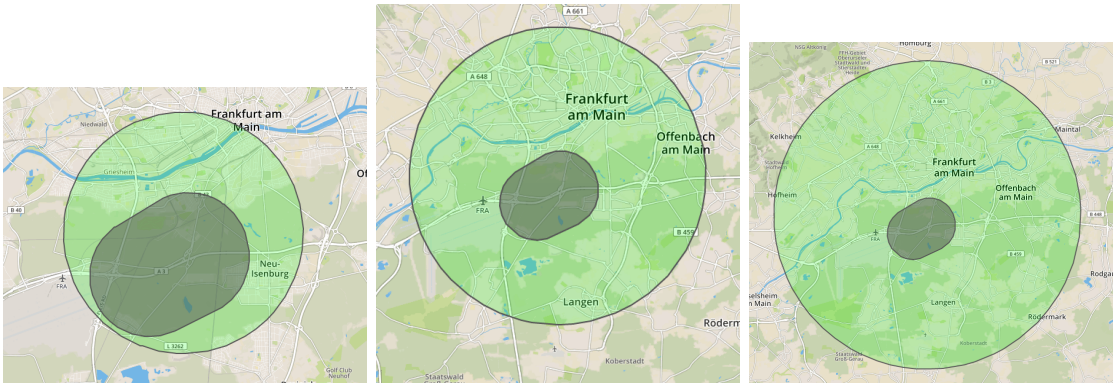


Figure 44: Radius comparison for *Traffic Congestion Area* and *Flight Delays Event*.



(a) *External Event* radius = 5000. (b) *External Event* radius = 10000. (c) *External Event* radius = 15000.

Figure 45: *cce Traffic Congestion Area* and *Football Match Event* at 1 hour offset.

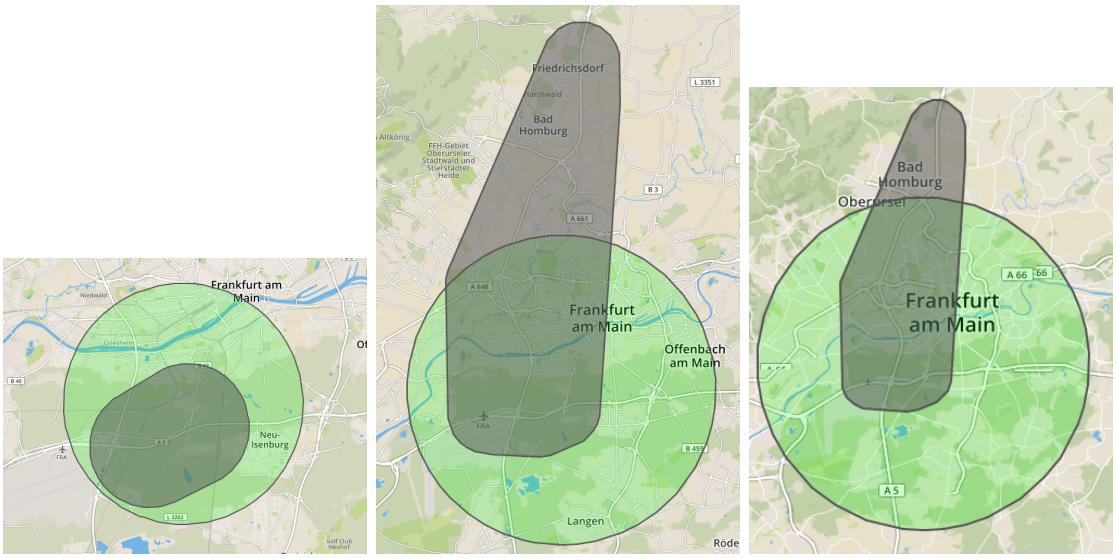
Finally, from Table 11 and Figure 43, it is concluded that the context for the *Traffic Congestion Area* at instant 2016-11-28T16:15:00+01:00 is the *Flight Delays Event* defined in the interval 2016-11-28T15:00:00+01:00 and 2016-11-28T16:30:00+01:00.

For the second *Flight Delays Event*, it was expected to overlap *Traffic Congestion Areas*, however no *Traffic Congestion Areas* were detected in the interval defined, further analysis can be done at *Traffic Congestion Event* level, in other words, at sensor level.

For the *Football Match Events* it was not expected to be related to any *Traffic Congestion Area*, as assumed in Section 4.6, for planned events, people tend to travel before and after the event. This is confirmed in Table 11, as no ST overlaps have been found with the test data set. Therefore $size(3)$ co-occurrence patterns are not found.

The assupmtion for the planned events is tested with a time offset of 1, 2, and 3 hours before and after the official timing, where ST overlaps were found, as shown in Table 11 and in Figures 45, 46 and 47. The variation of distance offset is also tested in combination with the time offset, first the distance offset is explored for each of the time offsets. Time offset variation is discussed with the radius of 15000 meters. It is assumed that the time shown in WeltFussbal.de is the time of kickoff and the football match last 2 hours including half time pause.

With a time offset of 1 hour before and 1 hour after the official timing, varying the distance offset, the result of the co-occurence mining algorithm is the same *Traffic Congestion Area*, as shown in Figure 45, however cce_{max} value changes while increasing the radius, as explained before, due to the relation between the intersection area and the union area of the events. This is expressed in Table 11.

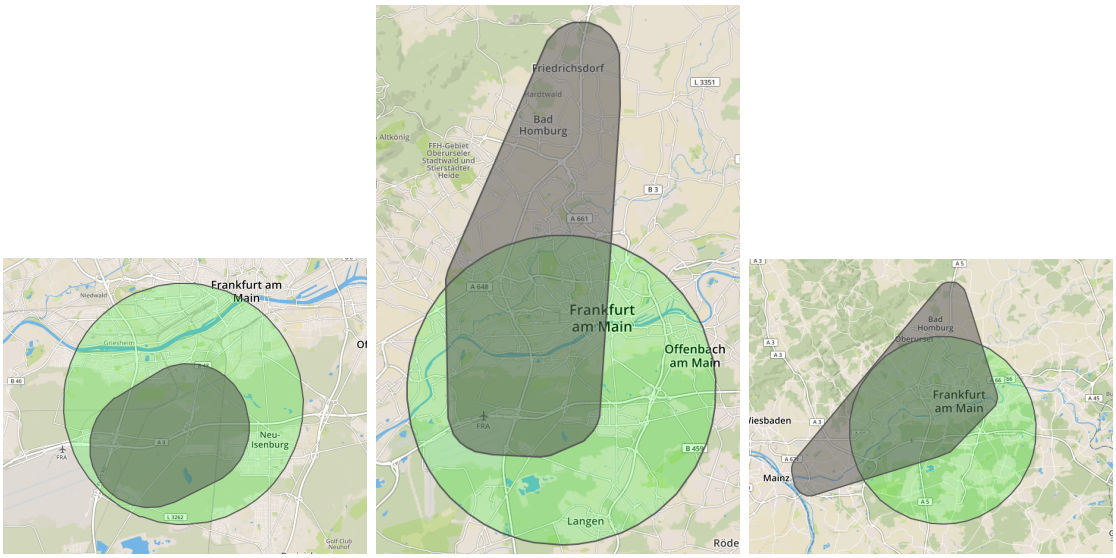


(a) *External Event* radius = 5000. (b) *External Event* radius = 10000. (c) *External Event* radius = 15000.

Figure 46: *cce Traffic Congestion Area* and *Football Match* at 2 hours offset.

Changing the value of the time offset to 2 hours and evaluating the three distance offsets, different *Traffic Congestion Areas* were detected in the cce_{max}, confirms that the variation in time and space influence how different event types are related, one of the objectives in this thesis. Figure 46 shows the map with both event types, *Football Match Event* and *Traffic Congestion Area*. Another value that has changed with the variation of offsets is the total count of overlaps, which is equivalent to the minutes when an overlap occurs. In Table 11 is shown that the increment on the distance offset represents an increment in the minutes of overlap.

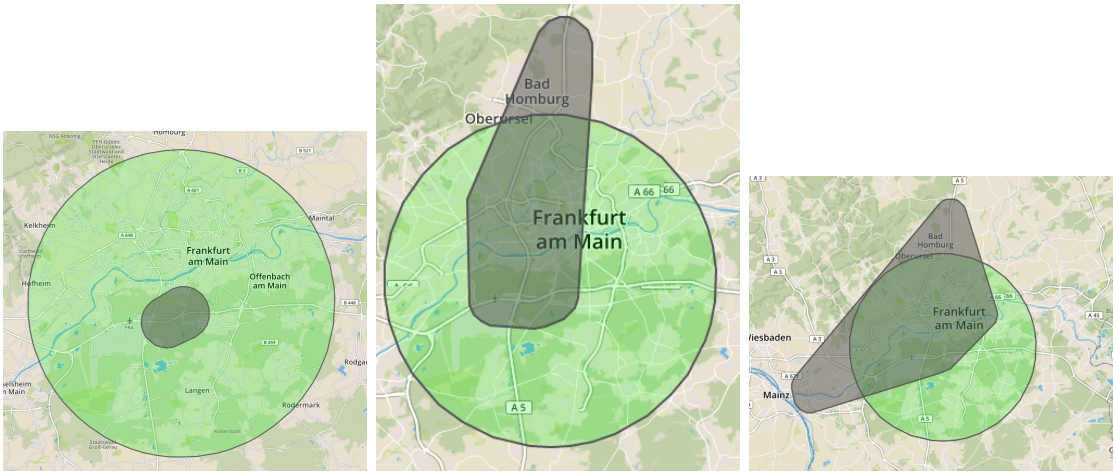
The last variation in time offset is increasing up to 3 hours before and after the *Football Match Events*. Figure 47 shows the cce_{max} for each radius, in comparison with the 1 hour and 2 hours offsets, in this case each variation results in a different *Traffic Congestion Area*. As in the previous explanation, the increment of overlaps with the increment of distance offset, it is shown in Table 11. The interpretation in this case is that there are possible *Traffic Congestion Areas* before which might be related to a *Football Match Event*.



(a) *External Event* radius = 5000. (b) *External Event* radius = 10000. (c) *External Event* radius = 15000.

Figure 47: *cce Traffic Congestion Area* and *Football Match Event* at 3 hours offset.

Finally a comparison between the three time offsets with constant radius, set to 15000 meters as the maximum in this thesis, shows that it is possible to detect different relationships between events in, however this specific data set shows that the closer to the official start time of the event, the less related the *Football Match Event* and the *Traffic Congestion Areas* are. For this case, the strongest relationship is almost three hours before the event starts.



(a) *External Event* time offset = 1 Hour. (b) *External Event* time offset = 2 Hours. (c) *External Event* time offset = 3 Hours.

Figure 48: cce_{max} *Traffic Congestion Area* and *Football Match* at 15000 meters.

In general, the examples of cce_{max} are only part of a *Traffic Congestion Area* that changes over time, those changes can be tracked with an algorithm, such as the MC1, presented in Section 4.4, which was not fully used in this evaluation, due to a flaw in the data set, which is not complete, and the results of the MC1 algorithm are sometimes limited to 3 or 4 consecutive instants. Such improvement, among others, is discussed in Section 6.2. Listing 6 shows an abstract of the resulting Cluster shown in Figure 48c. Line 1 and 2 are the date and time of start and end respectively, with a difference of 3 minutes.

```
1       ...
2       "date_time_start": "2016-12-09T17:40:00+01:00",
3       "date_time_end": "2016-12-09T17:43:00+01:00",
4       "Road": [
5         "A671",
6         "A5",
7         "A66",
8         "A648",
9         "A3",
10        "A661",
11        "A60"
12      ],
13      "description": "Area with 34 sensors reporting traffic congestion.
14      Roads: [A671, A5, A66, A648, A3, A661, A60]. OnRoad: [OnMainRoad].
15      HeadwayTime average in cluster:5.003083862419194.
16      Speed average in cluster:19.034723952218762.
17      HeadwayDistance average in cluster:23.249530417764205.
18      Quantity average in cluster:2281.7647058823522.
19      Density average in cluster:130.28585626074533. ",
```

```
      "cluster_type": "TrafficCongestion",
      "avgDensity": 130.28585626074533,
      "avgSpeed": 19.034723952218762,
      "avgQuantity": 2281.7647058823522,
      "cluster_id": 1481301780000004,
      "avgHeadwayTime": 5.003083862419194,
      "avgHeadwayDistance": 23.249530417764205,
      "cluster_elements": [
        "R2007302",
        "R2008313",
        "R2006112",
        "R2007288",
        "R2008079",
        "R2006793",
        "R2008134",
        "R2008111",
        ...
```


Listing 6: Traffic Congestion Area with cce_{max} within 3 hours time offset and 15000 meters from stadium.

The evaluation of the *Football Match Events* confirms that the use of a time offset in the planned events can give better chances of a co-occurrence. It is not concluding that the *Traffic Congestion Area* is generated by the people going to the *Football Match Event*, but it opens the possibility of further analysis and options. The final conclusions and the foreseen possibilities of the approach presented in this thesis are discussed in Section 6

6. Conclusion and Future Work

As mentioned in Section 1, the main objective of this thesis is to evaluate the use of *ST* algorithms in order to provide contextual data in the mobility domain. In order to achieve such objective, through this thesis fundamental concepts were presented in Section 2. The fundamentals on spatial dimension were explained at first, followed by the temporal dimension, wrapping up both with the explanation of what is known as Spatio-Temporal approach in Section 2.1.3, core of this thesis. This thesis is framed in a mobility use case, so that in Section 2.2 fundamentals of mobility were introduced, followed by basic parameters used in the traffic engineering field, which gives concrete methods to calculate and interpret traffic related data. Finally the focus of the thesis is on traffic congestions, so at the end of the section the reference method to calculate traffic congestions was presented.

Related work on mobility and specifically regarding traffic congestions was explored in Section 3. After an intensive study on previous works related to traffic congestions, at the end of the section weaknesses and gaps are summarized and further used as guidelines for the extended methodology and implementation of this thesis. Further in Section 6.1 those gaps are discussed whether they could be solved as a result of in this thesis.

In Section 4 the methodology is presented that achieves the objective of levering context information in the mobility domain, more specifically having a context for *Traffic Congestions*. The methodology take the most significant gaps detected in the related work as guideline to cover the gaps. Details of data sources and algorithms are mentioned in Section 4 as well. This section links the fundamental concepts with the desired results, it is presented as a chained process, where the output of one of the processes is the input for the following.

Finally Section 5 explains how this methodology was evaluated. The experiments are described, as well as the expected results, based in the fundamental concepts and the available data. This section is also presented as a chained process, where the result of one experiment is the input of the following one. Assumptions and explanations are provided if those are dependent on the test data set.

In this section a discussion about conclusions and future work is done. It takes references from the all the previous sections where questions or challenges are mentioned. The contributions of this thesis are mentioned when a challenge, problem, gap, or question is solved by the proposed methodology and implementation.

6.1. Conclusion

As shown in Figure 49, the process to include additional context in mobility data is divided into two steps: (i) Detection of Areas of Interest, *Traffic Congestion Areas* in this thesis; and (ii) Levering Context Information, using as an input the *Traffic Congestion Areas* and *External Events* as context. Based purely in the main objective of the thesis, the evaluation of *ST* approach in the mobility field was done. Exploring different algorithms, such as the *ST* Clustering in step (i) and the *ST* Mining algorithm in step (ii). Further in this section a deeper discussion of the evaluation is done, based on the results of the two experiments done.

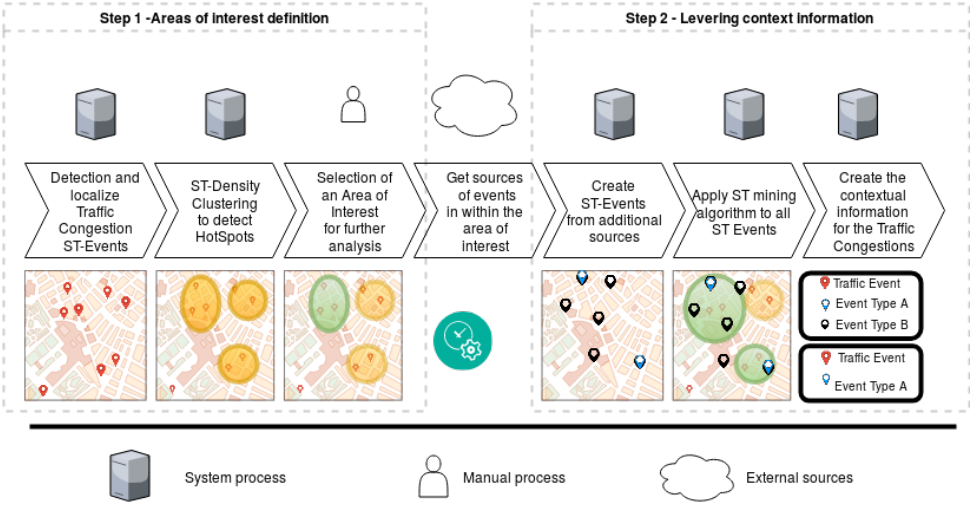


Figure 49: Overview of the process.

In Section 1 the following questions were pointed out and now they are put into perspective with the results of the thesis:

- Which areas may be interesting spots of traffic congestions for further analysis?

  The area selection criteria is detailed in Section 4.5. It suggest areas with higher concentration of sensors. Because the algorithm for clustering and detecting the *Traffic Congestion Areas* is based on density and it works better if a higher number of sensors are located in the area evaluated.

  Also, the area for further analysis should contain important landmarks, such as stadiums or airports, where concentration of people is high, as a precondition for

having mobility and vehicles sensed by the sensors. This condition is used in the evaluation to fix the center points of the *External Events* into real facilities.

Having different type of roads is also one of the criterias, because in contrast with previous works, the main focus is areas of traffic congestion, instead of single points or segments of the road. This criteria was followed to evaluate the methodology suggested in this thesis.

- How close in time and space is "close enough" to determine if an event is likely to trigger traffic congestions?

  Time and space closeness was explored in the Section 5.2, with the use of planned events data set with different time offsets (1, 2, and 3 hours). It was detected that at least 1 hour before and after the planned mass event was enough to detect traffic congestions. Increasing this value also increases the chance of relate different events to the traffic congestions, however it requires a larger study for a better tuning.

  In case of the space, spontaneous and planned events were evaluated with different distances, 5000, 10000, and 15000 meters. This point requieres more experimentation and support from a traffic engineering expert, because each event type is from different nature, in case of a big stadium is more likely to host more spectators than a small stadium, therefore the concentration of people changes and the traffic as well.

  The parameters to define time and distance closeness are set in the *ST* mining algorithm, so it is possible to vary them and do further experimentations.

- Which events are likely to trigger traffic congestions in a certain area?

  During the evaluation, in Section 5.2, it was expected that a *Football Match Event* could be associated to *Traffic Congestion Area*, due to the movement of espectators going to the stadium. However, no traffic congetions were detected, as assumed for the ongoing time of the match.

  The closest in time was almost one hour before the start of the match and within an area of 5000 meters, this is not concluding if that *Traffic Congestion Area* was associated to the match. A further study with longer period of *QnV Records* is required. Also other sources of mass events, such as concerts or demonstrations can be explored using the same methodology. can be verified

- What are the factors for such process in a Big Data environment, considering an increase in volume, velocity, and variety?

This was discussed in Section 4.8, where principles and architectures discussed. The use of distributed processing, fast communication channels, such as Publisher/Subscriber, and Real-Time Processing are elements towards a Big Data environment. Those elements were used in the implementation.

The experiments were set to test the functionality of the system as a priority, so that it is possible to show results by changing the parameters of the algorithms. Additional conclusions mentioned in Section 5.1 is as follows for the evaluation of *Traffic Congestion Areas Detection*:

- The *MinPts* and *eps* parameters that best were evaluated were *MinPts*=2 and *eps*=2500, as the criteria for selection of the parameters followed recommendations from [44]. As mentioned in [44] the selection of values are dependent on the data set and the field of the use case. An initial value of *MinPts*=4 as suggested did not work as expected in the test data set.

- Due to the nature of the data sets, parameters for clustering or the radius in the mining algorithm should be selected by an expert or by running experiments. The way the system was implemented it is possible to vary the parameters with no additional programming effort.

- *DBSCAN* is a non deterministic clustering algorithm, moreover when two core elements have a common border element, it only belongs to one cluster. In the case of the data set and the cluster implementation, the cluster detected were consistent after three runs of the algorithm, as a result the same number of clusters and elements for each one was found.

In case of the evaluation and discussion done in Section 5.2 for *Levering Context Information* the following points are additional conclusions:

- The algorithm worked as expected with the generated data of *Flight Delays* and the real data set of *Football Events* in combination with the *Traffic Congestion Areas*. However further study is suggested for the combination of the three event types, as no multiple event type matching was found.

- The evaluation for the planned event type, the *Football Match Event*, confirmed the that the use of a time offset is useful for events of this nature, where people travels before and after the event, and not during the event.

To sum up, this thesis proposed the use of a *ST* approach in the mobility field. The traffic parameters are used in all the steps of the process, from the transformation of road sensor data into a *Traffic Record*, which was defined in this thesis to contain all the basic traffic

parameters. The use of a clustering algorithm for the detection of *Traffic Congested Areas* instead of a single location or a road segment is a different perspective compared to the previous related work.

6.2. Future Work

As future work is is suggested to explore the following:

- Tune parameters of the clustering algorithm and the mining algorithm.
- Test the implemented system with data set of different regions, where the population changes or the landmarks are different.
- Discuss with an expert in traffic engineering the selection of parameters and the results obtained, as the perspective of a an expert in traffic engineering can open new possibilities. The system itself is flexible enough to be used with other parameters and run further experiments.
- Explore more sources of *External Events*, for instance the *Social Media APIs*, in order to test with larger number of *Event Types*.
- Test *ST* Clustering algorithm with other data sets. As the implementation generalizes the *STDI* it is possible to create wrappers fot the data types and test the same algorithm.
- Modify the MC1 Algorithm or complement it to avoid gaps when data is missing. Smoothing the data is an option. In this way it might be possible to track a *Traffic Congestion Area* that last for more than 30 minutes, for example and do further analysis.
- Automate the selection of *Traffic Congestion Areas of Interest* with the use of other parameters, such as the LCL Table to detect landmarks. As at the moment was done manually, an additional system can be used to select in an automated way the area, depending on the features to test.
- Explore different shapes of the *External Events*, such as polygons with inner gaps, so that the Jaccard Index consider the outer areas, closer to the roads and reduce the inner areas, that bias the result of the mining algorithm.
- Use different criterias to match the events, instead of Jaccard Index of areas, one option is the distance between centroids.

- A guideline is needed to incorporate new data sources to be ingested by the *ST* algorithms, and generalize such a process for different data sets and use cases, with the increasing geotagged data.

References

[1] Dragicevic S. Anton F. Sester M. Winter S. Coltekin A. Cheng T. AA. Li, S. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016.

[2] Schmidt E. Whittle S. Bradshaw R. Chambers C. Akida, T. The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment.*, 8(12):1792–1803, 2015.

[3] J.F. Allen. Maintaining knowledge about temporal intervals. *ACM*, 26(11):832–843, 1983.

[4] Breunig M. Kriegel H. P. Sander J. Ankers, M. Optics: Ordering points to identify the clustering structure. *ACM SIGMOD*, 1999.

[5] Karpatne A. Kumar V. Atluri, G. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, 1(1), 2017.

[6] Kut A. Birant, D. St-dbscan: An algorithm for clustering spatial–temporal data. *Data and Knowledge Engineering*, 60(1):208–221, 2007.

[7] Transportation Research Board. *Highway Capacity Manual.*, 2000. `http://hcm.trb.org`.

[8] Ester M. Qian W. Zhou A. Cao, F. Density-based clustering over an evolving data stream with noise. *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 328–339, 2006.

[9] European Commission. *Transport Eurobarometer: congestion and maintenance are the major challenges for EU roads.*, 2014 (accessed 2018-04-02). `https://ec.europa.eu/transport/media/news/2014-12-08-eurobarometer_en`.

[10] European Commission. *Europe on the move: Commission takes action for clean, competitve and connected mobility.*, 2017 (accessed 2018-05-18). `http://europa.eu/rapid/press-release_IP-17-1460_en.htm`.

[11] R. J. G. B. Campello A. Zimek D. Moulavi, P. A. Jaskowiak and J. Sander. Density-based clustering validation. *In Proceedings of the 14th SIAM International Conference on Data Mining (SDM).*, 2014.

[12] T. et al. Endriks. *Public health for mass gatherings: key considerations.*, 2015 (accessed 2018-09-01). http://apps.who.int/iris/bitstream/handle/10665/162109/WHO_HSE_GCR_2015.5_eng.pdf.

[13] Kriegel H. P. Sander J. Xu X. Ester, M. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96 Proceedings*, pages 226–231, 1996.

[14] Directorate General for Mobility and & TNS Opinion & Social. Transport. *Quality of transport report. Full Report.*, 2014 (accessed 2018-05-25). http://bookshop.europa.eu/uri?target=EUB:NOTICE:MI0614188:EN:HTML.

[15] Colombaroni C. Isaenko N. Fusco, G. Comparative analysis of implicit models for real-time short-term traffic predictions. *IET Intelligent Transport Systems*, 10(4):270–278, 2016.

[16] Bundesamt für Strassen (ASTRA). *Verkehrsfluss und Stauaufkommen - Definitionen*, (accessed 2018-05-18). https://www.astra.admin.ch/astra/de/home/themen/nationalstrassen/verkehrsfluss-stauaufkommen/definitionen.html.

[17] Bundesanstalt für Straßenwessen. *Qualitätssicherun der Location-Code-List, Typen und Untertypen der Lokationen.*, 2010 (provided 2017). http://bast.de.

[18] Bundesministerium für Verkehr und digitale Infrastruktur. *BMVI - Background.*, 2016 (accessed 2018-05-18). https://www.bmvi.de/SharedDocs/EN/Dossier/infrastructure/background.html.

[19] Butler H. Daly M. Doyle A. Schaub T. Gilles, S. *The GeoJSON Format*, 2016. https://tools.ietf.org/html/rfc7946.

[20] Hartmanis J. vanLeeuwen J. Etzion O. Jajodia S. Goos, G. *Temporal Databases: Research and Practice.* Springer Berlin Heidelberg., 1998.

[21] The PostGIS Development Group. *PostGIS Manual*, 2018 (accessed 2018-07-09). https://postgis.net/docs/index.html.

[22] Schneider M. Güting, R. H. Realm-based spatial data types: the rose algebra. *The VLDB Journal—The International Journal on Very Large Data Bases.*, 4:243–286, 1995.

[23] Schneider M. Güting, R. H. *Kurs 01676.*, (accessed 2018-04-09). https://www.fernuni-hagen.de/mathinf/studium/pdf/Leseprobe-komplett_01676.pdf.

[24] Norvig P. Pereira F. Halevy, A. The evolution of big data as a research and scientific topic: Overview of the literature. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[25] Vazirgiannis M. Halkidi, M. Clustering validity assessment using multi representatives. In *Proc. SETN Conf*, pages 237–249, 2002.

[26] Kamber M. Han, J. *Data Mining. Concepts and Techniques.* Morgan Kaufmanns., 2006.

[27] Bovy P. H. L. Hoogendoorn, S. P. State-of-the-art of vehicular traffic flow modelling. *Journal of Systems and Control Engineering.*, (215):283–303, 2001.

[28] Kanade T. Kittler J. Kleinberg J. Mattern F. Hutchison, DH. *Advances in Spatial and Temporal Databases.* Springer Berlin Heidelberg., 2005. Section: On Discovering Moving Clusters in Spatio-temporal Data.

[29] Christidis P. Ibañez Rivas, J. N. Measuring road congetion. Technical report, Institute for Prospective Technological Studies, 2012.

[30] B. Kener. Control of spatial-temporal congested traffic patterns at highway bottlenecks. *Journal of Physics A: Mathematical and General*, 35(3), 2003.

[31] B. Kener. Freeway traffic control based on three-phase traffic theory. *IFAC Proceedings Volumes*, 39(12):368–373, 2006.

[32] B. Kener. Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: A brief review. *Physica A: Statistical Mechanics and its Applications*, 392(21):5261–5282, 2013.

[33] Michels J. Kulkarni, K. Temporal features in sql:2011. *ACM SIGMOD Record.*, 41(3):34–43, 2011.

[34] K. Lemke. *The German Highway Capacity Manual HBS 2014.*, 2014. `https://nmfv.dk/wp-content/uploads/2014/04/The-German-Highway-Capacity-Manual-Kerstin-Lemke.pdf`.

[35] K. Lemke. The new german highway capacity manual. *International Symposium on Enhancing Highway Performance.*, 15:26–35, 2016.

[36] J. Lohmiller. *Qualität des Verkehrsablaufs auf Netzabschnitten von Autobahnen - Bewertung unter Berücksichtigung der Zuverlässigkeit und Analyse von Einflussfaktoren.* Universität Stuttgart., 2014.

[37] Rao K. Mathew, T. *Introduction to Transport Engineering.*, 2007. `http://nptel.ac.in/courses/105101087/`.

[38] Bouaziz R. Moalla M. Mkaouar, M. Querying and manipulating temporal databases. *International Journal of Database Management Systems.*, 3(1):1–17, 2011.

[39] Halevi G. Moed, H. The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends*, pages 3–9, 2012.

[40] Demiryurek U. Shahabi C. Pan, B. Utilizing real-world transportation data for accurate traffic prediction. *2012 IEEE 12th International Conference on Data Mining*, pages 595–604, 2012.

[41] Demiryurek U. Shahabi C. Gupta C. Pan, B. Forecasting spatiotemporal impact of traffic incidents on road networks. *2013 IEEE 13th International Conference on Data Mining*, pages 587–596, 2013.

[42] W. Pietsch. *Big Data – The New Science of Complexity.*, 2012 (accessed 2018-08-13). `http://www.wolfgangpietsch.de/pietsch-bigdata_complexity.pdf`.

[43] Angryk R.A. Banda J. M. Schuh M. A. Wylie T. Pillai, K. G. Spatio-temporal co-occurrence pattern mining in data sets with evolving regions. *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 805–812, 2012.

[44] Sander J. Ester M. Kriegel H. P. Xu X. Schubert, E. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, 42(3), 2017.

[45] Helbing D. Schönhof, M. Criticism of three-phase traffic theory. *Transportation Research Part B*, 43(21):784–797, 2009.

[46] Wong S.C. Xu J.M. Guan Z.R. Zhang P. Tan, M.C. An aggregation approach to short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 10(1):60–59, 2009.

[47] Walter Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography, Supplement: Proceedings. International Geographical Union.*, 46:234–240, 1970.

[48] Pinelli F. Nanni M. Giannotti F. Trasarti, R. Mining mobility user profiles for car pooling. *ACM SIGKDD*, pages 1190–1198, 2011.

[49] Puican F.C. Apostu A. Velicanu M. Ularu, E. G. Perspectives on big data and big data analytics. *Database Systems Journal*, 3(4):3–14, 2012.

[50] Hemsen H. Hennig L. Herth D. Merz M. Reinhard N. Xu F. von Bünau, P. *SD4M Broschuere Web.*, 2017 (accessed 2018-07-25). `https://www.sd4m.net/sites/default/files/publications/SD4M_Broschuere_Web.pdf`.

[51] Deng D. Demiryurek U. Shahabi C. Schaar M. v d. Xu, J. Mining the situation: Spatiotemporal traffic prediction with big data. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):702–715, 2015.

[52] Wu Z. Wang S. Wang Y. Ma. X. Yu, H. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors*, 17(1501), 2017.

Appendix

A. Appendix

A.1. XML Hessen Mobil Files

```
<measurementSiteReference targetClass="MeasurementSiteRecord" id="XX00000X" version="3"/>
<measurementTimeDefault>YYYY-MM-DDT08:25:00+01:00</measurementTimeDefault>
  <measuredValue index="1">
    <measuredValue>
      <basicData xsi:type="TrafficFlow">
        <vehicleFlow>
          <vehicleFlowRate>540</vehicleFlowRate>
        </vehicleFlow>
      </basicData>
    </measuredValue>
  </measuredValue>
  ...
  <measuredValue index="5">
    <measuredValue>
      <basicData xsi:type="TrafficSpeed">
        <averageVehicleSpeed>
          <speed>93.0</speed>
        </averageVehicleSpeed>
      </basicData>
    </measuredValue>
  </measuredValue>
/siteMeasurements>
```

Listing 7: QnV XML Sample.

```
  <measurementSiteRecord id="XX00000X" version="3">
    <measurementEquipmentTypeUsed>
      <values>
        <value lang="de">TLS-Erfassung</value>
      </values>
    </measurementEquipmentTypeUsed>
    <measurementSiteIdeQnVXMLntification>A5/1ALN</measurementSiteIdentification>
    <measurementSpecificCharacteristics index="1">
      <measurementSpecificCharacteristics>
        <period>60</period>
        <specificLane>lane1</specificLane>
        <specificMeasurementValueType>trafficFlow</specificMeasurementValueType>
        <specificVehicleCharacteristics>
          <vehicleType>QnVXMLanyVehicle</vehicleType>
        </specificVehicleCharacteristics>
      </measurementSpecificCharacteristics>
    </measurementSpecificCharacteristics>
    ...
```

```
      <measurementSiteLocation xsi:type="Point">
        <supplementaryPositionalDescription>
          <affectedCarriagewayAndLanes>
            <carriageway>mainCarriageway</carriageway>
            <lane>lane1</lane>
            <lane>lane2</lane>
          </affectedCarriagewayAndLanes>
        </supplementaryPositionalDescription>
        <alertCPoint xsi:type="AlertCMethod4Point">
          <alertCLocationCountryCode>D</alertCLocationCountryCode>
          <alertCLocationTableNumber>01</alertCLocationTableNumber>
          <alertCLocationTableVersion>15.1</alertCLocationTableVersion>
          <alertCDirection>
            <alertCDirectionCoded>positive</alertCDirectionCoded>
          </alertCDirection>
          <alertCMethod4PrimaryPointLocation>
            <alertCLocation>
              <specificLocation>00000</specificLocation>
            </alertCLocation>
            <offsetDistance>
              <offsetDistance>000</offsetDistance>
            </offsetDistance>
          </alertCMethod4PrimaryPointLocation>
        </alertCPoint>
        <pointByCoordinates>
          <pointCoordinates>
            <latitude>50.000</latitude>
            <longitude>9.000</longitude>
          </pointCoordinates>
        </pointByCoordinates>
      </measurementSiteLocation>
    </measurementSiteRecord>
```

Listing 8: QnV Location XML sample.

```
{
    "geometry": {
      "coordinates": [
        8.6478,
        50.1773
      ],
      "type": "Point"
    },
    "type": "Feature",
    "properties": {
      "traffic_flow_[vehicles/hour]": 2580,
      "LCL_location": 12213,
      "short_event_description_en": "Unstable Traffic Flow. Slow Traffic.",
      "level_of_service": "F",
```

```
      "period[min]": 1,
      "sensor_id": "R2018308",
      "short_event_description_de": "Verkehrsqualitats Unzureichend. Stockender Verkehr.",
      "density_[vehicles/km/all_lanes]": 92.95019157088123,
      "date_time": "2018-08-24T12:45:00+02:00",
      "road": "A661",
      "traffic_flow_[vehicles_in_last_1.0_min]": 43,
      "car_ratio_[%cars_over_anyVehicle]": 0.9347826086956522,
      "space_mean_speed_[[km/hr]/km]": 27.756801319043692,
      "on_road": true,
      "time_headway_[seconds]": 2.7906976744186047,
      "time_mean_speed_[km/hr]": 22.75,
      "distance_headway_[metres]": 21.51690024732069,
      "offset_distance": 104,
      "lane_quantity": 2,
      "density_[vehicles/km/lane]": 46.475095785440615,
      "timestamp": 1535107500000,
      "direction": "positive"
    }
  }
```


Listing 9: Traffic Record as GeoJSON.