A review of (dis)aggregation and decomposition methods in traffic assignment

By
Mark P.H. Raadsen, Michiel C.J. Bliemer and Michael G.H. Bell

Institute of Transport and Logistics Studies (ITLS),
The University of Sydney Business School, Australia

April 2019

ISSN 1832-570X
A review of (dis)aggregation and decomposition methods in traffic assignment

In this study we provide a comprehensive review of the existing literature on (dis)aggregation and decomposition methods in traffic assignment. The study takes on two different perspectives. First, we explore existing methods and relate them to their impact on one or more of the components involved in traffic assignment. It is found that there exists a clear separation between a demand modelling point of view, i.e. travel demand and (geographical) zoning, and supply modelling oriented methods, i.e. network topology and network loading. At the same time, the interface between demand and supply, i.e. connector and centroid placement which is to be considered a special type of aggregation, has received conspicuously little attention in this context, even though it is shown to be of significant impact on modelling results. The second perspective in this study places the discussed aggregation methodologies in the broader perspective of clustering procedures. It is shown that most existing methods can be classified as supervised – or classification based – clustering procedures while relatively few studies explore other known approaches such as semi-supervised or unsupervised clustering techniques, at least from a purely traffic assignment model point of view. Lastly, we discuss how aggregation techniques can be deployed to construct multi-scale modelling environments. There is however a lack of methodology to construct such models consistently. It is hoped that this work provides a first step in the direction of such developments by providing an objective classification framework for existing (dis)aggregation and decomposition methods.

Network aggregation, zonal aggregation, decomposition methods, traffic assignment, multi-scale transport models

AUTHORS: Raadsen, Bliemer and Bell

CONTACT: INSTITUTE OF TRANSPORT AND LOGISTICS STUDIES (H73)
The Australian Key Centre in Transport and Logistics Management
The University of Sydney NSW 2006 Australia
Telephone: +612 9114 1824
E-mail: business.itlsinfo@sydney.edu.au
Internet: http://sydney.edu.au/business/itls

DATE: April 2019
1 Introduction

Transport policy makers frequently rely on forecasts of transport models to make decisions, propose legislation, or appraise infrastructure projects. This is still a recent trend given that widespread adoption of transport models only started in the middle of the twentieth century. Adoption of transport models went hand in hand with technological advances, making it possible to design, operate and utilise complicated transport models in a practical setting. The symbiosis between transport model complexity and available computing power has been a balancing act ever since. Often, the adopted level of complexity in a transport model is based on the time it takes for the model to run. The maximum amount of time deemed acceptable varies, but typically ranges from a few hours up to several days. Many practitioners only accept models that can be solved within a single overnight run, i.e. the time between leaving work and turning up the next day. Even so, transport models for middle-to-large cities can still take days to run and this is unlikely to change soon. To reduce run times, aggregation and decomposition methods are utilised to - in part - address the increasing complexity and computation time of such models. They do so by altering the representation of the original model’s inputs and/or procedures.

All representation altering methods balance two conflicting objectives; the alternative model representation should ideally be as capable as the original while at the same time be as minimal as possible. To quantify how capable a transport model representation is, we can measure the amount of information loss suffered compared to the original. Similarly, we can quantify how minimal a transport model representation is by measuring the magnitude of simplification, also referred to as scaling. The term information loss should not be confused with information loss in physical processes, e.g. loss of data on a computer hard disk, it only reflects the error or difference introduced caused by the simplification of the model. The term scaling is used in various disciplines as well, in the context of aggregation methods however, scaling signifies the extent to which model input data (and possibly model procedures) are aggregated, it should not be confused with the concept of scale-free methods used in, for example graph theory.

The magnitude of information loss and scaling are relative measures requiring the original transport model as a reference point. Figure 1 shows the schematic relation of a transport model’s original representation, its alternative representation(s), and how to conceptually assess the impact of any differences between them. The magnitude of the scaling and information loss is determined by the differences between the original and alternative model representation. When one constructs multiple alternative model representations, consistency between alternatives is an additional consideration that might needs to be considered. One can for example construct two alternative representations that are both regarded capable and minimal given their respective application, but this does not necessarily mean we can compare results between these two alternative representations due to potential consistency issues. This is a growing problem in for example multi-scale environments where one operates multiple different transport models alongside each other at different levels of detail but on the same spatial domain. An obvious issue that might occur here is that differently constructed alternative models (with different levels of detail) adopt identical inputs, i.e. operate on the same spatial domain, but yield very different results. When these differences are the result of inconsistencies between the two models, rather than only stemming from the difference in granularity, it is no longer possible to determine which of the two results can be regarded as more accurate, possibly rendering both model outputs useless.
2 Scope, contributions and outline

In this study we only focus on aggregation and decomposition methods that construct a single alternative representation given their reference model. We do so for three reasons; (i) the vast majority – if not all - of the existing aggregation and decomposition methods are of this type, (ii) there is a severe lack of literature/knowledge on the consistency aspect of alternative model representations, (iii) to be able to assess the consistency between alternative representations we require some way to objectively compare them. By classifying existing methods in an objective manner - as part of this study – we hope to make a first valuable step in this direction.

2.1 Scope – traffic assignment

We choose to limit the discussion of aggregation and decompositions methods solely to traffic assignment models because this is arguably the most widely used transport model in existence. While most aggregation and decomposition methods in the literature only impact on the representation of a single aspect of the traffic assignment model or its input(s), we propose to take on a more holistic view when discussing these works and therefore first introduce all components of the traffic assignment model that can be made subject of representation altering methods.

Traffic assignment models consist of two main components: A demand model and a supply model. The demand model is responsible for estimating and/or providing the travel demand based on some zoning system. This results in trips between zones. Trips use the physical road infrastructure to travel across the network. The supply model is responsible for providing the level of service of this infrastructure. The interaction between supply and demand, i.e. network and trips, results in the demand being distributed across eligible paths conditional on the level of service of the network. This process is known as traffic assignment. The procedure that governs the traffic assignment process typically consists of two components; network loading and path choice. Network loading is part of the supply side, it loads the given path flows onto the transport network resulting in generalised costs. These costs are then used in the path choice model to update path flows. Path choice is considered part of the demand side and determines the number of trips assigned to each path conditional on the generalised costs. Apart from path choice, there are many other choices individuals can make when it comes to travel. In traffic assignment we typically limit ourselves to path choice, possibly extended with departure time choice and/or mode choice. The traffic assignment model is considered solved when demand...
and supply are in equilibrium. Figure 2 depicts the traffic assignment model, its inputs and outputs. The demand side inputs comprise traffic demand and the zoning system. The zoning system contains the Travel Analysis Zones (TAZs), also referred to as zones. A Zone serve as an origin, i.e. departure location, and destination, i.e. arrival location, of trips. A zone constitutes a geographical area. The design of the zoning system is a complex process and, in a traffic assignment context, assumed to be exogenous and given. In case the zoning system would be completely disaggregate, it would contain a zone for each household, or person even. In a practical setting, a zoning system often adopts a zone per block, suburb, or village, depending on the granularity or geographical coverage of the model.

On the supply side, the physical road network is a fixed input. A network consists of nodes and links. Links represent a homogeneous stretch of road while nodes represent locations where links intersect. Centroids and connectors are part of the interface that allow the demand and supply to interact. A centroid replaces the zone with a single point of entry/exit. Connectors, also known as connector links, are virtual, non-physical, links between a centroid and a regular node. Centroid and connectors allow the traffic demand to enter, or leave, the physical road network during network loading. It is debatable on whether centroids and connectors are demand or supply side entities. They do however play an important, but underestimated, role in traffic assignment models. We argue they are neither demand nor supply side and classify them separately as the demand-supply interface. For a comprehensive review on traffic assignment we refer the reader to Bliemer et al. (2017).

In this study we make the following assumptions with respect to traffic assignment: (i) we only consider private vehicular traffic, (ii) we focus on strategic and tactical assignment models used for planning purposes, i.e models that do not require the highest level of detail, as they are best suited for applying aggregation methods, (iii) we only consider path choice and ignore departure time, mode choice, or other choice models. We do so since this is most common approach to traffic assignment as well as that most existing aggregation/decomposition methods focus either on model inputs or the network loading procedure and less on the choice modelling aspects of traffic assignment.
2.2 Scope – aggregation and decomposition

Aggregation and disaggregation methods are the two most common types of representation altering procedures used in traffic assignment. Aggregation methods simplify the original representation while disaggregation methods enhance the level of detail in the model. They can be applied to model inputs, procedures, or both. Rogers et al. (1991) discuss a general framework for disaggregation-aggregation analysis in the context of optimisation problems, see Figure 3. The error analysis component for example, relates to the amount of information loss suffered, while the scaling magnitude can be regarded as the result of adopting a particular (dis)aggregation procedure.

In aggregation, individual data points are grouped together and the aggregate result, i.e. group or cluster, is then considered to be a single data point. The procedure then uses the remaining aggregate data points to solve the problem. Aggregation is generally applied to either reduce computational complexity or improve the reliability of the results when data is scarce. The latter is not very common in the context of traffic assignment. Conversely, disaggregation methods expand data points, hoping to reduce the error. Disaggregation comes at the cost of a higher model complexity and increased simulation times. Aggregation and disaggregation can also be applied on the procedural aspects of traffic assignment, for example modelling individual cars in a microscopic model can be replaced by taking on a more aggregate view in a macroscopic model, where traffic is modelled using average flow rates. This means that at the procedural level, a simplifying assumption is applied, resulting in individual data points (vehicles) being aggregated, inevitably causing some information loss.

Alternatively, decomposition methods can be used to speed up computations. In decomposition, an original model procedure is broken up in smaller parts. The idea being that solving each decomposed part separately and combining the results is more effective than solving the problem as a whole, i.e. the dimensionality of the problem is reduced. This in contrast to aggregation where the problem is solved the same way only its inputs are aggregated. Another difference with (dis)aggregation methods is that in decomposition, the original input is preserved and not replaced. Finally, in decomposition methods, the identified sub-components may partially overlap to exchange information, see for example Flötteröd and Osorio (2017), something uncommon in (dis)aggregation. Figure 4 provides a visual interpretation of the differences between decomposition and (dis)aggregation: decomposition methods do not change the underlying assumptions that are made (on the example spatial, temporal, and behavioural categories) but merely focus on changing the procedure to find the same, or similar solutions, at a lesser cost. Aggregation methods on the other hand do change the underlying assumptions to achieve their objectives yet suffer from information loss. In contrast, decomposition methods can be lossless although this might not always be the case hence the additional dashed line in Figure 4(b).
2.3 Contributions and outline

In this work we make the following contributions: (i) discuss the current state of the literature that proposes (dis)aggregation and decomposition methods in the context of traffic assignment, (ii) classify existing methods based on the impacted components of traffic assignment, (iii) identify research opportunities based on the gaps in the literature, and (iv) place existing aggregation techniques in the broader context of clustering procedures.

The remainder of this paper is organised as follows: In Section 3 we introduce the various existing types of aggregation and decomposition that currently exist in the literature in relation to traffic assignment and its inputs. We then discuss each of these types separately in Sections 4 to 8, e.g. zonal aggregation in Section 4, connector and centroid representation in Section 5, network aggregation in Section 6, procedural aggregation in Section 7; Error! Reference source not found., and decomposition in Section 8. We then shift focus to the relation between aggregation methods and clustering in Section 9, followed by a summary of our findings and conclusions in Section 10.

3 Types of (dis)aggregation and decomposition in traffic assignment

Early aggregation methods in transport mainly focussed on removing links in a network. Goldman (1966), for example, removed links from a network based on the criteria of unchanged shortest paths. This process of object removal is referred to as *extraction*. In contrast to *extraction* methods, *abstraction* methods not only remove parts of an original representation, but replace them with a proxy, to mimic the original behaviour. Abstraction based methods are generally thought of as the more desirable approach (Chan, 1976), see also Section 6. The replacement of local roads in a zone with connector links is an example of an abstraction method commonly applied in traffic assignment. Chan (1976) also differentiated between *uniform* and *non-uniform* aggregation methods (or ‘focal’ aggregation as the author originally refers to it). In the context of networks, uniform aggregation methods impact on the whole network while non-uniform methods only affect part(s) of the network. Aggregation methods can aggregate *data* (model inputs) but can also be applied to *procedures*. As mentioned, this
A review of (dis)aggregation and decomposition methods in traffic assignment
Raadsen, Bliemer and Bell

does not mean it groups procedures, but rather changes the underlying model assumptions based on aggregation principles. We refer to the opposite of abstraction as expansion, while we propose the term insertion to reflect the opposite of an extraction process, both of which are types of disaggregation methods. Based on the above classification, the following general definitions – extended from the informal description in Friesz (1985) - are proposed to describe aggregation, disaggregation, respectively in the context of traffic assignment models:

**Definition I: Traffic assignment model aggregation**
The process of decreasing the complexity of a traffic assignment model by aggregating its data, procedures, or both, either uniformly or not, by: (i) extracting original components, (ii) abstracting original components, or (iii) a combination of (i) and (ii).

**Definition II: Traffic assignment model disaggregation**
The process of increasing the complexity of a traffic assignment model by disaggregating its data, procedures, or both, either uniformly or not, by: (i) expanding original components, (ii) inserting new components, or (iii) a combination of (i) and (ii).

While aggregation can impact on both data and procedures, decomposition methods are always procedural. Most existing decomposition methods in traffic assignment pertain to the network loading procedure. Also, although it is theoretically possible to apply decomposition methods non-uniformly, decomposition is generally applied to the entire problem, i.e. uniformly. Figure 5 depicts the taxonomy of the aforementioned (dis)aggregation and decomposition method aspects. This taxonomy does not reflect an ordering, it merely shows how one can categorise existing methods by traversing the tree from the root to a leaf.

![Traffic assignment (dis)aggregation and decomposition](image)

**Figure 5: General taxonomy for traffic assignment (dis)aggregation and decomposition methods.**

### 3.1 Traffic assignment component specific aggregation and decomposition types
In the traffic assignment literature, the most common type of aggregation methods impact on model inputs, and the zoning and network topology specifically. Unsurprisingly, a clear distinction is therefore made between zonal aggregation and network aggregation.

Zonal aggregation scales the zoning system, i.e. the geographical areas. This means that given some original representation existing zones are grouped (or split up in case of disaggregation). In Openshaw and Taylor (1979) this is referred to as the scaling effect of zoning. Besides
scaling effects there also exist zoning effects. Zoning effects refer to changes in model results when one adopts different representations at the same level of scaling. Zoning effects provide insight in inconsistencies that can arise between different model representations. In traffic assignment we assume the base zoning system is given and therefore most methods focus on scaling the zoning system. However, since the choice of the base zoning system has significant impact on the modelling results, we choose to also include a section dedicated to zoning effects as well (Section 4.1). Further, we find that the term zonal aggregation is used rather loosely in the traffic assignment context because it can refer to both aggregation of centroids as well as zones. This is unfortunate since centroids do not equate to zones, i.e. the latter is a demand side entity while the former is part of the demand-supply interface.

Network aggregation is concerned with the grouping links, nodes, and sometimes connectors in the physical network. Connector representation is discussed separately because it has distinct features relating to cost and placement that differ significantly from ordinary links. Therefore, we refrain from using the term aggregation here, because connectors are the result of more than only grouping physical links. Similarly, the placement of centroids representing the underlying geographical zone is also treated separately, as a special case of aggregation, both of which require separate attention and are the topic of interest in the supply demand-interface in Section 5. Procedural aggregation and decomposition methods are relatively rare in traffic assignment. They do however exist and typically affect the network loading possibly often by manipulating the network topology or the traffic flow propagation procedure itself. Let us however start with the discussion of the existing zonal (dis)aggregation methods.

4 Zonal (dis)aggregation

Zonal (dis)aggregation originally emerged as a natural extension to statistical methods with a spatial component. It involves spatial data that needs to be pooled for a variety of reasons, for example to ensure the privacy of individuals, guarantee statistical significance, or a lack of resources to deal with the data in a disaggregate manner. Aggregating such data points can have unexpected and undesirable side effects. The first person to notice that the choice of geographic area used to aggregate – often socio-demographic - data points had an impact on the aggregate results was Dr Henry Sheldon in 1931 (cited in Gehlke and Biehl, 1934). Gehlke and Biehl (1934) confirmed this by showing that the correlation coefficients on the underlying data between geographic areas changed depending on the chosen size. Over the years this effect has been reconfirmed in many different areas and approaches (Kwigizile and Teng, 2009; Openshaw and Rao; 1995; Openshaw, 1977), and has since been referred to as the Modifiable Area Unit Problem (MAUP).

4.1 Modifiable area unit problem and zoning effects

First, we emphasize that MAUP and zoning effects are typically not of direct concern in traffic assignment modelling due to the assumption that the base zoning system is assumed given. However, the methods that exist to address MAUP and zoning effects when constructing an initial/base/generic zoning system are very much of interest because they can be used equally well to modify a given zoning within the traffic assignment model context. Secondly, it is worthwhile including this literature as the negative impacts of adopting a poor base zoning system are often overlooked by practitioners familiar only with the supply side of traffic assignment. Third, some of the literature with respect to creating zoning systems explicitly takes measured/inferred trip data into account as - part of - the objective function or as a constraint (Hagen-Zanker and Jin, 2012; Openshaw, 1977; Baass, 1981) revealing the direct
relation that exists between travel demand within and between zones and the adopted zoning structure.

When constructing the shape and sizes of a geographical zoning system there are an infinite number of possibilities. Typically, each geographical area is constructed based on the location and types of data points within this area. The number of combinations of grouping these data points is finite but it is extremely large in all but the simplest cases, resulting in difficult to solve combinatorial optimisation problems. Finding acceptable (base) zoning structures without spending excessive amounts of time is therefore hard. It is also important to realise that, unless each household is modelled separately, any traffic assignment model remains conditional on the chosen zoning system and is therefore affected by the MAUP to some extent. Therefore, traffic assignment results just reflect one of many possible outcomes (Páez and Scot, 2004; Weeks, 2004).

Addressing the MAUP has traditionally been a statistical exercise more than anything else (Wong, 2001). However, in the last three decades Geographical Information Systems (GIS) have become increasingly popular. Today, they supplement, or even replace, traditional methods when analysing and constructing zoning systems. This seems especially the case in transport related applications. So far, this has resulted in a list of somewhat pragmatic criteria that a zoning system should comply with. These criteria are based on qualitative and quantitative analyses (Ding, 1998; O’Neill, 1991; Baass, 1981) and have also found their way to practice, where they are frequently mentioned as part of, for example, industry guidelines, government recommendations and textbooks (Aecom, 2007; Ortuzar and Willumsun, 2002). Some argue that GIS based approaches are not only a step forward, but can also be regarded as a step backwards, mainly due to a lack of knowledge on the part of the practitioners using the, sometimes, complex software tools required to implement GIS solutions (Fotheringham, 2000). Nevertheless, the impact of GIS inspired methods cannot be ignored. An overview of the most commonly referred to criteria when constructing zones is given in Table 1.

Table 1: Commonly adopted zoning design criteria.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within zone data homogeneity</td>
<td>Low data variance means little error in aggregate form</td>
</tr>
<tr>
<td>Between zone data homogeneity</td>
<td>Comparable errors between zones</td>
</tr>
<tr>
<td>Minimise intrazonal trips</td>
<td>Maximise modelled travel demand</td>
</tr>
<tr>
<td>Adopt census boundaries</td>
<td>Data availability is mostly tied to this spatial unit</td>
</tr>
<tr>
<td>Adopt physical, political and historical boundaries</td>
<td>Implicit information is assumed to be contained within these boundaries</td>
</tr>
<tr>
<td>Convex area shape, i.e. no “holes”</td>
<td>Convex shape attributes to spatial homogeneity</td>
</tr>
<tr>
<td>Contiguous zones, i.e. no gaps</td>
<td>Capture all data within the study area</td>
</tr>
<tr>
<td>Within zone connectivity</td>
<td>If disconnected travel time homogeneity suffers</td>
</tr>
</tbody>
</table>

It is worth noting that some of the criteria conflict with each other or are not agreed upon. For example, it is argued that census boundaries can be a poor choice to delineate zones (Openshaw and Rao; 1995). In Martínez et al. (2009) the authors also highlight this issue and instead propose to construct a zoning system that is not based on the available census boundaries but use a dataset of smoothed geocoded travel demand (converted into a density surface) to construct their zoning system. This seems a natural approach, because when the underlying assignment method relies on travel demand, why not construct the zones based on this same metric? Openshaw and Rao (1995) argue that data should be released at a highly detailed level.
A review of (dis)aggregation and decomposition methods in traffic assignment
Raadsen, Bliemer and Bell

(or possibly disaggregated) and modellers should then construct appropriate zoning systems specific to their application instead of relying on rigid census boundaries. They refer to this application driven approach as optimal zoning, although this is a confusing term since the zoning does not appear to be optimal in terms of guaranteeing an optimal solution but instead indicates it is optimised given some follow-up application. While some might argue that “optimal zoning” will become synonymous with fully disaggregate approaches - with the emergence of big data and agent-based simulation - we argue there remains a need for constructing aggregate zoning systems for traffic assignment. Especially for long-term forecasts aggregate approaches remain important to avoid false accuracy and be able to test out scenarios for which base year data is not available.

4.2 Zonal scaling
Unlike zoning effects when constructing a base zoning system, zonal aggregation or zonal scaling is mainly adopted to reduce computational cost. An exception to this is proposed by Daganzo (1980a), who discusses a disaggregation method, increasing the number of zones in order to improve the accuracy of results. Daganzo argued that the edges of the original zones are misrepresented by a coarse zoning structure, leading to a “spatial aggregation problem”. The method works in conjunction with the Frank and Wolfe algorithm (1956) solving a traditional static assignment. Daganzo also points out that there is a lack of well understood rules on how to construct zoning systems and more specifically, where to locate, and how to connect centroids to the physical road network. Further, he mentions that there is a lack of understanding on what elements of the road network should be included or excluded. So, researchers have long been aware of the fact that all spatial components play a role in choosing an appropriate granularity for the traffic assignment model representation, even when not addressing the issue. It seems that since then very little has changed.

Instead of only disaggregation, one can also opt for a combined aggregation-disaggregation approach. Ruddel and Raith (2013) for example propose a method to reduce computational costs by obtaining a better initial solution. This initial solution is based on an aggregated zoning system and then feeds the result back into the original (disaggregate) problem.

Bovy and Jansen (1983) provide an early example of a more conventional zonal aggregation method. The authors first employ a network aggregation scheme to simplify the road network and - based on these results - aggregate the zoning system which, in their case, is represented by centroids, see also Jeon et al. (2012) for a similar approach. Based on their results, Bovy and Jansen concluded that: (i) Outcomes of traffic assignment are significantly influenced by the level of detail available in the network, (ii) increasing the level of detail in the zoning system and network representation improves traffic assignment results, i.e. reduces information loss, and (iii) this effect becomes marginal beyond a certain level. These findings are in line with Daganzo’s earlier observations.

A recent, more sophisticated approach is proposed by Hagen-Zanker and Jin (2015) who propose an adaptive zoning scheme (Hagen-Zanker and Jin, 2012) where, dependent on the interaction between two zones, a choice is made on the granularity of the departure and arrival zone being modelled within the assignment, where there exist multiple levels of detail for each zone (see also van Steijn, 2016). Results show a marked improvement in accuracy for a given computational budget. Observe that this implicitly also affects the original network representation since coarser zones bypass some of the road network. All the methods discussed here are examples of uniform aggregation methods although one could argue that in some sense
the adaptive zoning method is non-uniform albeit that its non-uniformity is determined in a uniform way. Non-uniform zonal aggregation methods are mainly used when one is only interested in a sub-area of the network. The sub-area is kept at the original higher level-of-detail while the surrounding network is removed or severely reduced and the roads to and from the non-sub-area network are replaced with external zones representing everything beyond the modelled area. These methods’ primary objective is generally to estimate an appropriate travel demand for the external zones rather than devising methodology for the aggregation of zones itself, see for example Zhou et al. (2006) or Larsson et al. (2002), it is therefore of lesser interest in the context of this work.

5 Supply-demand interface: Connector and centroid placement

Once the zoning system is in place, the interface to the underlying road network needs to be established. This involves choosing the location of centroids, the number and/or location of connectors, and estimating the cost/allocated demand of connectors (if fixed). The representation of this interface has significant consequences for modelling results. It also combines aggregation with a design problem since local roads are abstracted out and replaced with a limited number of – assumed representative - connectors and a single centroid.

There are various “best-practices” mentioned in the literature as to where place centroids. These approaches are listed in in Table 2 based on their increasing level of data required to calibrate/choose their location. We found that there has been little attention in the literature for this topic. The sources cited here consider the placement of centroids, but even so only mention these approaches informally and as a side note to their main topic. In most existing literature, authors simply assume centroid locations given or do not even specify the chosen approach. This is surprising because in most traffic assignment applications the location of centroids indirectly dictates the travel time from the zone to the zone boundary, which – especially for short trips - makes up a significant portion of the total travel time.

Table 2: Centroid placement approaches.

<table>
<thead>
<tr>
<th>Centroid placement</th>
<th>Required data</th>
<th>Assumption</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometric centre</td>
<td>Zone shape</td>
<td>Uniform distribution of trips</td>
<td>Friedrich and Galster (2009), Khatib et al. (2001)</td>
</tr>
<tr>
<td>Centre of gravity of weighted network nodes</td>
<td>Internal zone topology</td>
<td>Distribution of trips unknown</td>
<td>Friedrich and Galster (2009)</td>
</tr>
<tr>
<td>City location</td>
<td>City locations</td>
<td>Population density is representative for trips</td>
<td>Khatib et al. (2001)</td>
</tr>
<tr>
<td>Node with highest accessibility becomes centroid</td>
<td>Internal zone topology and node accessibility indices</td>
<td>Accessibility is representative for trips</td>
<td>Friedrich and Galster (2009)</td>
</tr>
<tr>
<td>Population-weighted centre of gravity</td>
<td>Spatial population density</td>
<td>Population density is representative for trips</td>
<td>Khatib et al. (2001)</td>
</tr>
<tr>
<td>Household-weighted density</td>
<td>household locations</td>
<td>household density is representative for trips</td>
<td>Khatib et al. (2001)</td>
</tr>
<tr>
<td>Centre of gravity of disaggregate trips</td>
<td>Trip origins and destinations</td>
<td>Actual trips are representative for trips</td>
<td>Friedrich and Galster (2009)</td>
</tr>
</tbody>
</table>
The placement and construction of connector costs received slightly more attention. A good introduction to issues regarding connector costs and their respective placement is given in Friedrich and Galster (2009), they investigated five different schemes for the cost and placement of connector links. They compare results from a (more) detailed model to some more simplified approaches, where local roads are replaced by connectors. However, as the authors point out, they only consider travel time, not link volumes. Also, their zoning system is assumed fixed and given. They found that within the tested approaches there was no definitive best solution, suggesting they either did not consider enough methods, metrics, or scenarios, or there are simply multiple ways of designing connectors in an acceptable fashion. More research is needed to provide an answer to this question. We also refer the reader to earlier but conceptually interesting work by Daganzo (1980b) where a method for estimating connector costs (termed access costs) is embedded in a traditional static traffic assignment procedure. It is based on an approach where demand in zones is modelled as a continuum, inspired by earlier work of Newell (1980).

A more pragmatic approach to connector placement is proposed by Mann (2002). In this work, existing connector end nodes are upgraded to a sub-centroid, termed B-node. Each B-node is then assigned a fixed portion of the total demand in order to improve accuracy. This approach is used in practice as well, in the Aimsun traffic simulation software (Aimsun, n.d.), one can split existing zones in sub-zones, with new connectors, and/or assign a fixed portion of demand to each existing connector. Clear downsides to this approach are the increased computational cost, possible reliance on pre-existing connector placement decisions, and the difficulty of constructing a justifiable distribution of demand across the sub-centroids.

Both Jafari et al. (2015) and Qian and Zhang (2012) confirmed that not only the zoning system but also the choice of connectors can severely impact traffic assignment results. In Qian and Zhang, the authors add/remove connectors and identify the impacts on reference link volumes. Jafari et al. (2015) not only investigate the effect of connectors on results, but also propose a method they call bi-level selection. They argue that when network detail is added, original connectors might bypass this added detail which is deemed undesirable. They argue network detail should match the zoning as well as the supply-demand interface and propose a connector design with additional connectors close to the centroid. They then distribute demand across the two levels. Although the number of added connectors, as well as the choice for two levels is rather arbitrary, the observation of matching the level of detail between supply, demand and supply-demand interface echoes the findings of the other discussed works.

All these approaches determine connector costs and/or demand distribution deterministically. Benezech (2011) approaches the problem from a more probabilistic perspective. His approach has conceptual similarities compared to Jafari et al. (2015) but instead of two levels, four anchor points (connectors) are defined around each zone and a logit model is adopted to estimate the distribution of demand across these points. The utilities in the logit model are obtained from an underlying - more detailed - network that is used to determine the travel times to the various anchor points. As the author also points out, the number of four is arbitrary. The results – akin to Mann (2002) and Jafari et al. (2015) - do not result in a connector cost, but instead split each centroid/zone in four parts.

1 With connector cost we mean the cost that attributes to the total path costs of paths traversing the connector which in turn are used to feed into the path choice model.
All in all, existing literature on connector cost and placement is limited. While there exists agreement on the fact that connector placement and costs are important, existing methods often lack justification for the choices that are being made. As we can observe from Table 3, connector placement and the number of connectors is virtually always chosen arbitrarily or based on an assumption that is not verified. Also, centroid design is hardly ever considered in conjunction with connector design even though connector costs rely on the location of the centroid. Yet, the validity of the adopted centroid placement is never questioned in these works.

**Table 3: Overview of approaches regarding centroid and connector placement. Colours: red – not quantitatively justified, grey - assumed given, green – quantitatively justified.**

<table>
<thead>
<tr>
<th>Approach description</th>
<th>Connector placement</th>
<th>Connector cost</th>
<th>Number of connectors</th>
<th>Centroid design</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connector based disaggregation</td>
<td>Assumed given</td>
<td>Assumed given</td>
<td>Assumed given</td>
<td>Disaggregate to connector nodes</td>
<td>Mann (2002)</td>
</tr>
<tr>
<td>Single closest node</td>
<td>Closest node to centroid</td>
<td>Zero cost</td>
<td>1</td>
<td>Assumed given</td>
<td>Friedrich and Galster (2009)</td>
</tr>
<tr>
<td>Single concentric sectors</td>
<td>Closest node to sector centroid</td>
<td>Zero cost</td>
<td>1 (per sector)</td>
<td>Assumed given</td>
<td>Jafari et al. (2015)</td>
</tr>
<tr>
<td>Advanced single concentric sectors</td>
<td>Closest node to sector centroid</td>
<td>Regression analysis data (land use/traffic states)</td>
<td>1 (per sector)</td>
<td>Assumed given</td>
<td>Friedrich and Galster (2009)</td>
</tr>
<tr>
<td>Double concentric sectors</td>
<td>Closest node to sector centroid</td>
<td>Zero cost</td>
<td>1 (per sector)</td>
<td>Bi-level centroid disaggregation</td>
<td>Jafari et al. (2015)</td>
</tr>
<tr>
<td>Equal travel time isochrones</td>
<td>Equal travel time to centroid</td>
<td>Average travel time, potentially based on simulation</td>
<td>Topology dependent</td>
<td>Assumed given</td>
<td>Friedrich and Galster (2009)</td>
</tr>
<tr>
<td>Stochastic disaggregation</td>
<td>“variety of directions”</td>
<td>Stochastic, distribution-based estimation</td>
<td>4</td>
<td>Disaggregate to “anchor” points</td>
<td>Benezech (2011)</td>
</tr>
</tbody>
</table>

### 6 Network aggregation

On the supply-side of traffic assignment, network aggregation emerged as a popular method to reduce computational complexity of a model. Early approaches (Bovy and Jansen, 1983; Long and Stover, 1967; Goldman, 1966) mainly considered network extraction methods. There are some more recent examples of network extraction as well, see for example Jeon et al. (2012) or Cui (2016). Early on, it was quickly established that extraction methods are undesirable (Chan et al., 1968) for a number of reasons. It causes (i) reduced capacity on the network, (ii) unrealistic diversion of traffic to remaining arcs, and (iii) reduction, or absence, of network connectivity. While this has been known for decades, practitioners still use ad-hoc extraction/insertion methods to match network detail to their choice of traffic assignment model, highlighting one of the key problems in the current state of practice. This observation is not at all new, Friesz (1985) already pointed out the need for novel methodology to change...
the perspective on the use of network aggregation, an observation that arguably still rings true today.

Early network abstraction methods are found in for example Chan (1976) and Zipkin (1980), where original links, or nodes, are replaced by a simplified proxy. In Chan (1976) this results in three types of proxies: bypass links, interzonal links and intrazonal links. Error analyses demonstrated that total travel time remained unaffected under this scheme. Other more recent network abstraction methods are found in for example Connors and Watling (2015, 2008), or Jafari and Boyles (2016) which try to replace the network topology with a functional form instead. A downside of these approaches, apart from solely considering network representation, is that they are tied to traditional static traffic assignment formulations, or that they adopt highly simplified cost functions that can limit their use in practice.

6.1 Macroscopic fundamental diagrams
The Macroscopic Fundamental Diagram (MFD), also known as the Network Fundamental Diagram (NFD) is an attempt to uniquely capture the relationship between the number of vehicles in an area, referred to as accumulation, and the trip completion rate, i.e. vehicles leaving the network, within this area (Daganzo, 2007). Originally discussed in Godfrey (1969), this concept gained renewed attention in recent years, mainly in the context of control and Intelligent Transport Systems (ITS), where they aim to regulate the number of vehicles in a prespecified area via perimeter control (Geroliminis and Sun, 2011; Daganzo and Geroliminis, 2008; Geroliminis and Daganzo, 2008). The last few years have seen alternative uses, ranging from taxi dispatch control strategies (Ramezani and Nourinejad, 2017) to MFD based assignment procedures such as the ones presented in Kim et al. (2018), Zhang et al. (2015), and Knoop and Hoogendoorn (2013). The latter case is the one that suits the scope of this study best where we consider the proposed MFD approaches as a combination of network aggregation and procedural aggregation, see also the next section. It is network aggregation because part of the road network is replaced with a single “reservoir” like link while at the same time the original network loading procedure is replaced with a simplified version where the resulting reservoir propagates traffic flow on an aggregate level based on the principles proposed in the Cell Transmission Model (Daganzo, 1995, 1994).

Challenges with this approach also remain. The functional relationship is very much topology dependent. Thus, it cannot be extracted without reverting to simulation or empirical data analysis (if at all). Knoop et al. (2015) also noticed that a homogenous layout reduces the scatter on the MFD, this means it only works reliably when the infrastructure characteristics, as well as the traffic states within the area, are relatively homogeneous. As a result, there is a need for aggregation, i.e. clustering, techniques to identify suitable MFD regions, one of which is for example proposed in Saeedmanesh and Geroliminis (2016). Concluding, given the dominant focus on using MFDs for control purposes, the relatively recent (re-)emergence of this field of research, and the promising but few studies on utilising MFDs as an aggregation method for traffic assignment, indicate there is both a need as well as a possibility for interesting and relevant research on this topic.

2 Alternatively, one can also consider the average link densities and link (out)flows on the links within the MFD region instead.
6.2 Non-uniform network aggregation methods

Virtually all existing network aggregation methods are uniform by nature. Likely because when the objective is to reduce computational cost applying the method to the complete topology is most beneficial. However, some of the methods discussed can be used for non-uniform applications in case this is desired. Such examples can be found when the network aggregation procedure is complemented with decomposition techniques, see also the next section, as found in for example Jafari et al. (2016), Jafari and Boyles (2016), Boyles (2012). Alternatively, the choice of simplifying the network locally by an MFD is also regarded as a non-uniform application of network aggregation (Zhang et al., 2015).

7 Procedural aggregation

Procedural aggregation is not a method but rather a change in the underlying model assumptions. When applied uniformly it reflects a choice for a (simulation) modelling paradigm. For example, choosing macroscopic network loading model, i.e. flow rate based propagation over a microscopic network loading model, i.e. individual vehicle propagation. In that case an explicit method to conduct a transformation on a base model/data is absent. It is however important to realise that the choice for a (network loading) procedure brings assumptions with it, impacting on the fidelity and realism of the results. Recall the discussion on consistency between model representations. One aspect of consistency between model procedures is to verify if the underlying assumptions of the various model representations are the same. If they are not and some of the assumptions make the model comparatively more complex while others simplify it, result in the inability to compare these models in a meaningful way. In other words, due to the inconsistency in model assumptions either model can be regarded better or worse depending on which change in assumption is dominant.

In recent years, hybrid simulation models have gained traction, where parts of the network are simulated at a higher fidelity than others, not (only) in terms of network detail but mainly in terms of the traffic flow propagation procedure. This is therefore an example of non-uniform procedural aggregation. In a practical setting, hybrid approaches typically simulate the majority of the network with a coarse(r) network loading model while some parts of the network - that require a higher level-of-detail such as complex intersections, motorways on/off ramps, or parts of city centres - are simulated using microscopic modelling. This might require a conversion mechanism between for example aggregate flow rates and individual vehicles at the boundaries between the different network loading fidelities (Barceló et al., 2005). Some existing approaches adopt individual vehicle propagation for the entire network but simplify their behaviour outside of the more detailed areas by simplifying the lane-changing and/or car-following models (Joueiai et al., 2015; Casas et al., 2011; Burghout et al., 2005). The latter model type is often termed mesoscopic since it resides in between macroscopic and microscopic. Others developed hybrid macroscopic-microscopic simulation models which are more difficult since they require a conversion between two different representations of traffic flow (Bourrel et al. 2003; Helbing et al. 2002). The earlier mentioned hybrid MFD-macroscopic model of Zhang et al. (2015) is also an example of this type of approach. An ongoing challenge for hybrid models is to determine where to impose the boundaries to maximise the capability of the model while minimising the additional computational cost. These hybrid simulation models are also referred to as multi-scale simulation. To avoid confusion we distinguish between two types of multi-scale approaches, where multi-scale traffic assignment procedures allude to the above hybrid simulations, allowing for different granularities within a single assignment, while multi-scale traffic assignment models are models that represent the same spatial domain but operate separately from one another, having...
A review of (dis)aggregation and decomposition methods in traffic assignment
Raadsen, Bliemer and Bell

different networks, zoning systems, network loading and path choice components that are consistent within the model have a different level-of-detail between models. Literature on the latter topic is scarce, although Raadsen (2018) provides some thoughts on this issue mainly from a consistency point-of-view.

8  Decomposition methods
Decomposition in traffic assignment is rare compared to applying aggregation methods, also it is often used in a slightly different context such as network design problems (Barmann, 2015). The most common decomposition methods in traffic assignment are typically combined with the traditional static traffic assignment paradigm since this model type has desirable properties suitable for deploying specific decomposition methods. The most well-known example of decomposition in traffic assignment however is the approach by Beckmann et al. (1956) who showed that one can replace the high-dimensional deterministic path-choice model with a link separable alternative that obviates the need for path enumeration. This decomposition led to the first efficient (static) traffic assignment method to adhere to Wardrop’s first principle (Wardrop, 1952). Akamatsu (1997) did the same only for the stochastic – logit based – equivalent. Other well-known examples of decomposition methods in static traffic assignment can be found in for example Hearn et al. (1984) and Larsson and Patriksson (1992) who apply simplicial decomposition methods, i.e. column generation, to more quickly solve the underlying problem. More interesting from a network perspective however are the methods such as presented in Hearn (1984), who proposed a method to decompose the network physically in a “focus area” and “rest of the network” with interconnecting virtual links. More recent works explore spatial decomposition by constructing partially overlapping subnetworks (Flötteröd and Osorio, 2017) or decomposing the network based on the underlying travel time function under the assumption of a fixed path set (Raadsen et al., 2016).

A contemporary and increasingly more conventional way of deploying decomposition in traffic assignment is to use it in conjunction with parallel computing. For this purpose, one might for example decompose the network in spatially disjoint areas where one looks to minimise the boundary interactions such that the overhead for exchanging information between computational threads is minimised. Examples of this type of decomposition that utilise clustering methods to identify the eligible areas can be found in Yahia et al. (2018) and Jafari et al. (2017, 2016). In those works, each of the areas is simulated at a higher or lower fidelity depending on whether the local area is modelled independently or as part of the whole network. Local areas can be distributed for speedup. These examples showcase possibilities for research in this area by deploying a combination of both aggregation methods (via clustering procedures) and/or decomposition approaches.

In Table 4 we classify – discussed - existing literature that either directly or indirectly affects traffic assignment. We do so on a per component basis (as discussed in Sections 4 to 8) as well as based on the various approaches of aggregation and decomposition discussed in Section 3. Note that most studies focus on a subset of traffic assignment components. When they do consider both the demand and supply-side aspects, they typically reflect studies with a comparative component instead of adopting a truly holistic approach in terms of methodology. Note that some of the mentioned studies are to be discussed in the coming section.
Table 4: Classification of aggregation and decomposition methods by impacted traffic assignment components. Colours: Red – not considered, green – considered (to some extent).

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Travel demand</th>
<th>Zoning effects</th>
<th>Scaling effects</th>
<th>Connectors</th>
<th>Centroids</th>
<th>Network</th>
<th>Path set</th>
<th>Path choice</th>
<th>Network loading</th>
<th>Method</th>
<th>(Dis)Aggregation type</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gehlke and Biehl</td>
<td>1934</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Beckmann et al.</td>
<td>1956</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>decomposition</td>
<td>-</td>
<td>uniform</td>
</tr>
<tr>
<td>Goldman</td>
<td>1966</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>extraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Long and Stover</td>
<td>1967</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction + extraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Chan</td>
<td>1976</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Openshaw</td>
<td>1977</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Openshaw and Taylor</td>
<td>1979</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Daganzo</td>
<td>1980a</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>disaggregation</td>
<td>expansion</td>
<td>uniform</td>
</tr>
<tr>
<td>Daganzo</td>
<td>1980b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>disaggregation</td>
<td>expansion</td>
<td>uniform</td>
</tr>
<tr>
<td>Zipkin</td>
<td>1980</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Baass</td>
<td>1981</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Bovy and Jansen</td>
<td>1983</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction + extraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Heymann</td>
<td>1984</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>decomposition</td>
<td>-</td>
<td>non-uniform</td>
</tr>
<tr>
<td>Hearn</td>
<td>1984</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>decomposition</td>
<td>-</td>
<td>uniform</td>
</tr>
<tr>
<td>Larsson and Patriksson</td>
<td>1992</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>decomposition</td>
<td>-</td>
<td>uniform</td>
</tr>
<tr>
<td>Openshaw and Rao</td>
<td>1995</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Akamatsu</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>decomposition</td>
<td>-</td>
<td>uniform</td>
</tr>
<tr>
<td>Ding</td>
<td>1998</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Khatib et al.</td>
<td>2001</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>2002</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Helbing</td>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction + extraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Mann</td>
<td>2002</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>aggregation</td>
<td>abstraction</td>
<td>non-uniform</td>
</tr>
<tr>
<td>Wei and Chai</td>
<td>2004</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Burghout et al.</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>non-uniform</td>
</tr>
<tr>
<td>Zhou et al.</td>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>non-uniform</td>
</tr>
<tr>
<td>Aecom</td>
<td>2007</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(dis)aggregation</td>
<td>abstraction/expansion</td>
<td>non-uniform</td>
</tr>
<tr>
<td>Connors and Watling</td>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Friedrich and Galster</td>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>aggregation</td>
<td>abstraction</td>
<td>uniform</td>
</tr>
<tr>
<td>Martinéz et al.</td>
<td>2009</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>aggregation</td>
<td>expansion</td>
<td>uniform</td>
</tr>
<tr>
<td>Benezech</td>
<td>2011</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>aggregation</td>
<td>expansion</td>
<td>uniform</td>
</tr>
<tr>
<td>Study</td>
<td>Year</td>
<td>Aggregation</td>
<td>Abstraction</td>
<td>Decomposition</td>
<td>Non-uniform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casas et al.</td>
<td>2011</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boyles</td>
<td>2012</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hagen-Zanker and Jin</td>
<td>2012</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jeon et al.</td>
<td>2012</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joshi et al.</td>
<td>2012</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qian and Zhang</td>
<td>2012</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knoop and Hoogendoorn</td>
<td>2013</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruddel and Raithi</td>
<td>2013</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al.</td>
<td>2014</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connors and Watling</td>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hagen-Zanker and Jin</td>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jafari et al.</td>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jouei et al.</td>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim and Mahmassani</td>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang et al.</td>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cui</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jafari and Boyles</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jafari et al.</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim et al.</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Steijn</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raadsen et al.</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saeedmanesh and Geroliminis</td>
<td>2016</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jafari et al.</td>
<td>2017</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flötteröd and Osorio</td>
<td>2017</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim et al.</td>
<td>2018</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yahia et al.</td>
<td>2018</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aimsun</td>
<td>n.d</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1While connectors replace the underlying network, they do so together (if there are multiple), hence it is neither a true aggregation nor decomposition method. We therefore refer to it as a placement/design approach in this context.
9 Aggregation methods from a clustering perspective

So far, we have discussed existing methods from a transport perspective where we related them to the traffic assignment component(s) they act upon. Let us now shift our focus by providing insight in the underlying methodologies that are used by these methods. While decomposing a traffic assignment procedure is largely a mathematical exercise and can either be done or not, the decision to (dis)aggregate model inputs and to what extent is more opaque. It is always possible to group either zones, links, nodes, centroids etc. and it depends on the underlying objective(s) if this is something worth pursuing. Hence, justification is needed both in terms of the application as well as the proposed procedure. From the existing literature we often find that (dis)aggregation methods are described in terms of (custom) algorithms but the broader context in which these procedures fit is often left untouched. We therefore revisit the aggregation methods discussed earlier to place them in the broader context of clustering methods which have a long history, broader, more general application domain and a literature that distinguishes between various (sub-)types that might prove useful in developing new methods within the context of traffic assignment.

All aggregation methods can be thought of as a clustering exercise with a consequence where we replace the found cluster with something else (or removing it entirely). Clustering methods, like aggregation methods, group data points, where each group becomes a cluster and a group of clusters is referred to as a clustering. Clustering procedures are classified as being either unsupervised, semi-supervised, or supervised. Supervised methods are sometimes, for example in machine learning, also referred to as classification-based clustering. Most of the discussed aggregation methods in transport are in fact classification-based clustering methods.

Classification based approaches rely on prior knowledge to assign components into known classes. For example, in network aggregation one can use road types to dictate which links are to be removed, or abstracted out (Jeon et al., 2012; Chan, 1976; Goldman, 1966). Similarly, classification-based methods can be used to place or (dis)aggregate centroids (Mann, 2002; Khatib et al., 2001; Bovy and Jansen, 1983), or where and how many connectors should be included (Jafari et al., 2015; Benezech, 2011; Friedrich and Galster, 2009). These methods are supervised because the criteria for clustering are known a-priori, so the clustering procedure operates under supervision of these criteria. This in contrast to semi-supervised and unsupervised approaches where only some, or even no, prior knowledge is embedded (Manning and Schütze, 1999). In this study we limit ourselves to the difference in clustering methods based on the inclusion – or absence - of this prior knowledge in the form of (spatial) constraints. Machine learning specific aspects such as the training of the models based on datasets is considered out of scope.

While the distinction between the types of clustering is agreed upon, no single definition of what clustering exactly stands for seems to exist. Here, we adopt the formulation in Pfitzer et al. (2009), who state that clustering is: “a process in which the members of a data set are divided into groups such that the members of each cluster, i.e. group, are sufficiently similar to infer they are of the same type and the members of the separate clusters are sufficiently different to infer they are of different types”. The similarity between cluster members can refer to anything that relates to the application context. Figure 6 shows an example where two different similarity measures result in two different clusterings on the same dataset. The chosen similarity measure drives the clustering result. When using a categorical scale (Figure 6(b)), solutions are generally easier to obtain as the classes, i.e. labels, must be known. Finding
appropriate clusterings when the measure of similarity is continuous, possibly dependent on multiple variables, and/or multiple data points, significantly complicates matters.

In those circumstances more sophisticated clustering techniques come into play possibly providing ideas/opportunities when applied to a transport context. It is worth pointing out that these more advanced clustering methods can already be found in land use modelling or GIS oriented transport application, where they are concerned with – for example - creating socio-demographic based zoning systems. However, these zoning systems are typically not constructed with traffic assignment in mind. Transport modellers often have no choice but to adopt these zoning systems in their application because the demand data is only available at this level. There is potential to adapt, extend, or develop these methods by assessing the methods underpinning these existing approaches and use them to, at the very least, improve the (base) zoning systems used in traffic assignment by considering some supply side aspects, and at the very best, construct these zoning systems in conjunction with the underlying traffic assignment methods.

9.1 Hierarchical and partitional clustering
Given that even comprehensive reviews on clustering techniques only discussing a small subset of the literature on the subject (Jain et al., 1999), we only introduce core concepts of (non-classification based) clustering and directly relate them to the transport domain when possible. A clustering technique is either hierarchical or it is not. Hierarchical clustering methods, also known as agglomerative clustering, have a nested structure where each level closer to the root, merges the clusters of the previous level until only a single partition – the root - containing all data points remains (Ward, 1963). In non-hierarchical methods only a single level of partitioning results, hence the name partitional clustering. Most zoning systems operated by cities, states, and countries are hierarchical in nature. For traffic assignment however, modellers typically settle for a single zonal granularity by extracting a single partitional zoning system. This zoning is then often used in various applications by making local changes as deemed necessary. Consequently, virtually all transport related methods for constructing a (base) zoning system discussed earlier are examples of partitional clustering since they operate on this – assumed given – preconstructed partitional zoning system. Only a few examples of hierarchical clustering in a traffic assignment context exist, such as the adaptive zoning approach in Hagen-Zanker and Jin (2015). Hierarchical clustering can have significant drawbacks. It, for example, imposes a rigid structure between the different levels of detail such that the original – top level - zone boundaries always remain. Based on our earlier discussion on MAUP we know that it is unlikely that, at a different level of detail, these boundaries remain the best choice. Constructing a hierarchical clustering is also computationally costlier than adopting a single level partitioning approach because each additional level of zoning takes up resources to compute and store results. Hence, it depends on the application and how they are used if they are a suitable choice.
A special type of partitional clustering worth mentioning is spectral partitioning (Alpert and Yao, 1995). Spectral partitioning is a graph-based clustering technique that uses eigenvalues and eigenvectors to partition data points. It is considered attractive because it allows for a concise mathematical formulation. Also, it seems to be efficient in terms of computational cost which has led to a small, but growing, literature on the subject in the transport domain, see for example, Yahia et al. (2018), Bell et al. (2017) or Ruddell and Raith (2013). A similar technique termed Non-negative matrix factorisation has been used in the context of MFDs to cluster homogeneous regions in the work of Saeedmanesh and Georliminis (2016).

9.2 Unsupervised Clustering techniques overview

Most traditional unsupervised partitional clustering techniques rely on heuristics, i.e. there exists an underlying objective that one tries to satisfy but the method does not guarantee an optimal solution. The best-known example of an unsupervised heuristic clustering technique (without an explicitly formulated objective) is the k-means algorithm (MacQueen, 1967). In k-means, a dataset is partitioned in k sets of arbitrary size. It requires an initial location for each of the k clusters, the algorithm then assigns each data point to the closest cluster. After each iteration, the average location of each cluster is updated based on the locations of its assigned data points until convergence is reached. It is elegant and simple but has two major drawbacks. Results are strongly dependent on the initial locations and the number of clusters is fixed. The k-means algorithm is an example of crisp clustering, sometimes also referred to as hard clustering, where each data point is assigned to exactly one cluster. For a comprehensive overview of unsupervised crisp partitioning clustering algorithms – among other things – we refer the reader to Gan et al. (2007). In contrast to crisp clustering, fuzzy clustering does not necessarily assign data points fully to one cluster but instead utilises a membership probability. A well-known example is found in the Expectation Maximisation (EM) algorithm which can be considered as the fuzzy equivalent of k-means (Kearns et al., 1997). So far, fuzzy clustering is a largely ignored technique in the context of traffic assignment but given the interest in various forms of uncertainty, stochasticity, and (un)reliability in traffic assignment either on the network loading side (Flötteröd and Osorio, 2017; Jabari et al., 2013), or the demand side (Duthie et al., 2011; Waller et al., 2001), fuzzy clustering might provide a different take on this subject, especially from a zoning perspective. One could for example explore uncertainty of inputs in terms of membership functions rather than injecting statistical elements into the network loading itself. The only publications utilising fuzzy clustering in a context vaguely related to transport seem to stem from the land use modelling field relating activities to spatial areas, see for example Soto and Frias-Martinez (2011) and Demissie et al. (2015).

Spatial (unsupervised) clustering is a specific type of clustering where data points are clustered based on nothing but geographical characteristics. One of the most well-known algorithms in this area is DBSCAN (Ester et al., 1996). It proposes to cluster points based on density and requires just one parameter which acts as the constraint to delineate clusters. A benefit of this approach is that it does not require a predetermined number of clusters. From a transport perspective however, considering only the density of (spatial) data points limits its applicability and additional steps are needed. For example, Kim and Mahmassani (2015) extend DBSCAN to identifying vehicle trajectories by combining it with a hierarchical clustering post-processing step.
9.3 Semi-supervised clustering techniques

Semi-supervised clustering techniques are a hybrid form of their unsupervised and supervised counterparts. This branch of methods differs from traditional unsupervised clustering techniques by incorporating some background knowledge (labelled data) into the method. This background knowledge can be included via different types of constraints, or additional distance-based metrics (Basu et al., 2004). This is particularly useful in case contextual information is available, but this information is insufficient to revert to a classification-based approach. In these situations, one can construct an objective function subject to these constraints, in turn providing the opportunity to cast the problem as – for example – a constrained optimisation problem (with its many existing solution methods). In our context these additional constraints are mostly spatial in nature because they either are imposed on zones or the network.

Wagstaff et al. (2001) were the first to introduce background knowledge into the k-means algorithm by proposing two types of constraints that are especially suitable in a spatial context, namely a must-link constraint and a cannot-link constraint. Both constraints impose a limitation on the relation between two data points and are therefore known as pair-wise constraints, or instance-level constraints. Observe that, because of their pair-wise nature, these constraints can be constructed a-priori and serve as additional input to the clustering algorithm. Wagstaff et al. (2001) use their method for lane identification based on GPS data. Some adaptations of clustering methods using instance-level constraints can be found in Ruiz et al. (2009), Lelis and Sander (2009), or Klein et al. (2002) who extend the k-means and DBSCAN methods with aforementioned constraints. Observe that when constructing a zoning system with contiguity constraints, these constraints are in fact pair-wise, spatial, can(not)link constraints between adjacent zones.

In addition to instance-level constraints there also exist constraints that act upon a cluster rather than on two individual data points. In Joshi et al. (2012) this type of constraint is termed cluster-level constraint. The first to incorporate this type of constraint, as far as the authors are aware, are Davidson and Ravi (2005), who impose a so-called minimum-separation-constraint where they ensure that all data points in the cluster can no further be apart from any other data point in the cluster than a predefined value. Verifying compliance on cluster-level constraints requires the construction of a cluster first, making this type of constraint more computationally costly to consider compared to instance-level constraints. Like pair-wise constraints, they are typically enforced as (binary) hard constraints, i.e. they are either satisfied or not. Alternatively, one can incorporate such constraints as soft constraints. In that case they are placed in the objective function and a penalty is imposed based on the magnitude of the constraint violation. Figure 7 depicts the main categories of clustering techniques and the constraint types that are associated with them. There exist some examples of semi-supervised clustering approaches in the transport context although they are often not classified as such. They are mostly concerned with the demand modelling side and focus on constructing zoning systems (Hagen-Zanker and Jin, 2015, 2012; Martinez et al., 2009; Ruiz et al., 2009; Openshaw and Rao, 1995; Baass; 1981).
9.4 Clustering and optimization

Because clustering problems that incorporate constraints are a type of constrained optimisation problem it is worthwhile exploring some well-known solution methods used in this area which might be of interest in a traffic assignment context. Solutions to constrained optimization problems are either \textit{optimal} or not. The simplest way to guarantee an optimal solution is to adopt a brute-force method searching the complete solution space. This however might not be feasible due to the size of this space. On can dynamically reduce the search space by for example adopting a \textit{branch-and-bound} type approach (Morrison et al., 2016), this reduction stems from bounding the space by cutting branches that cannot yield an improvement in the solution. Branch-and-bound approaches can be optimal or non-optimal depending on their design. They are also suitable for cluster-based problems (Koontz et al., 1975) although not many transport examples seem to exist to date.

In transport planning, we often deal with large networks and even larger zoning systems (>1000 zones are common), the solution space of clustering problems is therefore often too large to solve optimally even when the solution space is dynamically bounded. This is where heuristic approaches become relevant, providing solutions of acceptable quality, i.e. close to optimality, in a reasonable amount of time. Numerous heuristic approaches exist, and new approaches still emerge frequently. Besides simple greedy heuristics and various incarnations of Monte-Carlo simulation, the, arguably, most popular type of a generally applicable heuristic algorithm are \textit{metaheuristics}. Metaheuristics propose a high-level solution approach that can be applied to a wide variety of optimisation problems. They are sufficiently general so that they require little to no knowledge of the underlying problem that is being solved. As a result, they are useful both in an unsupervised as well as a semi-supervised environment. For a general introduction on metaheuristics see for example Voß (2000). To increase the effectiveness of metaheuristics they are often adapted or combined (Puchinger and Raidl, 2005). Some well-known examples of metaheuristics are \textit{Genetic and evolutionary algorithms} (Hruschka et al., 2009), \textit{simulated annealing} (Aarts et al.,2005), \textit{tabu search} (Gendreau and Potvin, 2005), or \textit{swarm intelligence} approaches such as ant-colony optimisation (Dorigo et al., 2006).

In transport modelling, metaheuristics are, again, mostly used to construct zoning systems, since these are highly non-linear complex combinatorial problems. In land use modelling zoning systems sometimes drop the term zone and refer to \textit{regions} instead. For example, in Li et al. (2014) the so called \textit{p}-compact regions problem is specified where the authors aim to create \textit{p} regions out of a given number of smaller polygonal units. The resulting regions are
then used in urban economic models which in turn incorporate a traffic assignment model. The authors define an objective function with multiple criteria and constraints. They use metaheuristics to find a near optimal solution. Analogous to other demand driven models in the literature, they consider socio-economic indicators in addition to shape constraints (compactness measures) to determine their regions and largely ignore supply-side aspects such as capacity constraints, travel times etc. Some other examples of (meta)heuristics to construct geographical areas – or clustering problems in general - can be found in Schockaert et al. (2011), Wei and Chai (2004), or Taillard (2003). It is also worth noting that because metaheuristics can be kept quite generic that unoptimized versions of them can be outperformed by tailor made algorithms as shown by Kim et al. (2016) who demonstrate that their custom zoning algorithm outperforms a generic simulated annealing approach.

In Table 5 the classification of the discussed methods that implicitly or explicitly adopt some form of clustering is provided. It is also indicated if they relate to transport and if so, if they concern the construction of a (base) zoning system, i.e. zoning effects, or not. We found that clustering techniques - in traffic assignment - are predominantly classification based, partitional, crisp, and adopt hard constraints. Exceptions do exist, yet mainly concern the construction of a base zoning system. In that case, semi-supervised and unsupervised methods appear to be the preferred method of choice. However, this often occurs in a more general-purpose context rather than having traffic assignment in mind specifically. Further, if they are constructed in a transport context, they do not consider supply side effects. For example, when constructing the zones, the “closeness” between zones is often dominated by the distance (possibly supplemented with other variables). However, it is well-known that people do not consider the distance but rather travel time when making travel related decisions and as such homogeneity in zones based on distance is arguably rather meaningless. To incorporate travel time in constructing zoning systems, we need to consider supply side aspects and therefore consider traffic assignment results in such procedures. This poses a problem since traffic assignment requires zones, but to obtain “good” zones we require traffic assignment. Research is needed to develop computationally acceptable methods that can deal with this interaction effect. To date very few, if any, such methods exist.

### Table 5: Classification of clustering methods in – predominantly - transport related research.

<table>
<thead>
<tr>
<th>Source</th>
<th>Year</th>
<th>Related</th>
<th>Structure</th>
<th>Membership</th>
<th>Prior knowledge</th>
<th>Constraints</th>
<th>Spatial constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long and Stover</td>
<td>1967</td>
<td>yes</td>
<td>partitional</td>
<td>crisp</td>
<td>supervised</td>
<td>hard</td>
<td>-</td>
</tr>
<tr>
<td>Macqueen</td>
<td>1967</td>
<td>no/-</td>
<td>partitional</td>
<td>crisp</td>
<td>unsupervised</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kooztz</td>
<td>1975</td>
<td>no/-</td>
<td>partitional</td>
<td>crisp</td>
<td>unsupervised</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chan</td>
<td>1976</td>
<td>yes/no</td>
<td>partitional</td>
<td>crisp</td>
<td>supervised</td>
<td>implicit hard</td>
<td>-</td>
</tr>
<tr>
<td>Openshaw</td>
<td>1977</td>
<td>yes/yes</td>
<td>partitional</td>
<td>crisp</td>
<td>semi-supervised</td>
<td>hard (+ soft)</td>
<td>pair-wise</td>
</tr>
<tr>
<td>Baass</td>
<td>1981</td>
<td>yes/yes</td>
<td>hierarchical</td>
<td>crisp</td>
<td>semi-supervised</td>
<td>hard + soft</td>
<td>pair-wise + cluster-level</td>
</tr>
<tr>
<td>Bovy and Jansen</td>
<td>1983</td>
<td>yes/no</td>
<td>partitional</td>
<td>crisp</td>
<td>supervised</td>
<td>implicit hard</td>
<td>-</td>
</tr>
<tr>
<td>Openshaw and Rao</td>
<td>1995</td>
<td>yes/yes</td>
<td>partitional</td>
<td>crisp</td>
<td>semi-supervised</td>
<td>hard + (soft)</td>
<td>-</td>
</tr>
<tr>
<td>Ester et al.</td>
<td>1996</td>
<td>no/-</td>
<td>partitional</td>
<td>crisp</td>
<td>unsupervised</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ding</td>
<td>1998</td>
<td>yes/yes</td>
<td>partitional</td>
<td>crisp</td>
<td>semi-supervised</td>
<td>hard + soft</td>
<td>pair-wise + cluster-level</td>
</tr>
<tr>
<td>Khatib et al.</td>
<td>2001</td>
<td>yes/yes</td>
<td>partitional</td>
<td>crisp</td>
<td>supervised</td>
<td>hard</td>
<td>-</td>
</tr>
<tr>
<td>Wagstaff</td>
<td>2001</td>
<td>no/-</td>
<td>partitional</td>
<td>crisp</td>
<td>semi-supervised</td>
<td>hard</td>
<td>pair-wise</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>2002</td>
<td>yes/yes</td>
<td>partitional</td>
<td>crisp</td>
<td>supervised</td>
<td>hard</td>
<td>-</td>
</tr>
<tr>
<td>Klein et al.</td>
<td>2002</td>
<td>no/-</td>
<td>partitional</td>
<td>crisp</td>
<td>semi-supervised</td>
<td>hard</td>
<td>pair-wise</td>
</tr>
</tbody>
</table>
A review of (dis)aggregation and decomposition methods in traffic assignment

Taillard 2003 no/- partitional crisp unsupervised - -
Wei and Chai 2004 yes/yes partitional crisp semi-supervised hard + soft pair-wise
Davidson and Ravi 2005 no/- partitional crisp semi-supervised hard pair-wise + cluster-level
Lelis and Sander 2009 no/- partitional crisp semi-supervised hard pair-wise
Martínéz et al. 2009 yes/yes partitional crisp semi-supervised hard + soft pair-wise + cluster-level
Ruiz et al. 2009 no/- partitional crisp semi-supervised hard pair-wise
Schokaert et al. 2011 no/- partitional crisp + fuzzy semi-supervised hard + soft pair-wise
Soto and Frias-Martinez 2011 no/- partitional fuzzy unsupervised - -
Hagen-Zanker and Jin 2012 yes//yes hierarchical + partitional crisp semi-supervised hard + soft pair-wise
Jeon et al. 2012 yes/no partitional crisp supervised implicit hard -
Joshi et al. 2012 yes/yes partitional crisp semi-supervised hard + soft pair-wise + cluster-level
Raddel and Raith 2013 yes/no partitional crisp supervised hard pair-wise
Li et al. 2014 yes/no partitional crisp semi-supervised hard pair-wise
Demissie et al. 2015 no/- partitional fuzzy unsupervised - -
Hagen-Zanker and Jin 2015 yes/yes hierarchical + partitional crisp semi-supervised - -
Kim and Mahmassani 2015 yes/no partitional crisp unsupervised hard none
Jafari et al 2016 yes/no partitional crisp predetermined n/a n/a
Kim et al. 2016 yes/yes partitional crisp semi-supervised hard pair-wise
Raadsen et al. 2016 yes/no partitional crisp supervised hard pair-wise
Saeedmanesh and Geroliminis 2016 yes/no partitional crisp unsupervised soft -

10 Summary and Discussion

In this study we explored and classified the existing literature on aggregation and decomposition methods in relation to the traffic assignment modelling paradigm, whether it be inputs (zones, network, paths, demand), procedures (network loading, path choice) or a combination of both. In addition, we introduced the reader to clustering methodology, its use in a broader context and relating it to transport – and traffic assignment – specific aggregation methods to this more general methodological framework. Throughout this work we highlight gaps, potential research directions, and some more subjective observations with respect to what is thought to be of interest to our community. We briefly reiterate these findings below.

First, existing research is scattered across many different fields of research (land use modelling, traffic assignment, operations research, GIS, machine learning, data analytics). The result is the use of different terminology, but more importantly, an often overly narrow focus constrained to the authors’ field of research and/or expertise. In the traffic assignment context this results in the observation that most existing methods consider only small subset of the components often only considering either demand side, or supply side aspects but not both. There exists research on connector costs without considering the zoning system, research on centroid placement without considering network design and research on network design without considering zoning, connectors, and/or centroids. Interestingly, the literature also clearly suggests that altering the representation of a traffic assignment model should be considered in an integrated fashion. Choices regarding the granularity of one component impact on other components and therefore holistic methodology that considers all components – both on the demand and supply side - is long overdue. Yet, little progress has been made in this respect apart from the reconfirmation that it is needed.
Third, there is a gap in the literature regarding the supply-demand interface. Justification on the placement and number of connectors for a given zoning system is lacking, while at the same time it has been recognised that connectors (and their costs) significantly impact traffic assignment results. The same goes for the placement of centroids given the underlying geographical area.

Fourth, the literature on zoning systems and trip matrices considers socio-economic data, statistical information such as census boundaries, and household travel survey results. This is demand side information. The literature does not seem to consider supply side information (travel times, congestion levels, path flows) in this context. This seems odd given the fact that the level of service of a network impacts on travel choices and could - and arguably should - be considered when constructing these long-term zoning systems and their related trip matrices.

Fifth, the last two decades gave rise to an increased interest in clustering methods. The similarities between aggregation and clustering methodologies provide opportunities to exploit these developments and bring them into the transport domain. Currently, in the context of traffic assignment, there is only a relatively small literature that touches on this topic, mainly related to constructing base zoning systems.

Lastly, we see that multi-scale transport models, i.e. different granularities on the same spatial domain, are gaining traction in practice, yet little is known on how to construct such models consistently, regarding their inputs as well as model procedures. By discussing and classifying existing aggregation methods in traffic assignment - both from an input as well as a procedural perspective - in a structured fashion, we hope to have contributed towards a first step in this direction. Ideally, our findings act as a stepping stone, generating novel ideas and/or methodology allowing us to generate/extract consistent multi-scale traffic assignment models at any level-of-detail given the requirements of the modeller or application at hand.

References


造成的交通模型的重新配置和削减方法研究

Raadsen, Bliemer and Bell


https://doi.org/10.1080/13658816.2015.1031671


Raadsen, M.P.H., Bliemer, M.C.J., Bell, M.G.H. (2016). A Lossless Spatial Aggregation procedure for a class of traffic assignment models with point queues. Presented at TRB 95th Annual Meeting


