Journal of Geovisualization and Spatial Analysis (2020) 4:15
https://doi.org/10.1007/s41651-020-00057-4

Citywide Traffic Analysis Based on the Combination of Visual and Analytic Approaches

Feng Liu[1] · Gennady Andrienko[2] · Natalia Andrienko[2] · Siming Chen[2] · Davy Janssens[1] · Geert Wets[1] · Yannis Theodoridis[3]

Abstract

A method for citywide traffic analysis is introduced based on the combination of visual and analytical approaches. Large volumes of GPS data collected from urban vehicles are utilized. In the method, a traffic condition map is constructed, composed of five different layers featuring traffic conditions, road linkage, travel patterns, congestion zones, and traffic flows, respectively. Based on the map, specific transport situations surrounding the congested areas are examined and ways of reducing congestion are suggested. The method is evaluated in the aggregated metropolitan area of Athens and Piraeus in Greece, and the potential and the effectiveness of this technique in analysing traffic are demonstrated. With more and more urban vehicles being equipped with GPS devices, the method can be easily transferable to other regions, paving the way for the adoption of the approach for an up-to-date, spatial-temporal sensitive, visual and analytic method for traffic monitoring that supports the establishment of a more sustainable urban transportation system.

Keywords Visual analytical approaches · Traffic conditions · Travel patterns · Congestion · GPS data

Introduction

Batty (2013) considers a city a system composed of movement flows (between locations and between activities), networks of relationships, and interactions among various entities. For understanding the urban movement flows and traffic, both *visual analytical approaches* (VAA) and *statistical and algorithmic approaches* (SAA) have been developed. The methods analyse a variety of *big mobility data*, including GPS (global positioning systems) (e.g. Chen et al. 2019; Liu et al. 2019), mobile phone (e.g. Skupin 2013), smart card (e.g. Itoh et al. 2016), and space- and time-referenced social media (e.g. Chae et al. 2014; Lansley and Longley 2016) data. Big mobility data (particularly GPS data) offer detailed movement trajectories, enabling the examination of urban mobility in highly spatial-temporal resolution and across a large part of the transport network, having great potential in supporting policies.

However, while a multitude of VAA and SAA has been adopted for mobility and traffic analysis, there are few studies on the combination of these two types of approaches. The existing VAA (and their implementation software programs) are usually manageable for a limited size of data (e.g. a few

✉ Feng Liu
Feng.liu@uhasselt.be

Gennady Andrienko
gennady.andrienko@iais.fraunhofer.de

Natalia Andrienko
Natalia.andrienko@iais.fraunhofer.de

Siming Chen
Siming.Chen@iais.fraunhofer.de

Davy Janssens
davy.janssens@uhasselt.be

Geert Wets
geert.wets@uhasselt.be

Yannis Theodoridis
ytheodoridis@gmail.com

[1] Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6, 3590 Diepenbeek, Belgium

[2] Fraunhofer Institute IAIS - Intelligent Analysis and Information Systems, Schloss Birlinghoven, 53757 Sankt-Augustin, Germany

[3] University Piraeus, 80 Karaoli & Dimitriou Str., 18534 Athens, Greece

Published online: 18 June 2020

gigabytes (GB)), but would be very slow or even fail when being applied to extremely large volumes of data (e.g. dozens of GB). Image processing requires a lot of memories and displays even on a middle level of resolution cannot show millions or even billions of movement points at the same time without completely cluttering the screen (e.g. Andrienko et al. 2013b). Thus, SAA, which are capable of efficiently processing a large size of data, are sorely needed to reduce the data volume by data aggregation and selection and to bring out the important properties. However, such statistical and algorithmic methods, though being able to generate tables and figures or to display the analytical results on a dynamic map (e.g. in a route planner), are inefficient in enabling interactive exploration of the analytical process under various parameter settings and scenarios. The methods also have limitations in supporting traffic managers to maximally utilize the derived results for gaining more insights and designing more effective policy measures. This is exactly the challenge, and if a methodology (or a process) can be found which tightly integrates the VAA and SAA techniques into a common development framework, the integrated method would utilize the strength of both approaches and enable urban mobility and traffic analysis in an interactive manner while based on large volumes of GPS data.

To this end, a visual and analytical method combining both VAA and SAA techniques has been developed in this paper, with the goal of interactively systematically examining traffic conditions and analysing congestion problems across the entire urban area. Compared with existing VAA and SAA, the proposed method offers the following advantages. (1) It integrates VAA and SAA into a common framework, enabling a visual and analytical analysis on urban traffic based on large sizes of GPS data. (2) The visualization allows interactive inspection into traffic situations, facilitating analysts and traffic managers acquiring more insights during the development process of the approach and the exploration of the derived results. (3) Due to the utilization of massive data, the derived results are more objective and representative. (4) A unique citywide traffic condition map is constructed, comprised of five different layers featuring key aspects (i.e. traffic conditions, road linkage, travel patterns, congestion zones, and traffic flows) of the road network respectively. This enables comprehensive analysis on urban traffic by means of various combinations of these layers. (5) Specific transport situations surrounding the congestion areas are examined and ways of reducing congestion are suggested.

The rest of this paper is organized as follows. The “Literature Review” section gives literature reviews, while “The Method” section details the proposed method. A case study is performed in “The Case Study” section, and finally the “Discussions and Conclusions” section ends this paper with major discussions and conclusions.

Literature Reviews

Visual Analytics for Urban Mobility Analysis

The science of visual analytical approaches develops principles, methods, and tools to enable synergistic work between humans and computers through interactive visual interfaces (Thomas and Cook 2005). Such interfaces support the unique capabilities (e.g. flexible application of prior knowledge and experiences, creative thinking, and insights) of humans, and couple these capabilities with machines’ computational strength, enabling the generation of new knowledge from large volumes of and complex data recorded from urban areas (e.g. Xie et al. 2019).

Various VAA have been developed, and they explore mobility and mobility-related phenomena from different perspectives. A review by Andrienko and Andrienko et al. (2013a) considered approaches *from the data processing perspective*. It includes analysing individual trajectories (e.g. trajectory cleaning and clustering) and transforming trajectories into data of various spatial-temporal resolution. A more recent review (Andrienko et al. 2017) outlined approaches *from the travel and driving behaviour perspective*. In terms of travel behaviour, it consists of understanding the details of individual movement traces and the variety of taken routes and assessing the dynamics of aggregated movement of a group along routes of a single road, between origins and destinations, or across the entire study area. With respect to driving behaviour, the review describes methods of detecting movement events (e.g. harsh brakes and sharp turns) and analysing the context, impact, and risk of the events. Moreover, Markovic et al. (2019) presented a viewpoint *from the traffic modelling and management perspective*, specifying the following problems of interest: estimating travel demand, analysing and predicting traffic conditions, designing public transit, improving road safety, and assessing the impact of traffic on the environment.

Statistical and Algorithmic Approaches on Traffic Studies

Alongside visual analytics, there is a rich literature on statistical and algorithmic approaches for urban mobility and traffic studies based on big mobility data. The most commonly used technique is *Floating Car Data* which employs a number of GPS-equipped vehicles (e.g. taxis) as moving sensors to build *Intelligent Transportation Systems* (ITS), in order to monitor real-time traffic conditions on the roads and predict short-term traffic evolution (e.g. congestion) (e.g. Dias et al. 2019). The information is further used for dynamic routing and navigation in a route planner (e.g. the google map https://www.google.be/maps/@50.9304038,5.3372893,12z). ITS provides a real-time solution to an end-user, enabling individuals to effectively manage the current traffic and optimize the utilization of existing transport network capacities. Apart from real-time analysis, efforts have also been devoted to examine major traffic parameters during a certain time period (e.g. several weeks, months, or years) using historical GPS data. The goal is to systematically analyse, uncover, and solve traffic problems that are persistent for a long time, in order to support long-lasting transport improvement. The typical research includes the following. Zeng et al. (2019) developed a Cellular Automaton model to estimate traffic flows for single and two lane highways, while Yuan et al. (2011) proposed an approach of deriving travel time of *landmarks* that are road links frequently traversed by taxis. Moreover, Ruan et al. (2019) constructed models of *traffic density* (the number of vehicles per unit length of the roadway) and further predicted congestion levels and effects of weather on congestion. Cui et al. (2016a) developed urban mobility demand models and detected a serious mismatch between mobility demand and road network services, uncovering major causes of congestion in the study area.

The Method

Overall Structure of the Method

The method introduced here consists of four major steps: (1) The entire study area is divided into zones, and three matrices characterizing traffic conditions, road linkage, and travel patterns of each zone or zone pair are constructed and visualized; (2) based on the matrices, congestion zones are detected; (3) major sources of traffic flows in the obtained congestion zones are analysed; (4) specific transport situations surrounding the congestion areas are further examined and ways of alleviating congestion are explored. Prior to these steps, a preliminary step is conducted for raw GPS data preprocessing and trip extraction.

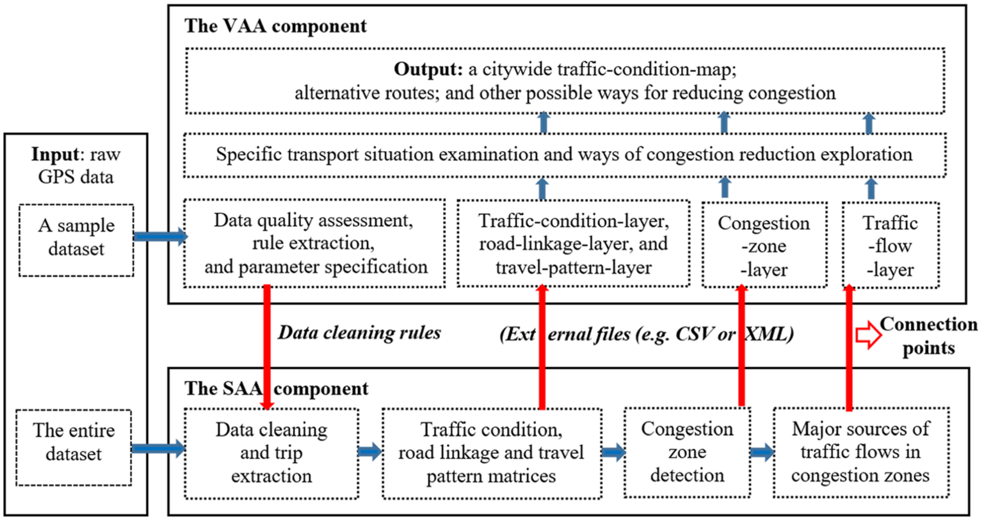
To realize the above steps, two components, namely *VAA* and *SAA* components, are adopted, accommodating the VAA and SAA techniques, respectively. In the preliminary step, the VAA component is applied to *a small sample dataset* drawn from *the entire dataset* (the whole data used in the analysis), in order to interactively examine the data, extract rules for removing data errors, and specify parameter values for trip extraction. Based on these results, the entire dataset is cleaned and trips are derived in the SAA component. In the next steps (steps 1–3), three matrices for traffic conditions, road linkage, and travel patterns are constructed using the obtained trips in SAA, and congestion zones and major sources of traffic flows in the congestion areas are further recognized. All the results are then transferred to VAA, in which they are visualized on different layers of *a map* (*the traffic condition map*). In the last step, specific transport situations surrounding the congestion zones are further examined based on the map, and ways of alleviating congestion are investigated. The connection between the two components includes 4 parts: the extracted rules, the three matrices, the detected congestion zones, and the major sources of traffic flows in the congestion zones. The first part (the extracted rules) is linked from VAA to SAA through knowledge transfer, while the last three parts (the computed results) are connected from SAA to VAA by means of data transfer (i.e. by external files of CSV or XML). The overall structure of the method is shown in Fig. 1, where the connection points (parts) between VAA and SAA are indicated with the red solid arrows between the two components.

GPS Data Preprocessing and Trip Extraction

Rule Extraction and Parameter Specification in VAA

Definition 2.1 A GPS trajectory from a vehicle can be described as a sequence of time-ordered GPS points, i.e. *Tra* =

Fig. 1 Overall structure of the method


The VAA component
Output: a citywide traffic-condition-map; alternative routes; and other possible ways for reducing congestion
Specific transport situation examination and ways of congestion reduction exploration
Input: raw GPS data
A sample dataset
Data quality assessment, rule extraction, and parameter specification
Traffic-condition-layer, road-linkage-layer, and travel-pattern-layer
Congestion -zone -layer
Traffic -flow -layer
Data cleaning rules
(External files (e.g. CSV or XML)
Connection points
The SAA component
The entire dataset
Data cleaning and trip extraction
Traffic condition, road linkage and travel pattern matrices
Congestion zone detection
Major sources of traffic flows in congestion zones

$p_1(id_1, L_1, t_1, a_1)$-... -$p_K(id_K, L_K, t_K, a_K)$; where p_k $(k = 1\ldots K)$ represents a point with its identification (ID) as id_k, longitude and latitude as $L_k = (X_k, Y_k)$, time stamp as t_k, and attributes as a_k (e.g. instant speed v_k and heading h_k), and K is the total number of the points.

Raw GPS data usually contain bad records caused by random errors (e.g. clouding); the existing visual analytical approach on data quality assessment and trip extraction (Andrienko et al. 2016) is adopted. In the method, errors that may occur in all aspects of the data, namely time and space (e.g. time intervals and distances between two consecutive points), and other attributes (e.g. instant speeds and headings) are examined. For trip extraction, the following steps are undertaken:

1. For a group of consecutive points, e.g. $p_i(id_i, L_i, t_i, a_i)$-... -$p_j(id_j, L_j, t_j, a_j)$ $(j > i)$, if each of the points has $v_k = 0$ $(k = i\ldots j)$ and if the stop duration t_{stop} ($t_{\text{stop}} = t_j - t_i$) is longer than a parameter TH_{Act}, the points are clustered into a *stop-location* where the person stays for doing activities.
2. The points in between two adjacent stop locations are extracted as a candidate trip; if the distance (computed according to Formula 1) between the first and last points of the trip is larger than a threshold TH_{Trip}, the trip is retained.

The above method is applied to the sample dataset, by which data errors are identified and rules for eliminating these errors are generated. Moreover, appropriate values for TH_{Act} and TH_{Trip} are set up.

Data Cleaning and Trip Derivation in SAA

Based on the above-derived rules and parameter settings, the entire dataset is cleaned and trips are derived in SAA. For each of the obtained trips, i.e. *Trip* = $p_s(id_s, L_s, t_s, a_s)$-...$p_m(id_m, L_m, t_m, a_m)$...$p_n(id_n, L_n, t_n, a_n)$ ...-$p_e(id_e, L_e, t_e, a_e)$ $(s \le m < n \le e)$, the travel time t_{mn}, travel distance d_{mn}, and travel speed v_{mn} of the segment of the trip between p_m and p_n are computed as follows.

$$
\begin{aligned}
t_{mn} &= t_n - t_m \\
d_{mn} &= \sum_{k=m}^{n-1} [D(p_k, p_k + 1) \\
v_{mn} &= \frac{d_{mn}}{t_{mn}}
\end{aligned}
\qquad (1)
$$

where $D(p_k, p_{k+1})$ is the geographic distance between two consecutive points p_k and p_{k+1} derived from the haversine formula (https://en.wikipedia.org/wiki/Haversine_formula).

Citywide Traffic Condition, Road Linkage, and Travel Pattern Characterization

Study Area Division and Trip Projection in SAA

The entire study area is divided into *Grid_X* × *Grid_Y* disjoint zones using a grid-based method, with each zone being identified as Z_i (i = 1,…, *Grid_X* × *Grid_Y*) or $Z_{i1,i2}$ (i_1 = 1,…, *Grid_X* and i_2 = 1,…, *Grid_Y*). The temporal dimension is classified into different *time periods of a day* (*TimeP*) and different *types of the day* (*DayT*), including weekdays, weekends, and public holidays. Based on the spatial and temporal division, each point of each trip is projected into the corresponding zones and time periods.

Definition 2.2. For each trip *Trip* = $p_s(id_s, L_s, t_s, a_s)$-... - $p_e(id_e, L_e, t_e, a_e)$, all the GPS points along the trace are matched into zones, generating a trip at the zone level as *Trip-zone* = $Z_s(t_s)$-... -$Z_e(t_e)$, with Z_i ($i = s\ldots e$) referring to the corresponding zone. Moreover, the consecutive zones that are identical are further combined, leading to a *travel path* of the trip, i.e. *Path* = Z_s-... -Z_e, with Z_j ($j = s\ldots e$) denoting the distinct zone passed by the trip (see Fig. 2).

Matrix Construction in SAA

Based on the *trip-zone* or *path* of all the trips, three matrices characterizing traffic conditions, road linkage, and travel patterns of the study area are constructed respectively. First, *the traffic condition matrix TC*(Z_i, *TimeP*, *Day*, *DayT*) accommodates all trip segments that traverse each zone Z_i within *TimeP* on the *Day* with the type of *DayT*. Two variables are derived for each matrix element, namely the total number (M_i) of trip segments that traverse Z_i during the time of *TimeP*, *Day*, and

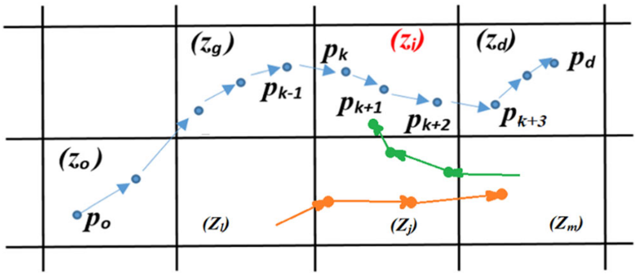


Fig. 2 The zones and trips. *Note*: the black grids represent zones, the curves are trips, and the dots along the curves denote GPS points. The blue curve (trip) starts and ends at p_o and p_d, respectively, covering 4 zones of Z_o, Z_g, Z_i, and Z_d, i.e. *Path* = Z_o-Z_g-Z_i-Z_d. For *TC*, the segment of the trip traversing Z_i (in red) is between p_k and p_{k+3}; for *RL*, the segment passing Z_o, Z_g, and Z_i is between p_o and p_{k+3}; for *TP*, the entire trip between p_o and p_d is adopted. Moreover, the orange and green trips describe the trip segment traversing Z_l, Z_j, and Z_m as well as Z_m, Z_j, and Z_i, indicating the existence of road connectivity Z_l- > Z_j- > Z_m and Z_m- > Z_j- > Z_i, respectively. However, these two connections are not linked inside Z_j, leading to the absence of the connectivity Z_l- > Z_j- > Z_i

DayT and the average travel speed (V_i) over all the segments. The latter variable is computed by Formula 2.

$$V_i = \frac{\sum_{w=1}^{M_i} v_m{'}n'^{w}}{M_i} \tag{2}$$

where w represents a trip that passes Z_i, $p_{m'}$ and $p_{n'}$ are the first points of the trip in Z_i and out of Z_i, respectively (*see* Fig. 2), and $v_{m'n'^{(w)}}$ is the speed of the segment between $p_{m'}$ and $p_{n'}$ (*see* Formula 1).

Secondly, *the road linkage matrix RL*(Z_i, Z_j, Z_l, *TimeP, Day, DayT*) describes trip segments that successively pass three zones of Z_i, Z_j, and Z_l in *TimeP* on the *Day* of *DayT*. Two features are extracted for each matrix element, namely the total number (M_{ijl}) of the segments and the average speed (V_{ijl}) over the segments. V_{ijl} is computed according to Formula 3.

$$V_{ijl} = \frac{\sum_{w=1}^{M_{ijl}} v'\prime n''^{(w)}}{M_{ijl}} \tag{3}$$

where w represents a trip that passes Z_i, Z_j, and Z_l, $p_{m''}$ and $p_{n''}$ are the first points of the trip in Z_i and out of Z_l, respectively (*see* Fig. 2), and $v_{m'\prime n''^{(w)}}$ is the speed of the segment between $p_{m''}$ and $p_{n''}$. *RL* characterizes *road connectivity* (Z_i- > Z_j- > Z_l) between Z_i, Z_j, and Z_l, with the value of M_{ijl} larger than 0 signifying the existence of road connections (in the corresponding direction) among these zones. It should be noted that, despite the existence of road connections among three neighbouring zones, e.g. Z_l- > Z_j- > Z_m among Z_l, Z_j, and Z_m and Z_m- > Z_j- > Z_i among Z_m, Z_j, and Z_i, these two connections may not be linked inside the middle zone Z_j, leading to the absence of the connectivity Z_l- > Z_j- > Z_i, as illustrated in Fig. 2.

Thirdly, *the travel pattern matrix TP*(Z_o, Z_d, *r*, *TimeP*, *Day*, *DayT*) represents all trips that originate from Z_o, end in Z_d, travel along the *r*th distinct path, and start within *TimeP* on the *Day* of *DayT*. Let Num_{od} be the total number of the distinct paths; thus, $r = 1 \dots Num_{od}$. Each element of the matrix is characterized with five variables, namely the travel path ($Path_{odr}$), the total number (M_{odr}) of trips along the path, and the average travel speed (V_{odr}), travel distance (L_{odr}), and travel time (T_{odr}) over the trips. V_{odr}, L_{odr}, and T_{odr} are obtained as follows.

$$\begin{aligned} V_{odr} &= \frac{\sum_{w=1}^{M_{odr}} v_{m'\prime n'\prime^{(w)}}}{M_{odr}} \\ L_{odr} &= \frac{\sum_{w=1}^{M_{odr}} v_{m'\prime n'\prime^{(w)}}}{M_{odr}} \\ T_{odr} &= \frac{\sum_{w=1}^{M_{odr}} v_{m'\prime n'\prime^{(w)}}}{M_{odr}} \end{aligned} \tag{4}$$

where $v_{m'\prime n'\prime^{(w)}}$, $d_{m'\prime n'\prime^{(w)}}$, and $t_{m'\prime n'\prime^{(w)}}$ are the travel speed, travel distance, and travel time of the entire trip w (i.e. between the first and last points of $p_{m'''}$ and $p_{n'''}$) respectively (*see* Fig. 2).

Matrix Visualization in VAA

Once the matrices are constructed, they are imported and visualized in VAA. Specifically, for a chosen time period and day (or day type), the variables from each of the matrices, including M_i and V_i from *TC*, M_{ijl} and V_{ijl} from *RL*, and $Path_{odr}$, M_{odr}, V_{odr}, L_{odr}, and T_{odr} from *TP*, are visualized on three different layers (including *the traffic condition layer, road linkage layer*, and *travel pattern layer*) of the traffic condition map, respectively, for a part or the entire study area.

Detection of Congestion Zones

Congestion Zone Detection in SAA

To detect congestion zones, the method proposed in the literature (Cui et al. 2016a; Zheng et al. 2011) is adopted. This approach consists of two steps; the first is to identify zones with congestion each day, while the second is to examine the occurrence probability of the daily congestion over all observation days. The aim is to detect zones that regularly suffer from traffic problems:

1. To detect congestion on a day, the zones with $M_i > TH_{mz}$ and $V_i < TH_{vz}$ are selected, where M_i and V_i are the number of trip segments and average speed of the segments in Z_i during *TimeP* of the day, and TH_{mz} and TH_{vz} are the thresholds for the minimum number of trips and lowest speed in normal traffic situations in *TimeP*, respectively. The condition $M_i > TH_{mz}$ ensures that the detected zones have a high volume of traffic and they thus may accommodate important travel corridors or activity locations (e.g. high-density residence, employment, shopping, and/or leisure areas). This high volume also enables the value of V_i to be more accurate and better represent the traffic conditions of the corresponding zones.
2. To differentiate the situations between temporary congestion for only one or several days caused by anomaly events (e.g. traffic accidents or road construction work), and systematic problems that remain for a long time and constantly cause poor transport performance, *the occurrence probability* (P_i) of the congestion in Z_i over all the observation days of the same type is further computed by Formula 5.

$$P_i == \frac{\sum_{Day=1}^{Num(DayT)} I\left(M_i > TH_{mz\ and\ V_i < TH_{vz}}\right)}{N(DayT)} \tag{5}$$

where *Num(DayT)* denotes the total number of the days of *DayT*. *I(x)* is a Boolean function with *x* as the logical parameter; $I = 1$ if $x =$ true and $I = 0$ if otherwise. The zone with P_i larger than a threshold TH_{con} is chosen as a *congestion zone*, referred to as Z_{con}(*TimeP*, *DayT*).

Congestion Zone Visualization in VAA

All the detected congestion zones are visualized on the 4th layer (*congestion zone layer*) of the map, showing the geographic distribution of these areas. Detailed traffic conditions, road connectivity, and/or travel patterns of these zones can be observed from the previously constructed layers.

Major Sources of Traffic Flows in the Congestion Zones

Traffic Source Identification in SAA

Definition 2.3. Given a congestion zone Z_{con}(*TimeP*, *DayT*), all trips that start, end, or just pass Z_{con} in *TimeP* of *DayT* are extracted, and they are further classified based on the OD zones of the trips. In each class, if *the number of trips over the total number of the trips* ($M_{\text{od-con}}$) is larger than (or equal to) a threshold $TH_{\text{od-con}}$, the class is selected. The corresponding OD zones and the most frequently used route of the trips are defined as *a major source of the traffic flows in* Z_{con}, referred as *Source*($Z_{\text{o-con}}$, $Z_{\text{d-con}}$, $Path_{\text{od-con}}$). $Z_{\text{o-con}}$ and $Z_{\text{d-con}}$ represent the OD zones, and $Path_{\text{od-con}}$ is the travel path, i.e. $Path_{\text{od-con}} = Z_{\text{o-con}}$-... Z_k-...-$Z_{\text{d-con}}$, with $Z_{\text{o-con}} = Z_{\text{con}}$, $Z_{\text{d-con}} = Z_{\text{con}}$, or $Z_k = Z_{\text{con}}$ (*see* Fig. 3).

Traffic Source Visualization in VAA

A 5th layer (*traffic flow layer*) is constructed, on which all the major sources of traffic flows that start, end, or just pass a congestion zone Z_{con} (upon selected) are shown in terms of $Z_{\text{o-con}}$, $Z_{\text{d-con}}$, and $Path_{\text{od-con}}$.

Specific Transport Situation Examination and Ways of Congestion Reduction Exploration in VAA

Alternative Routes Avoiding the Congestion Zone Z_{con}

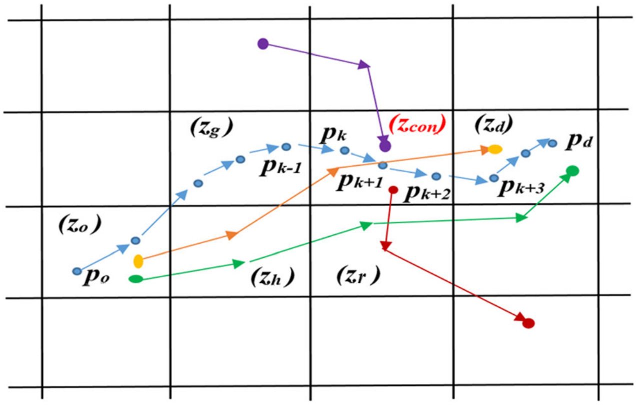


Fig. 3 Major sources of traffic flows in the congestion zone Z_{con}. *Note*: the black grids represent zones, and the grid Z_{con} (in red) is the congestion zone. The curves denote the trips; four of which start, end, or just pass Z_{con}, including the red trip starting in Z_{con} (classified into the 1st class), purple trip ending in Z_{con} (the 2nd class), and blue and orange trips passing Z_{con} (the 3rd class). Let $TH_{\text{od-con}} = 0.3$, the 3rd class is selected as the major source *Source*($Z_{\text{o-con}}$, $Z_{\text{d-con}}$, $Path_{\text{od-con}}$), with $Z_{\text{o-con}} = Z_o$, $Z_{\text{d-con}} = Z_d$, and $Path_{\text{od-con}} = Z_o$-$Z_g$-$Z_{\text{con}}$-$Z_d$ or $Path_{\text{od-con}} = Z_o$-$Z_h$-$Z_{\text{con}}$-$Z_d$. The remaining green trip ($Path = Z_o$-Z_h-Z_r-Z_d) starts and ends in the same zones as this major source but bypassing Z_{con}, and it is thus chosen as the alternative route between Z_o and Z_d

For each *Source*($Z_{\text{o-con}}$, $Z_{\text{d-con}}$, $Path_{\text{od-con}}$) of the flows in Z_{con}, if $Z_{\text{o-con}} \neq Z_{\text{con}}$ and $Z_{\text{d-con}} \neq Z_{\text{con}}$, i.e. if the trips (*passing trips*) between $Z_{\text{o-con}}$ and $Z_{d\text{-con}}$ just pass Z_{con} instead of starting or ending in Z_{con}, we search for *alternative routes* ($Path_{\text{alter}}$) that start and end in the same zones as $Z_{o\text{-con}}$ and $Z_{\text{d-con}}$ respectively but circumventing Z_{con}. The goal is to divert the passing trips to $Path_{\text{alter}}$ to avoid Z_{con}, thus to alleviate the traffic congestion inside Z_{con}. To this end, all the distinct paths ($Path_{odr}$, $r = 1 \ldots Num_{od}$) with $Z_o = Z_{o\text{-con}}$ and $Z_d = Z_{\text{d-con}}$, obtained from *TP*(Z_o, Z_d, *r*, *TimeP*, *Day*, *DayT*), are visualized (on the travel pattern layer). Among these paths, the ones that do not traverse Z_{con} are selected (if there is existence of such paths), and the corresponding values (M_{odr}, V_{odr}, L_{odr}, and T_{odr}) are listed in an attribute window. The path with the shortest travel time (or distance) is subsequently chosen as the alternative route $Path_{\text{alter}}$ (*see* Fig. 3).

Other Possible Ways to Further Alleviate the Congestion in Z_{con}

If $Path_{\text{alter}}$ is not found, i.e. if all trips between $Z_{\text{o-con}}$ and $Z_{\text{d-con}}$ have passed Z_{con}, this suggests that the congestion zone Z_{con} is critical for linking the corresponding OD zones, and that the traffic situations in the surrounding area of Z_{con} leave less (acceptable) choices for drivers to take other routes. It is likely that among all the potential routes between $Z_{\text{o-con}}$ and $Z_{\text{d-con}}$, the current path $Path_{\text{od-con}}$ (e.g. provided by a route planner) has the shortest travel time (or distance). In contrast, the other routes (if exist) may take too much longer travel time or distances than $Path_{\text{od-con}}$ to be considered by travellers. In this case, an in-depth examination into the specific road situations surrounding Z_{con} is conducted, and the matrix *RL*(Z_i, Z_j, Z_l, *TimeP*, *Day*, *DayT*) is utilized. The analysis focuses on the following aspects. (1) The road linkage (M_{ijl} and V_{ijl}) in *the* *surrounding area of* Z_{con} consisting of Z_{con} and its geographically adjacent zones, (2) the mismatch between the road connections and major sources of the flows, and (3) the possible new routes that could be generated in order to replace the current travel path.

The Case Study

The Data and Study Region

The data are provided by a large logistic company in Greece which monitors the travel of more than 6500 utility vehicles across the country via GPS devices installed in the vehicles. The devices collect information at an average rate of 0.48 min, generating data of 1.07 GB each day and 1.1 TB over 3 years. A range of variables is collected, including vehicle positions (e.g. GPS coordinates and recording time), drivers' driving patterns (e.g. instant speeds and headings), and vehicle conditions (e.g. engine status and fuel levels). The study region is *Attica Basin*, the conglomeration of Greek cities of Athens and Piraeus as well as their suburban areas, with a population of over 3.5 million inhabitants. According to the report (Tuszyńska 2018), logistics has been a core sector for the economic growth of this region, manifested by a high number of utility vehicles, i.e. 120 vehicles per 1000 residents. Accordingly, the data cover approximately 1.5% of the total fleet of this area. In this study, the data of 3 months between September and November of 2018 are used, and the total size of the data is 86.4 GB. The adopted variables include vehicle IDs, GPS coordinates and time, and instant speeds.

The VAA and SAA Components

While a range of VAA (e.g. Robinson 2017) and SAA (e.g. R and Python) tools can be utilized for the analysis, in this study, the two components are implemented by means of *V-Analytics* (the Geospatial Visual Analytics) and *SAS* (the Statistical Analysis System), respectively. V-Analytics (Andrienko et al. 2013a) is a software package composed of a set of comprehensive functionalities, realizing a number of state-of-the-art visual analytic approaches. It provides both a basic platform and more advanced techniques for interactively analysing mobility data and extracting knowledge from the data. SAS is a software suite for advanced analytics; it accommodates a multitude of classical and modern statistical and data management approaches. Its architecture designs enable the system to scale up to perform efficient analysis on large sizes of data (e.g. Pope 2017).

GPS Data Preprocessing and Trip Extraction

From all the experimental data, a sample dataset, consisting of 1300 (20% of the total) vehicles for a regular week with 1.3 GB in size, is generated. The examination into the sample data identified the following data problems. Approximately 0.08% of the points have the coordinate X_k or Y_k as zero, 1.14% have the instant speed v_k higher than 130 m/h (the highest driving speed limit in Greece), and 0.04% have duplicates of the same point IDs and timestamps (i.e. $id_k = id_{k+1}$ and $t_k = t_{k+1}$).

Moreover, problems were also revealed regarding the time intervals between consecutive points. Figure 4 describes the distribution of the intervals. It was noted that, while the average is 0.48 min (the typical data recording rate), the intervals vary over a range of 0–3 min, with the highest peak at 0.1 min. However, a second peak at 2 min was also observed, accounting for 2% of the total intervals. Further investigation shows that this peak is mainly caused by points that appear either at highways with some intermediate points being not recorded (e.g. due to bad satellite connections) or *at the border of the study area*. The latter phenomenon occurred when data for the study area are extracted from a larger dataset covering a more extended territory. To extract data from the larger dataset, the data provider just used a rectangle delimiting the study area and removed all GPS points that lie beyond this area. This leads to specific cases of data missing when vehicles temporarily move out of the enclosed area and return back after a few minutes, making artificially long time intervals of two consecutive points that consist of the last position before moving out and first position after returning back. Thus, the points with time intervals longer than (or equal to) 2 min and the OD trips that contain such points are both considered errors.

Based on the above-detected problems, the following rules are extracted: (1) $X_k \neq 0$ and $Y_k \neq 0$; (2) $v_k \leq 130\ km/h$; (3) if $id_k = id_{k+1}$ and $t_k = t_{k+1}$, p_{k+1} is deleted; and (4) if t_{k+1}-$t_k \geqslant 2\ min$, p_{k+1} is deleted.

In the trip extraction, suitable settings for TH_{Act} and TH_{Trip} depend on the research questions. In this study, TH_{Act} is set as

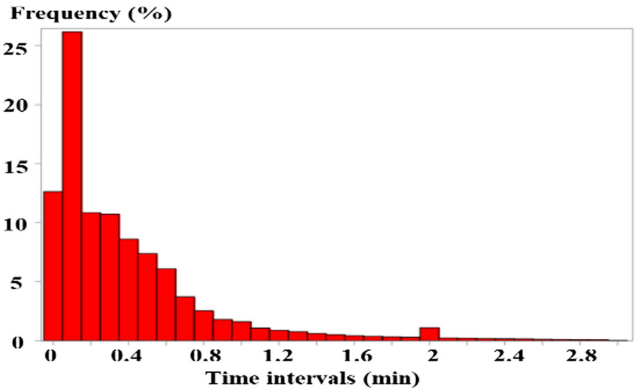


Fig. 4 Distribution of the time intervals between consecutive points

3 min and TH_{Trip} as 100 m. This is based on the data examination as well as the study (Cich et al. 2016) in which optimal values for TH_{Act} and TH_{Trip} are decided through the comparison between trips derived from GPS data and those recorded in the corresponding travel diaries (as ground-truth data). Under the former value, a stop for longer than 3 min is considered a stop location for activities (e.g. parcel deliveries); while according to the latter, the movement longer than 100 m in distance is regarded as a trip with clear destinations (thus reflecting travel patterns). These parameter values, along with the extracted rules, are applied to the entire dataset for data cleaning and trip derivation in SAA. Each of the derived trips is further examined; if the original GPS trace of the trip contains consecutive points with the time intervals not less than 2 min, the trip is removed. Finally, 5.1 million trips were derived from a total of 91 days, with 8.63 trips on average for each vehicle each day.

The left and right of Fig. 5 describe the distributions of the average of v_k and the average number of points per half an hour over all the weekdays, respectively. From the left figure, clear variations in driving speeds across different time periods of the day were revealed, leading to the splitting of a day into 7 periods, including 7–9 am, 9–10:30 am, 10:30 am–15 pm, 15–17 pm, 17–19 pm, 19–21 pm, and 21 pm–7 am. The average speed in each of these periods is 35.6, 37.4, 34.2, 38.8, 40.8, 37.8, and 43.7 (km/h), respectively. According to the right figure, the average number of points over the above-obtained periods is 19.2%, 20.5%, 22.7%, 20.6%, 10.8%, 6.0%, and 0.2%, respectively. Large differences in travel demand are demonstrated among these time periods, underlying the causes for the observed speed deviations.

Traffic Condition, Road Linkage, and Travel Pattern Matrices

Division of the Study Area

The study area is located in the range of longitude (19.870, 26.587) and latitude (35.046, 41.479). During the construction of matrices, the entire area is divided into zones; *Grid_X* and *Grid_Y* decide the total number of the study units. The larger these values are, the higher the spatial resolution reaches, but the less the number of observed GPS points in each zone and between zones. In order to derive results that are statistically sound and representative, these two variables are set as 400, respectively, resulting in a total of 1.6×10^5 zones with each being 1.62 km^2 in size. When this size is compared with other existing studies, it was noted that the average size of the units varies depending on study areas and travel modes. It ranges from 2.14 km^2 in the Twin Cities for car-based transport analysis (Anderson et al. 2013) to 0.15 km^2 in Denizli of Turkey for transit studies (Gulhan et al. 2013), and to 0.03 km^2 and 0.05 km^2 in two towns of Sweden for traffic studies on a combination of bikes and autos (Makrí and Folkesson 1999).

The Matrices

Based on the spatial partition, along with the previously defined 7 temporal periods, three matrices are built, i.e. the traffic condition matrix *TC*(Z_i, *TimeP*, *Day*, *DayT*), road linkage matrix *RL*(Z_i, Z_j, Z_l, *TimeP*, *Day*, *DayT*), and travel pattern- matrix *TP*(Z_o, Z_d, *r*, *TimeP*, *Day*, *DayT*). Z_i, Z_j, Z_l, Z_o, and Z_d represent zones, with *i*, *j*, *l*, *o*, *d* = *1*…*400*; *TimeP* = *1*…*7*; *Day* = *1*…*91* (all the data collection days); and *DayT* = *weekdays*, *weekends*, and *public holidays*. A range of variables is derived from the matrices, including the total number of segments (M_i) in Z_i and average travel speed (V_i) of the segments; the total number of segments (M_{ijl}) successively passing Z_i, Z_j, and Z_l and average speed (V_{ijl}) of the segments; the total number of distinct travel paths (Num_{od}) between Z_o and Z_d, the r^{th} ($r = 1 \ldots Num_{od}$) specific path ($Path_{odr}$); the number of trips (M_{odr}) along the path; and the average travel speed (V_{odr}), distance (L_{odr}), and time (T_{odr}) over the trips. Between different time periods of the day or different types of days, traffic conditions in each zone or between zones could differ to a large extent. In the following, the traffic in the morning (7–9 am) of weekdays is analysed, but the same process can be repeated to the remaining periods and day types. The procedure is identical, but the results are likely to be different.

Figure 6 demonstrates the visualization of *TC* for a part of the study area. Each zone Z_i (on the middle panel) is described as a rectangle (a grid), and the values of the variables (*attributes*) M_i and V_i of the zone are represented by different colours of the grid lines or area. *Conditions* on each attribute

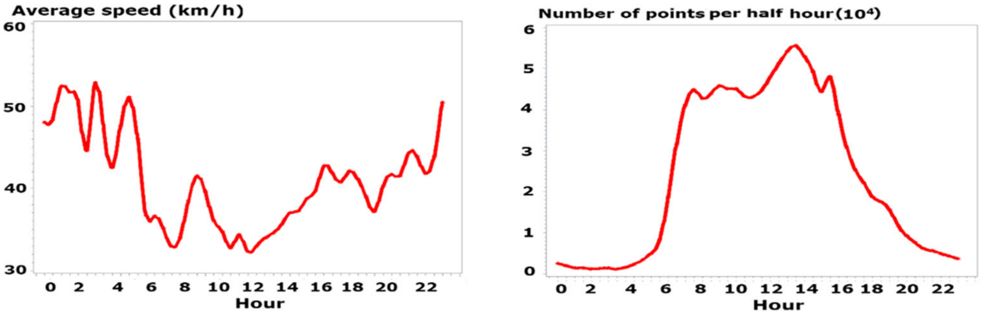


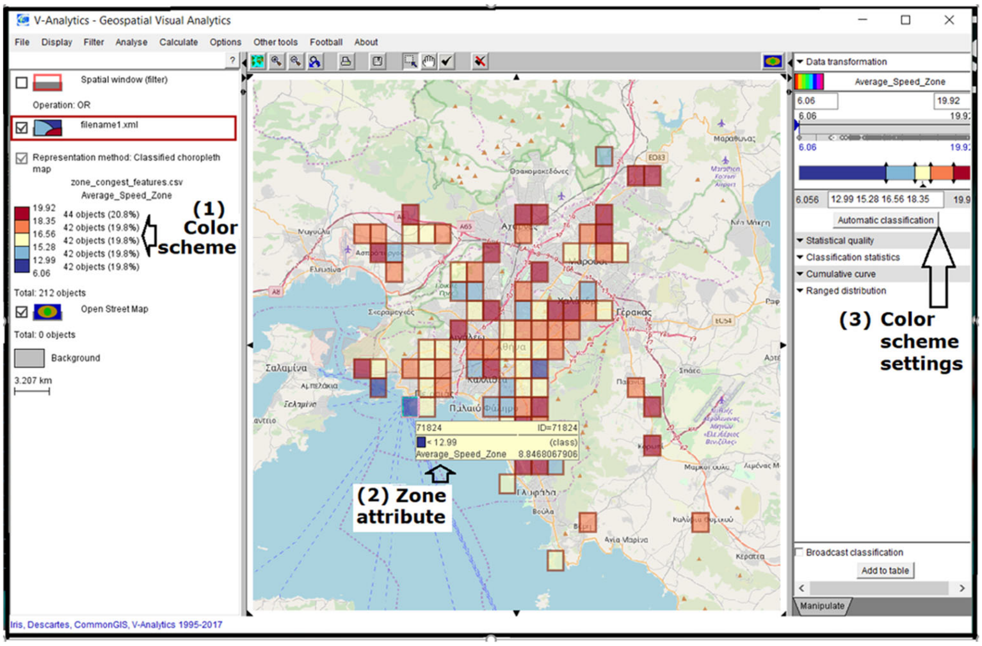
Fig. 5 Distributions of the average of v_k (left) and average number of points (right) over each half an hour

or on the combination of multiple attributes can be specified such that only the zones that satisfy these conditions are shown (e.g. the zones with $V_i < 20$ km/h being visualized in Fig. 6). From the displayed zones, one can be further selected, and the specific attribute values of the zone are presented in an *attribute window*. On the left panel, the colour scheme is described; on the right, the scheme can be re-defined. In this study, only relevant visualization elements are introduced; more details about V-Analytics can be referred to in the literature (Andrienko et al. 2013a).

Figure 7 describes the visualization of *RL*, on which road connections in the surrounding area of a selected zone Z_i (i.e. $Z_{224,189}$, labelled as $Z_{i,j}$ in the black grid) are shown. This (surrounding) area consists of Z_i and its eight adjacent zones (labelled as $Z_{i'j'}$, with $i\text{-}1 \le i' \le i + 1$ and $j\text{-}1 \le j' \le j + 1$). Figure 7 a shows road connections that start from the study zone $Z_{i,j}$, while Fig. 7 b–f illustrate the connections that begin from each of five (out of the eight) surrounding zones, including $Z_{i\text{-}1,j}$, $Z_{i+1,j}$, $Z_{i\text{-}1,j+1}$, $Z_{i,j+1}$, and $Z_{i+1,j+1}$ respectively. On each of these figures, the zone where the connections start is highlighted in the purple grid, and the connection among three zones is expressed as a curve that starts and ends in the first and last zones respectively while passing the second zone. Take the connection $Z_{i,j}$- > $Z_{i,j\text{-}1}$- > $Z_{i+1,j\text{-}1}$ in Fig. 7a as an example, the purple curve describes the connection that begins from the small green circle in $Z_{i,j}$, passes $Z_{i,j\text{-}1}$, and ends at the end (arrow) of this curve in $Z_{i+1,j\text{-}1}$. The colour and/or thickness of the curve reflects the average speed (V_{ijl}) and/or the number of trips (M_{ijl}) along the connected roads. Apart from the black and purple grids, the red grids represent the congestion zones (detected in the "Major Sources of Traffic Flows in the Congestion Zones" section).

rom Fig. 7, road connections starting from a zone and variations in the connections across zones were observed. For example, there are only 2 connections starting from $Z_{i+1,j}$ (Fig. 7c), while 15 and 7 connections originate from $Z_{i\text{-}1,j}$ (Fig. 7b) and $Z_{i,j+1}$ (Fig. 7e) respectively. However, despite multiple connections from a certain zone, there could be a lack of roads in certain directions, e.g. no connections of $Z_{i\text{-}1,j}$- > $Z_{i\text{-}1,j+1}$- > $Z_{i,j+1}$ in Fig. 7b or $Z_{i,j+1}$- > $Z_{i\text{-}1,j+1}$- > $Z_{i\text{-}1,j}$ in Fig. 7e. Furthermore, the three zones involved in a connection may not be geographically adjacent; there could be intermediate zones in between, e.g. $Z_{i\text{-}1,j+1}$- > $Z_{i+1,j+1}$- > $Z_{i+2,j+1}$ (Fig. 7d), $Z_{i,j+1}$- > $Z_{i+1,j+1}$- > $Z_{i+4,j}$ (Fig. 7e), and $Z_{i+1,j+1}$- > $Z_{i+3,j}$- > $Z_{i+5,j}$ (Fig. 7f), with the zone $Z_{i\text{-}1,j+1}$, zones $Z_{i+2,j+1}$ and $Z_{i+3,j}$, and zones $Z_{i+2,j}$ and $Z_{i+4,j}$ being skipped respectively. This has resulted from two consecutive GPS points that skip one (or a few) (geographically adjacent) zone(s) in between these points, indicating a high likelihood that travel speeds of the points are high and that the connected roads are featured with a high speed limit. Further investigation discloses that there is indeed a highway going through the neighbouring zones $Z_{i\text{-}1,j+1}$, $Z_{i,j+1}$, and $Z_{i+1,j+1}$.

All the above observations demonstrate the ability of Fig. 7 in depicting road network situations in the surrounding-area of a selected (congestion) zone $Z_{i,j}$. This information, combined with the major traffic sources of the congestion zone (as displayed in Figs. 11 and 12), becomes an important tool for the identification of mismatch between traffic flows and road network services surrounding the congestion area, enabling further examination into the traffic problems.



□ Spatial window (filter)

Operation: OR

☑ filename1.xml

☑ Representation method: Classified choropleth map

zone_congest_features.csv
Average_Speed_Zone

🟥	19.92	44 objects (20.8%)
🟧	18.35	42 objects (19.8%)
🟨	16.56	42 objects (19.8%)
🟦	15.28	42 objects (19.8%)
🟦	12.99	42 objects (19.8%)
	6.06	

Total: 212 objects

☑ Open Street Map

Total: 0 objects

Background

3.207 km

Iris, Descartes, CommonGIS, V-Analytics 1995-2017

71824 ID=71824
< 12.99 (class)
Average_Speed_Zone 8.8468067906

Average_Speed_Zone
6.06 19.92
6.06 19.92
6.06 19.92
6.056 12.99 15.28 16.56 18.35 19.9

□ Broadcast classification

Fig. 6 Visualization of the traffic condition matrix *TC* (traffic condition layer). *Note*: on the middle panel, each grid represents a zone, with the colour of the grid area reflecting the categories of V_i of the zone. The small (attribute) window indicates the original value of V_i for the selected zone

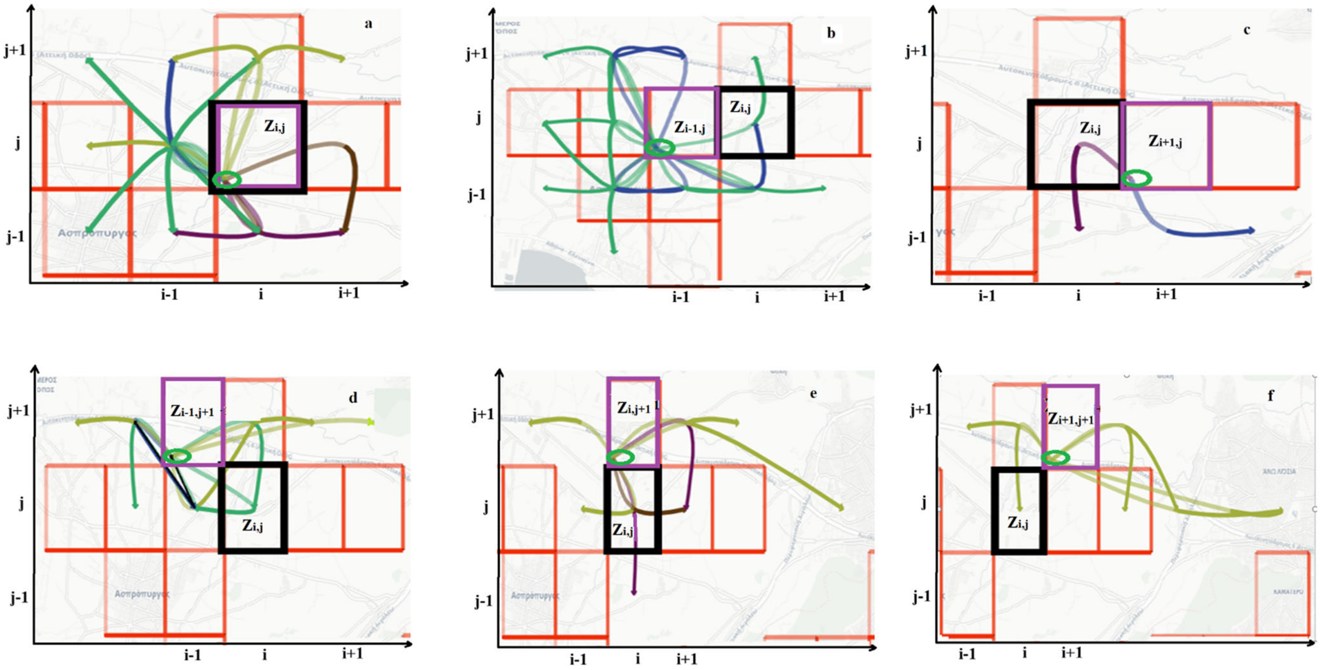


Fig. 7 Visualization of the road linkage matrix *RL* in the surrounding area of $Z_{i,j}$ (road linkage layer). *Note*: the black, purple, and red grids denote the selected zone $Z_{i,j}$ (i.e. $Z_{224,189}$), the zone where the connections start, and the congestion zones, respectively. Each curve describes the connection among 3 zones, starting from the green circle in the 1st zone, passing the 2nd zone, and ending at the end of the curve in the 3rd zone. The colour of the curve reflects V_{ijl} of the roads. Figure 7a-f describe the connections starting from $Z_{i,j}$, $Z_{i-1,j}$, $Z_{i+1,j}$, $Z_{i-1,j+1}$, $Z_{i,j+1}$ and $Z_{i+1,j+1}$ respectively

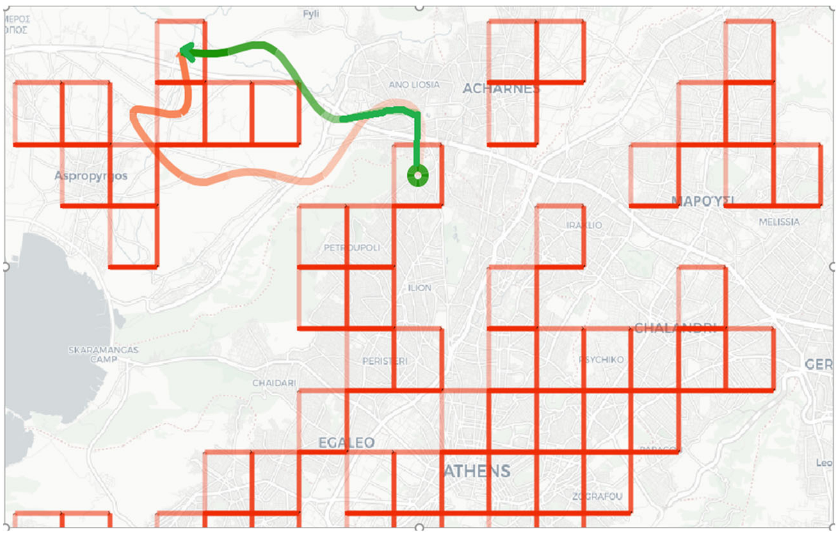
This will be elaborated in the "Other Possible Ways to Alleviate the Congestion" section.

Figure 8 illustrates the visualization of *TP*, on which travel patterns between a selected zone pair Z_o and Z_d are displayed as curves representing all the travel paths (i.e. $Path_{odr}$, $r = 1 \ldots Num_{od}$) between the corresponding zones. The colour and/or thickness of the curves reflects the values of the travel speed (V_{odr}), distance (L_{odr}), and/or travel time (T_{odr}) along these paths.

Congestion Zone Detection

In the detection, TH_{mz}, TH_{vz}, and TH_{con} are designated as 40 20 km/h and 0.8, respectively, based on the literature (Cui et al. 2016a; Zheng et al. 2011) in combination with the data sampling rate and number of sampling vehicles in the experimental data. This leads to 184 zones being filtered out as the congestion zones. These places have undertaken at least 40 trips in the morning per weekday and suffered from the

Fig. 8 Visualization of the travel pattern matrix *TP* (travel pattern layer). *Note*: the purple grids are the selected origin and destination zones $Z_{229,188}$ and $Z_{224,190}$, respectively, and the black grid is $Z_{224,189}$. The curves represent all the (two) travel paths between the OD zones, with the colour denoting the travel speed



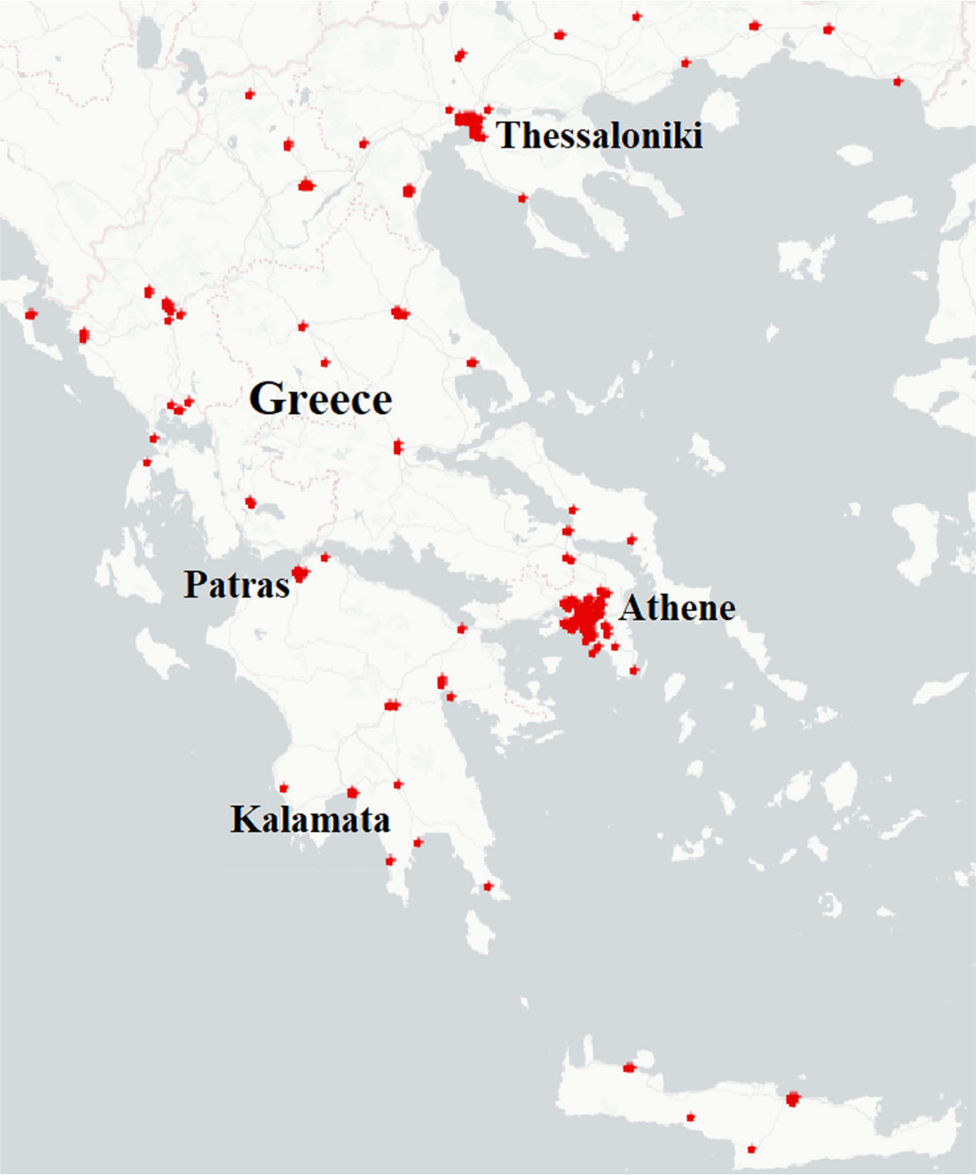


Fig. 9 Geographic distribution of all the congestion zones in the study area (congestion zone layer). *Note*: each red grid represents a congestion zone

average travel speeds lower than 20 km/h at least for 4 in 5 of the days. Figure 9 describes the geographic distribution of all the congestion zones.

Major Sources of Traffic Flows in the Congestion Zones

To analyse traffic flows in the congestion zones, a typical congestion zone $Z_{224,189}$ was used as a demonstration. This zone is located in the commercial area of Athens, and the average speed and number of trips in the morning each day in this area are 18.6 km/h and 201.9, respectively. All the morning trips over all the 65 experimental weekdays are generated by 380 distinct OD pairs. Among these trips, 20.5% both start and end in $Z_{224,189}$, 13.9% and 21.9% only start or end in this zone, and the remaining 43.7% just pass this area.

Let $TH_{od\text{-}con} = 0.05$, 8 OD pairs (2.1% of all the ODs) are filtered out; trips between each of these pairs contribute to at least 5% of the total flows in $Z_{224,189}$. Table 1 presents these ODs, and Fig. 10 further illustrates the common path ($Path_{\text{od-con}}$) for each of these ODs. From these figures, it was noted that the major traffic sources in $Z_{224,189}$ are from the west to this zone (Fig. 10 c and g), from this zone to the east (Fig. 10d), from the east through this zone to the west (Fig. 10 e and f) or to the north (Fig. 10h), and from the north through this zone to the southeast (Fig. 10i).

Table 1 Major sources of traffic flows in $Z_{224,189}$*

OD-ID	$Z_{\text{o-con}}$	$Z_{\text{d-con}}$	$Path_{\text{od-con}}$	$M_{\text{od-con}}$ (%)	Flow
1	224,189	224,189	*(224,189)-(224,189)*	20.5	Start and end
2	221,189	224,189	(221,189)-*(224,189)*	10.4	End
3	224,189	225,189	*(224,189)*-(225,189)	7.1	Start
4	237,188	221,189	(237,188)-(237,187)-(236,187)-(236,186)-(235,187)-(234,187)-(234,188,)-(233, 188)-(232,188)-(231,188)-(230,188)-(230,189)-(229,189)-(228,189)-(227,189)-(226,190)-(225,190)-(224,190)-*(224,189)*-(223,189)-(222,189)-(221,189)	6.5	Pass
5	232,186	221,189	(232,186,)-(232,187)-(233,187)-(232,188)-(231,188)-(230,188)-(230,189)-(229,189)-(228,189)-(227,189)-(226,190)-(225,190)-(224,190)-*(224,189)*-(223,189)-(222,189)--(221,189)	5.3	Pass
6	223,189	224,189	(223,189)-*(224,189)*	5.3	End
7	229,188	224,190	(229,188)-(229,189)-(228,189)-(227,188)-(226,188)-(225,188)-(224,188)-(223,189)-*(224,189)*-(224,190)	5.1	Pass
8	224,190	230,183	(224,190)-*(224,189)*-(223,189)-(223,188)-(222,188)-(222,187)-(222,186)-(223,186)-(223,185)-(224,185)-(225,185)-(226,185)-(227,185)-(227,184)-(228,184)-(229,184)-(229, 183)-(230,183)	5.0	Pass

The columns from the left to right represent the ID of the OD pair, origin zone, destination zone, common travel path, number of trips between the OD zones, and direction of the flows. Zones are identified with the grid numbers in the longitude and latitude dimensions respectively. The study zone (224,189) ($Z_{224,189}$) is in italics

Specific Transport Situations and Possible Ways of Reducing Congestion in $Z_{224,189}$

Alternative Routes

Out of all the major sources listed in Table 1, four (i.e. *OD-IDϵ{4,5,7,8}*) have passed $Z_{224,189}$. The examination into the trips discloses that only one of these sources, i.e. *OD-ID* = *7* from $Z_{229,188}$ to $Z_{224,190}$ ($Path_{\text{od-con}} = Z_{229,188}$-$Z_{229,189}$-$Z_{228,189}$-$Z_{227,188}$-$Z_{226,188}$-$Z_{225,188}$-$Z_{224,188}$-$Z_{223,189}$-$Z_{224,189}$-$Z_{224,190}$) (Fig. 10h), has one additional travel route that bypasses $Z_{224,189}$. This path, i.e. $Path_{\text{odr}} = Z_{229,188}$-$Z_{229,189}$-$Z_{228,189}$-$Z_{227,189}$-$Z_{226,190}$-$Z_{225,190}$-$Z_{224,190}$, is subsequently chosen as the alternative route ($Path_{\text{alter}}$) that could replace the original path. Figure 8 describes this route (the curve in green), and the orange curve on this figure is the original path ($Path_{\text{od-con}}$) that traverses $Z_{224,189}$.

The comparison between $Path_{\text{alter}}$ and $Path_{\text{od-con}}$ delimits the following differences (Table 2). (1) $Path_{\text{alter}}$ has a lower average speed (28.3 km/h) but shorter travel distance (10.9 km) than $Path_{\text{od-con}}$ featured with the speed and travel distance as 35.4 m/h and 12.7 km, respectively. (2) Both routes are mainly comprised of roads with high speed limits. However, due to heavy traffic in the morning rush hour, the average speeds of both routes do not reach their limits. Particularly, the part of $Path_{\text{od-con}}$ inside $Z_{224,189}$ suffers from heavy congestion and low driving speeds (i.e. the average speed of 18.6 km/h), leading to the overall travel time of $Path_{\text{od-con}}$ as 21.5 min. In contrast, although the whole travel time is 23.1 min (1.6 min longer than $Path_{\text{od-con}}$), $Path_{\text{alter}}$ has a stable traffic condition along this path, with the average speed at each zone passed being higher than 20 km/h. (3) Regardless of the above differences, more people have taken $Path_{\text{od-con}}$ than $Path_{\text{alter}}$, with the average number of trips per day as 10.3 (83.7%) along $Path_{\text{od-con}}$ but only 2 (16.3%) through $Path_{\text{alter}}$.

To have a further look into the traffic situations of the two paths, additional analysis is performed on the off-peak period of 9–10:30 am when the average driving speed is 37.4 km/h (*see* Fig. 5). During this period, the speed of $Z_{224,189}$ increases to 22.9 km/h. The traffic conditions of both $Path_{\text{alter}}$ and $Path_{\text{od-con}}$ have also improved, with the average speeds as 47.4 km/h and 57.5 km/h, respectively. This leads to the travel time as 13.8 and 13.3 min, and the number of trips per day as 3.1 (39.7%) and 4.7 (60.3%), respectively. Thus, in relation to the morning rush hour, $Path_{\text{alter}}$ is more closed to $Path_{\text{od-con}}$ in terms of travel time (only 0.5 min difference) and $Path_{\text{alter}}$ also undertakes more trips (23.4% more) in the off-peak period.

From the above analysis, it was noted that the majority (83.7% and 60.3%) of travellers between the OD zones (from

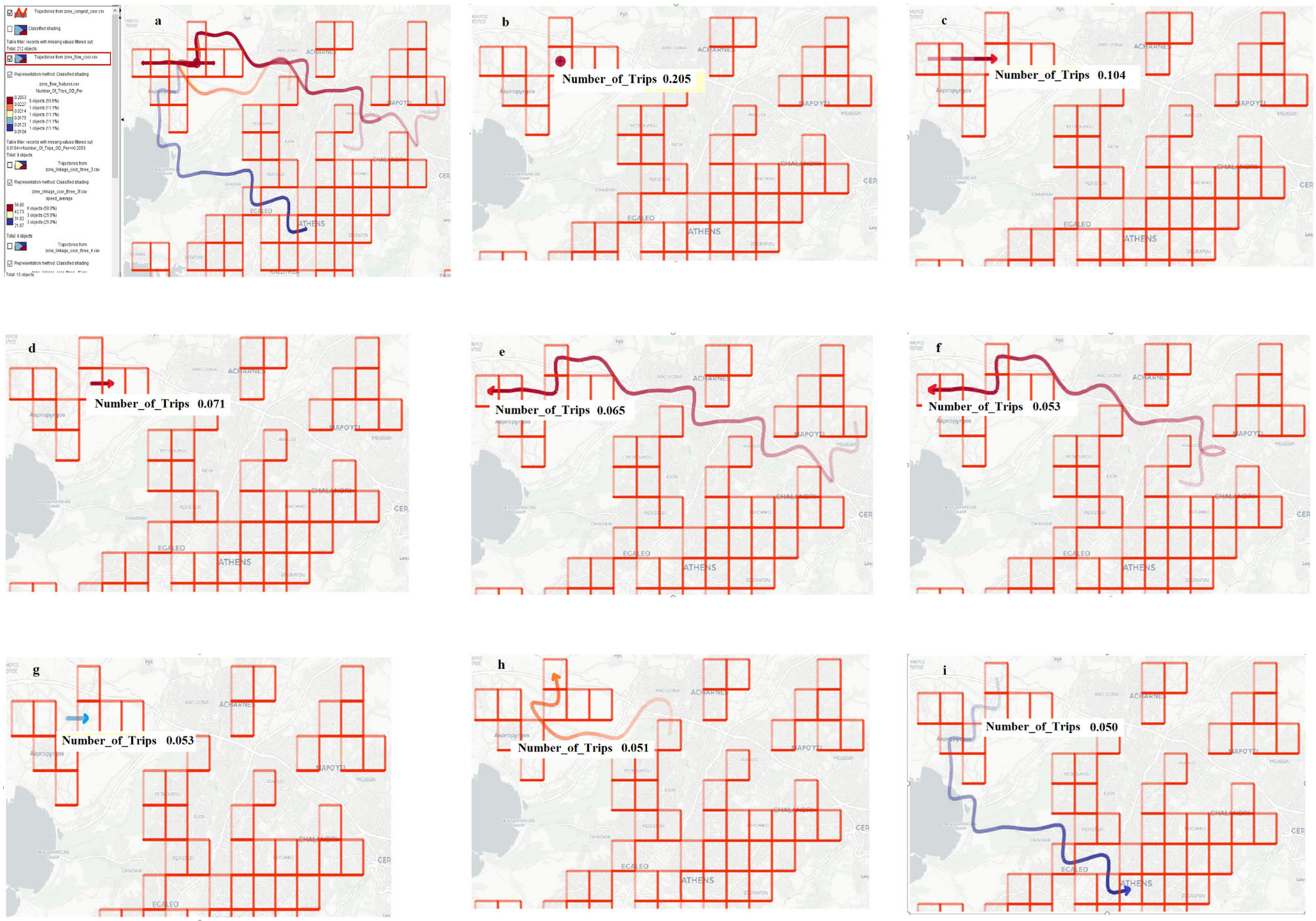


Fig. 10 Major sources of traffic flows in $Z_{224,189}$ (travel flow layer). *Note*: the black and red grids are $Z_{224,189}$ and congestion zones, respectively. The curve describes the typical path $Path_{\text{od-con}}$ and the attribute window lists the value of $M_{\text{od-con}}$. **a** displays all the 8 paths and **b**–**i** depict each of these paths

$Z_{229,188}$ to $Z_{224,190}$) adopt the original path $Path_{\text{od-con}}$ instead of $Path_{\text{alter}}$ (during 7–9 am and 9–10:30 am respectively). Apart from shorter travel time (1.6 min and 0.5 min shorter), there could be other factors contributing to the route choice (e.g. poor connectivity between the high-speed road and local roads inside $Z_{229,188}$ or $Z_{224,190}$ along the path $Path_{\text{alter}}$). Moreover, given that the experimental data are generated from utility vehicles for logistic purposes, the route patterns may not be representative of all trips between the OD zones. Private trips may exhibit different route choices from the observed ones. Thus, before any management decisions (e.g. replacing $Path_{\text{od-con}}$ with $Path_{\text{alter}}$ under certain enforcement measures) are made, a detailed inspection should be conducted regarding why $Path_{\text{od-con}}$ is preferred to $Path_{\text{alter}}$ and what is the representability of the observed route patterns.

Other Possible Ways to Alleviate the Congestion

For the remaining three sources (i.e. $OD\text{-}ID\epsilon\{4,5,8\}$) that pass $Z_{224,189}$, no other trips have been observed between the corresponding OD zones that circumvent the congestion zone. Figure 11 a describes these sources (in the red and blue dash curves), which are from $Z_{237,188}$ to $Z_{221,189}$, from $Z_{232,186}$ to $Z_{221,189}$, and from $Z_{224,190}$ to $Z_{230,183}$, respectively, accounting for 16.8% of the total flows in $Z_{224,189}$. The road connections (7 in total) starting from

Table 2 Comparison between the original path and the alternative route

Route	Number of zones passed	Number of congestion zones passed	Route distance (km)	Average speed (km/h)	Travel time (min)	Number of trips per day
$Path_{\text{alter}}$	5	0	10.9	28.3	23.1	2 (16.3%)
$Path_{\text{od-con}}$	7	1	12.7	35.4	21.5	10.3 (83.7%)

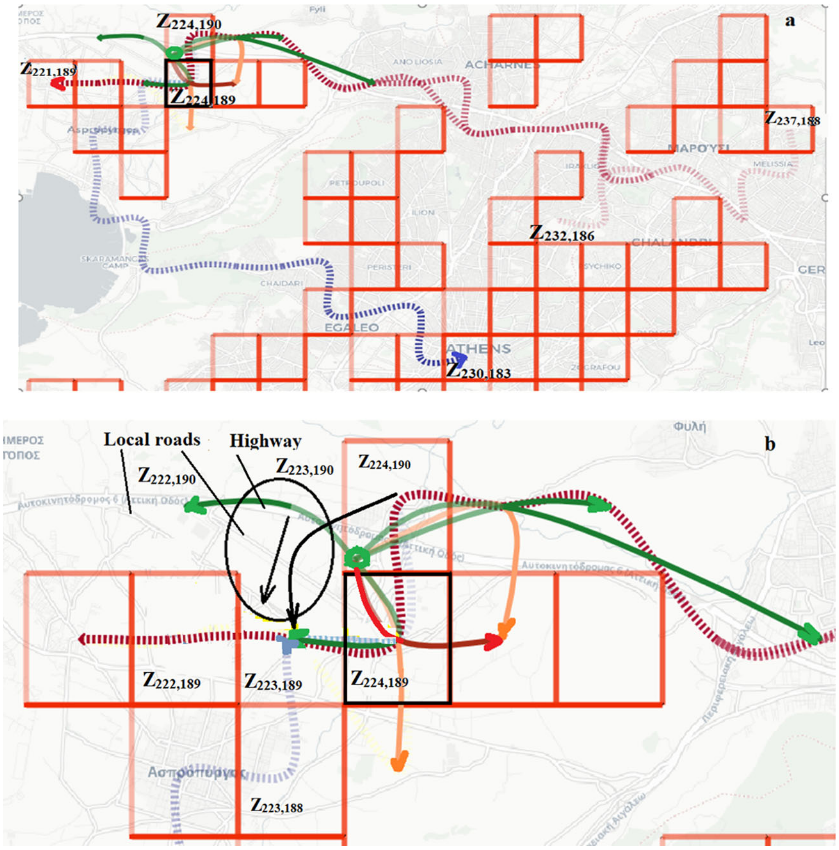


Fig. 11 The three major sources and the road linkage starting from $Z_{224,190}$. *Note*: in both **a** and **b**, the black grid represents $Z_{224,189}$, the dash curves in red and blue are the major traffic sources, and the solid curves in green, red, and orange are the road connections starting from $Z_{224,190}$. The black curve, circle, and arrow in **b** denote the missing link, the area containing the mismatch, and the proposed new road, respectively

$Z_{224,190}$ on the road linkage layer (described in Fig. 7e) are also displayed (in the green, red, and orange solid curves). From the zoomed-in map in Fig. 11b, it was observed that the connections begin from $Z_{224,190}$ and go to the northwest, northeast, southeast, and south respectively. However, no road connections are present towards the direction of the southwest, i.e. $Z_{224,190}$- > $Z_{223,190}$- > $Z_{223,189}$ (*the missing-link*) (shown in the black curve). This causes all the trips from these three sources taking the connection $Z_{224,190}$- > $Z_{224,189}$- > $Z_{223,189}$ (i.e. traversing $Z_{224,189}$ in order to reach $Z_{223,189}$ from $Z_{224,190}$). Due to the missing link, a mismatch is formed between the road connectivity (among the three zones $Z_{224,190}$, $Z_{223,190}$, and $Z_{223,189}$ in the corresponding direction) and these major flows. Further inspection reveals that there are roads available in the area of these zones (shown in the black circle), including a highway and several local roads. But there is a lack of connections between the high way and local roads inside $Z_{223,190}$. These roads are only linked in a further area of $Z_{222,190}$. This leads to none of the trips of these sources choosing the highway and local roads in order to get to $Z_{223,189}$ from $Z_{224,190}$ while avoiding $Z_{224,189}$ (i.e. none of the trips taking the connection $Z_{224,190}$- > $Z_{223,190}$- > $Z_{222,190}$- > $Z_{223,189}$). Thus, if new roads (shown in the black arrow) were built to connect the highway and local roads inside $Z_{223,190}$, the missing link ($Z_{224,190}$- > $Z_{223,190}$- > $Z_{223,189}$) could be generated, enabling these major flows to be diverted from the original connection ($Z_{224,190}$- > $Z_{224,18}$- > $Z_{223,189}$) to the new link, therefore bypassing $Z_{224,189}$.

Similar situations occur to the source of *OD-ID* = *7*. Apart from the alternative route identified in the "Alternative Routes" section, there could be other ways to divert the corresponding trips. Figure 12 a and b demonstrate this source (in the red dash curve from $Z_{223,189}$ to $Z_{224,190}$); the road linkage starting from $Z_{223,189}$ (depicted in Fig. 7b) is also shown. It was noticed that, despite various (15 in total) connections from $Z_{223,189}$, an important connection towards the northeast is missing, i.e. $Z_{223,189}$- > $Z_{223,190}$- > $Z_{224,190}$ (the missing link) (represented in the black curve in Fig. 12b). This leads to the corresponding trips passing $Z_{224,189}$ in order to reach $Z_{224,190}$ from $Z_{223,189}$ (i.e. taking the connection $Z_{223,189}$- > $Z_{224,189}$- > $Z_{224,190}$). As in Fig. 11b, a mismatch exists in the opposite direction of the road connectivity among the three zones $Z_{223,189}$, $Z_{223,190}$, and $Z_{224,190}$ (Fig. 12b). If new roads (denoted in the black arrow) were constructed connecting the highway and the local roads inside $Z_{223,190}$, the missing link ($Z_{223,189}$- > $Z_{223,190}$- > $Z_{224,190}$) could be created, enabling the trips to take the route along the new link, thus avoiding $Z_{224,189}$.

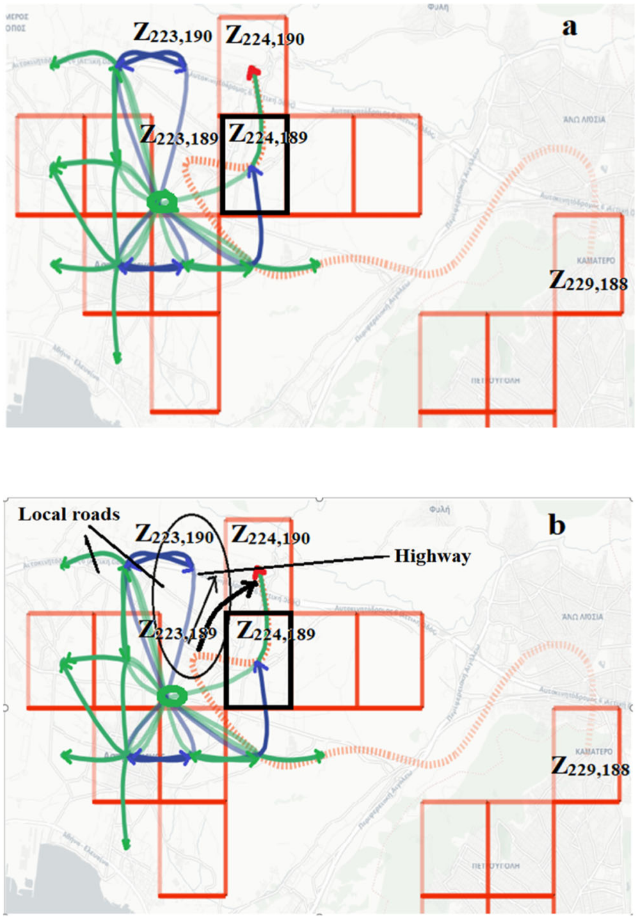


Fig. 12 The major source of OD-ID = 7 and the road linkage starting from $Z_{223,189}$. *Note*: in both **a** and **b**, the black grid represents $Z_{224,189}$, the red dash curve is the source, and the solid curves in green and blue are the road connections starting from $Z_{223,189}$. The black curve, circle, and arrow in **b** refer to the missing link, the area for the mismatch, and the proposed new road, respectively

Discussions and Conclusions

Due to the rapid development of technologies, an extensive growth in big mobility data exists and is continuously expanding. These technologies offer a solution to the challenges associated with conventional travel data. However, obtaining significant mobility knowledge from the data and using the knowledge for supporting traffic management decisions require detailed but efficient data analysis as well as good communication and exploration of results. Towards this perspective, this paper has developed a method for citywide traffic analysis based on the combination of visual and analytical approaches. Applications to large volumes of GPS data have demonstrated the potential and effectiveness of the approach in achieving such a goal.

The method and the results that arise in this study will help traffic managers improve the understanding of current traffic (e.g. through the traffic condition layer, travel pattern layer, and congestion zone layer), its major sources (through the traffic flow layer), and mismatch between traffic flows and road network services (through the road linkage layer). Based on the results, further investigation can be performed in order to design more effective policy measures. To this end, the proposed method does not only identify traffic problems, but more importantly, it provides a platform (by means of the traffic condition map) to assist managers in further exploring these problems and making appropriate decisions. Particularly, through the combination of the traffic flow layer and travel pattern layer (e.g. in Fig. 8), alternative routes ($Path_{\text{alter}}$) that start and end in the same places as the OD zones ($Z_{\text{o-con}}$ and $Z_{\text{d-con}}$) of the major traffic source of a congestion zone (Z_{con}) can be identified. Moreover, by overlaying the traffic flow layer and road linkage layer (e.g. in Figs. 11 and 12), a mismatch between road connectivity and major sources of traffic flows in Z_{con} is revealed. Based on the mismatch, possible ways (e.g. new roads) could be further suggested in alleviating traffic problems. Note that all the above analysis is performed in the VAA component, allowing interactive inspection into the traffic situations, facilitating managers gaining more insights into the problems and discoveries of new policy measures. Moreover, all the visualized results are derived from large volumes of GPS data, ensuring more objective and representative of the detected problems, laying a basis for the design of more effective measures.

A number of areas lie ahead for future improvement. Firstly, the traffic situations (e.g. the average speed V_i and major traffic source $Path_{\text{od-con}}$ of a congestion zone, and the travel pattern $Path_{\text{odr}}$ between two zones) could be dynamically visualized over the course of a day based on temporal information (e.g. across different time periods). Secondly, in this study, the transport situations concerning only passing trips (e.g. $OD\text{-}ID\epsilon\{4,5,7,8\}$) are examined and the ways of diverting the trips are investigated. Further work should be done regarding the other major sources that start or end in (rather than passing) the congestion zone Z_{con}. The goal is not to divert the trips from Z_{con}, but to examine if the capacity of the road connections between Z_{con} and its adjacent zones (e.g. the connections $Z_{224,189}$- > $Z_{225,189}$ in Fig. 10b and $Z_{223,189}$- > $Z_{224,189}$ in Fig. 10 c and g) sufficiently serve these major flows. Thirdly, out of all the congestion zones detected in the study area, 45 (21.2% of the total) are geographically distributed as single ones, 57 (26.7%) have 2 or 3 congestion zones connected, and the rest (51.9%) form large congestion areas with more than 3 zones (most of which are in the city centre). In this study, only a single congestion zone is analysed. In the future, the method could be applied to a larger place consisting of multiple congestion zones.

Fourthly, in the method a variety of parameters has been defined; the performance of the approach thus depends on the specification of these parameters, particularly the following ones. (1) The minimum duration (TH_{Act}) and distance

(TH_{Trip}) above which a stop and a trip are selected respectively; (2) the number of zones (*Grid_X* and *Grid_Y*) of the study area; (3) the minimum number of trip segments (TH_{mz}), lowest speed in normal conditions (TH_{vz}), and smallest probability (TH_{con}) which determine whether a zone is considered experiencing congestion; and (4) the minimum percentage ($TH_{\mathrm{od\text{-}con}}$) of trips between a pair of OD, above which the pair is selected as a major source of the traffic in Z_{con}. In specifying these parameters, higher values of TH_{Act} and TH_{Trip} would miss activities for a short duration (e.g. parcel deliveries) and trips with short distances (e.g. between geographically closed locations), but lower values may pick up non-activities (e.g. stops at traffic lights) and non-trips (e.g. movement under bridges), respectively. Furthermore, the combination of higher values of *Grid_X*, *Grid_Y*, TH_{mz}, TH_{con}, and $TH_{\mathrm{od\text{-}con}}$ and lower values of TH_{vz} would lead to the derived results in a more detailed spatial resolution, more statistically representative, and the detection of more severe congestion problems. But the stricter parameter values also call for more GPS data, thus requiring a higher data sampling rate and/or more sampling vehicles. In future research, investigation should be conducted regarding how and to what extent the derived results are influenced by these parameters and what is the minimum amount of GPS data needed for various combinations of the parameter values. This will provide an important guideline for the parameter selection issue.

Fifthly, in deriving the travel distance of a trip segment, the *geodetic distance* (d_{mn}) is used, which is the sum of the geographic distance ($D(p_k, p_{k+1})$) between two adjacent GPS points along the path (*see* Formula 1). The geodetic distance has also been adopted in a number of studies (e.g. Cui et al. 2016b; Zheng et al. 2011) to estimate distances of trips based on GPS data. When this distance is compared with the *network distance* (i.e. the length of the shortest path between two locations along the transport network), the former is fast and easy to compute, while the latter requires more time and complex analysis due to the utilization of network analysis techniques and geographic information systems (e.g. Cubukcu and Taha 2016; Okabe et al. 2006). Moreover, further examination into the experimental data reveals that the average difference between $D(p_k, p_{k+1})$ and the actual distance of the adjacent points (after map matching to specific roads) is 0.014 km, leading to the deviation between the average speed (derived from $D(p_k, p_{k+1})$) and the actual speed of the two points as 1.75 km/h (given the data sampling rate of 0.48 min). This suggests that the actual speeds are typically 1.75 km/h higher than the derived results, including v_{mn} (Formula 1), V_i (Formula 2), V_{ijl} (Formula 3), and V_{odr} (Formula 4) and that the actual threshold TH_{vz} for congestion detection is 1.75 km/h higher than the designated value (20 km/h). Thus, while geodetic and network distances produce different absolute values, they do not affect much the detection of congestion (i.e. with the actual threshold TH_{vz} being increased by a certain amount). The use of geodetic distances can generate sufficiently accurate results (Ivis 2006; Zheng et al. 2011); the higher the data sampling rate, the closer the two distances, and the more accurate the results are. Nevertheless, in order to obtain even more precise results, the network distance could be considered in the future, as the transport network topology varies across different areas. The results could be compared with the ones in the current study, and variations between them could be delimited.

Lastly, the method primarily focuses on systematically examining traffic conditions across the urban area, with the goal of uncovering traffic problems that are persistent over a long time. Due to the volume of the GPS data utilized and the scale of the analysis, the method is normally implemented off-line. However, the approach could also be applied to real-time analysis on a local area (in small sizes), through the construction and visualization of the matrices (*TC*, *RL*, and *TP*) and the detected congestion zones (alongside their major traffic sources) in this area for a short period (e.g. $TimeP = 5$ or 10 min) of the current day. Due to the small size and short time interval, the amount of data being analysed is considerably reduced, leading to the update of the computed results and corresponding map layers in a short time (e.g. 5 or 10 min), enabling near real-time analysis.

Funding Information The work was sponsored by the EU Horizon 2020 research and innovation program Grant 780754.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This paper does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Anderson P, Levinson D, Parthasarathi P (2013) Accessibility futures. Transactions in GIS 17(5):683–705

Andrienko G, Andrienko N, Bak P, Keim D, Wrobel S (2013a) Visual analytics of movement. Springer, Berlin

Andrienko G, Andrienko N, Hurter C, Rinzivillo S, Wrobel S (2013b) Scalable analysis of movement data for extracting and exploring significant places. IEEE Trans Vis Comput Graph 19(7):1078–1094. https://doi.org/10.1109/TVCG.2012.311

Andrienko G, Andrienko N, Fuchs G (2016) Understanding movement data quality. Journal of Location based services 10(1):31–46. https://doi.org/10.1080/17489725.2016.1169322

Andrienko G, Andrienko N, Chen W, Maciejewski R, Zhao Y (2017) Visual analytics of mobility and transportation: state of the art and further research directions. IEEE Trans Intell Transp Syst 18(8): 2232–2249. https://doi.org/10.1109/TITS.2017.2683539

Batty M (2013) The new science of cities. MIT Press

Chae JH, Thom D, Jang Y, Kim SY, Ertl T, Ebert DS (2014) Public behaviour response analysis in disaster events utilizing visual analytics of microblog data. Computers &Graphics, 38:51–60

Chen S, Andrienko G, Andrienko N, Doulkeridis C, Koumparos A (2019, 2019) Contextualized analysis of movement events. EuroVis Workshop on Visual Analytics

Cich G, Knapen L, Bellemans T, Janssens D, Wets G (2016) Threshold settings for TRIP/STOP detection in GPS traces. J Ambient Intell Human Comput 7:395–413

Cubukcu KM, Taha H (2016) Are Euclidean distance and network distance related? Asia-Pacific International Conference on Environment-Behavior Studies, Edinburgh University 2016. https://doi.org/10.21834/e-bpj.v1i4.137

Cui JX, Liu F, Hu J, Janssens D, Wets G, Cools M (2016a) Identifying mismatch between urban travel demand and transport network services using GPS data: a case study in the fast growing Chinese city of Harbin. Neurocomputing 181:4–18

Cui JX, Liu F, Janssens D, An S, Wets G, Cools M (2016b) Detecting urban road network accessibility problems using taxi GPS data. J Transp Geogr 51:147–157

Dias KL, Pongelupe MA, Caminhas WM, de Errico L (2019) An innovative approach for real-time network traffic classification. Comput Netw 158:143–157

Gulhan G, Ceylan H, Özuysal M, Ceylan H (2013) Impact of utility-based accessibility measures on urban public transportation planning: A case study of Denizli, Turkey. Cities 32:102–112

Itoh M, Yokoyama D, Toyoda M, Tomita Y, Kawamura S, Kitsuregawa M (2016) Visual exploration of changes in passenger flows and tweets on mega-city metro network. IEEE Trans, Big Data 2(1): 85–99. https://doi.org/10.1109/TBDATA.2016.2546301

Ivis F (2006) Calculating geographic distance: concepts and methods. Proc. 19th Conf. Northeast SAS User Group 2006

Lansley G, Longley P (2016) The geography of twitter topics in London. Comput Environ Urban Syst 58:85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Liu H, Jin SC, Yan YY, Tao YB, Lin H (2019) Visual analytics of taxi trajectory data via topical sub-trajectories. Visual Informatics 3: 140–149

Makrí MC, Folkesson C (1999) Accessibility measures for analyses of land use and travelling with geographical information systems. In Proceedings of 2nd KFB-Research Conference 1–17

Markovic N, Sekula P, Vander Laan Z, Andrienko G, Andrienko N (2019) Applications of trajectory data from the perspective of a road transportation agency: literature review and Maryland case study. IEEE Trans Intell Transp Syst 20:1858–1869. https://doi.org/10.1109/TITS.2018.2843298

Okabe A, Okunuki KI, Shiode S (2006) The SANET toolbox: new methods for network spatial analysis. Trans GIS 10(4):535–550

Pope P (2017) Big data analytics with SAS. https://www.bol.com/nl/p/big-data-analytics-with-sas/9200000084601370/?country=BE

Robinson A (2017) Geovisual analytics. In: Wilson JP (ed) The Geographic Information Science & Technology Body of Knowledge (3rd Quarter 2017 Edition). https://doi.org/10.22224/gistbok/2017.3.6

Ruan ZY, CSong CC, Yang XH, Shen GJ, Liu Z (2019) Empirical analysis of urban road traffic network: a case study in Hangzhou city, China. Physica A 527:121287

Skupin A (2013) A visual exploration of mobile phone users, land cover, time, and space. Pervasive Mobile Comput 9:865–880

Thomas J, Cook K (2005) Illuminating the path: the research and development agenda for visual analytics. IEEE

Tuszyńska B (2018) Transport and tourism in Greece. Directorate-General for Internal Policies of the Union. Publications Office of the European Union. https://publications.europa.eu/en/publication-detail/-/publication/ced95429-6548-11e8-ab9c-01aa75ed71a1

Xie C, Li MK, Wang HY, Dong JY (2019) A survey on visual analysis of ocean data. Visual Informatics 3:113–128

Yuan NJ, Zheng Y, Zhang LH, Xie X (2011) T-finder: a recommender system for finding passengers and vacant taxis, Proceedings of the 13th ACM international conference on ubiquitous computing (UbiComp 2011)

Zeng JW, Qian YS, Yu SB, Wei XT (2019) Research on critical characteristics of highway traffic flow based on three phase traffic theory. Physica A 530:121567

Zheng Y, Liu YC, Yuan J, Xie X (2011) Urban computing with taxicabs. UbiComp'11, Beijing, China

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.