Journal of Transport Geography 83 (2020) 102658

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling

Alexandros Sfyridis, Paolo Agnolucci*

Institute for Sustainable Resources, University College London, 14 Upper Woburn Place, WC1H 0NN London, United Kingdom

ARTICLE INFO

Keywords:
Annual Average Daily Traffic (AADT)
Clustering
K-prototypes
Support Vector Regression (SVR)
Random Forest
GIS

ABSTRACT

Collection of Annual Average Daily Traffic (AADT) is of major importance for a number of applications in road transport urban and environmental studies. However, traffic measurements are undertaken only for a part of the road network with minor roads usually excluded. This paper suggests a methodology to estimate AADT in England and Wales applicable across the full road network, so that traffic for both major and minor roads can be approximated. This is achieved by consolidating clustering and regression modelling and using a comprehensive set of variables related to roadway, socioeconomic and land use characteristics. The methodological output reveals traffic patterns across urban and rural areas as well as produces accurate results for all road classes. Support Vector Regression (SVR) and Random Forest (RF) are found to outperform the traditional Linear Regression, although the findings suggest that data clustering is key for significant reduction in prediction errors.

1. Introduction

Annual average daily traffic (AADT) is a measure of road traffic flow, defined as the average number of vehicles at a given location over a year[1] (Roess et al., 2011). AADT data are mainly collected by Automatic Traffic Counters (ATCs) where passing vehicles are monitored on a 24 h basis and are used for a number of applications in road transport studies, such as accident prediction (Çodur and Tortum, 2015), GHG emission estimation (Puliafito et al., 2015) and noise exposure estimation (Morley and Gulliver, 2016). AADT values are also fundamental for road construction, planning, maintenance and pavement design studies (Leduc, 2008). However, ATCs are normally not integrated throughout the road network. In the UK – as in most countries – ATCs are only installed at selected locations on major roads covering only a fraction of the network. Minor road counts are undertaken at selected locations as well, although counting is conducted manually. Moreover, manual counts for major and minor roads are undertaken seasonally and adjustments to estimate AADT are applied (Department for Transport, 2013).

Lack of traffic count measurements across the road network, underlines the need for a method to estimate these values as accurately as possible at all possible locations on the road network. To date, a number of attempts has been made, although research on AADT estimation exhibits limitations in several aspects. First of all, studies are usually limited within the boundaries of urban areas (e.g. Doustmohammadi and Anderson, 2016; Kim et al., 2016), or on particular types of roads (e.g. Caceres et al., 2012). Secondly, most studies estimate AADT on major roads, while minor roads are repeatedly excluded, with only a few studies incorporating them (e.g. Apronti et al., 2016; Morley and Gulliver, 2016). Finally, the majority of statistical models incorporates limited explanatory variables – so that many potentially affecting factors are not taken into account.

In this paper we present a methodology, based on Machine Learning and standard statistical methods, which can accurately estimate AADT values for all road types (major and minor) for the whole road network. We exclude motorways from the analysis considering that traffic on these roads is not directly affected from its surrounding characteristics (Eom et al., 2006; Zhao and Chung, 2001). The methodology, explores a comprehensive set of driving factors, while addressing the identified limitations of the modelling implemented so far. In order to do so, the method extracts information from a number of spatial and non-spatial datasets from different sources, with the datasets being manipulated in a GIS environment. England and Wales in the UK are used as an empirical study to demonstrate the methodology. The method can be used to provide outputs sat different geographical scales, so it can be used both for macro and micro analyses. As a word of clarification, it should be noted, that AADT estimation studies can generally be divided into current-year and future-year estimations (Castro-Neto et al., 2009), with the former using data from existing traffic counters to develop

* Corresponding author.
E-mail addresses: alexandros.sfyridis.15@ucl.ac.uk (A. Sfyridis), p.agnolucci@ucl.ac.uk (P. Agnolucci).

[1] AADT is given by (Leduc, 2008): $AADTi = \sum_{j=1}^{365} \frac{TC_{i,j}^{24}}{365}$, where $TC_{i,j}^{24}$ is the 24 h traffic count on road link i at day j.

https://doi.org/10.1016/j.jtrangeo.2020.102658
Received 9 July 2019; Received in revised form 24 January 2020; Accepted 25 January 2020

models capable of estimating AADT at locations where counts are not available when new data are used (Selby and Kockelman, 2013) and the latter incorporating historical traffic data, aiming to estimate short or long term future AADTs at the same locations. Our approach focuses on the former, where a model is developed and applied on data from existing traffic counters, so as to test its accuracy and potential application on street segments where counters are not available.

The paper is presented in six sections. Section 2 provides a review of the AADT estimation approaches found in the literature. Section 3 describes the datasets used, while section 4 presents the methodology for creating the variables and estimating AADTs for all roads. Finally, section 5 presents the results and section 6 concludes and discusses the findings and potential future works to improve our approach.

2. Literature review

AADT estimation is not a novel concept with analyses conducted for over 30 years now (e.g. Neveu, 1983; Fricker and Saha, 1987). To date, a number of approaches has been tested using known traffic counts and incorporating additional explanatory variables for prediction. In this section, the most common modelling approaches as well as identified characteristics potentially influencing traffic flows are presented.

2.1. AADT estimation models

In the literature, three principal approaches to estimate AADT can be identified, namely Linear Regressions, Spatial Statistics and several Machine Learning (ML) techniques. Linear regression models have been the most popular in the literature, with applications ranging from the very early to the latest stages of AADT estimation. For example, Mohamad et al. (1998) applied linear regression with 11 predictors in county roads in Indiana and Xia et al. (1999) used a linear regression model from a sample of 450 count stations in Florida. Zhao and Chung (2001) extended previous research by using a larger dataset, incorporating land use and accessibility variables in Florida. More recently, Doustmohammadi and Anderson (2016) applied linear regression with land use data in two small and medium sized cities in Alabama and Pun et al. (2019) applied a multiple linear regression model in Hong Kong.

Evolution in the field of spatial statistics and development of spatial datasets has led to application of spatial methods for AADT estimation. In these models spatial location and correlation are taken into account so that data points are weighted according to their distance from the location where the dependent variable is to be estimated (Loyd, 2007). For example, Wang and Kockelman (2009) applied Kriging interpolation with Euclidean distances among traffic count stations in Texas while Kriging with additional covariables (CoKriging) has been applied by Eom et al. (2006), Selby and Kockelman (2013) and Kim et al. (2016) in North Carolina, Texas and South Korea respectively. Geographic Weighted Regression (GWR) to estimate AADT has been used by Zhao and Chung (2001) and Zhao and Park (2004) in Florida and Selby and Kockelman (2013) in Texas.

More recently, application of Machine Learning (ML) algorithms has reached AADT estimation studies. To our knowledge, applications are mainly focused on production of AADT predictions based on historical data, while ML use has been scarce for estimation at unmeasured locations. ML applications can be found in Shojaeshafiei et al. (2017), where the K-STAR (K*) and Random Forest (RF) algorithms are applied to estimate AADT in Alabama, and in Fu et al. (2017) where Artificial Neural Networks (ANNs) are used to estimate AADTs in the Republic of Ireland. The latter also appears to be the first study attempting to extend the study area to country level. Cluster analyses have been conducted by Gecchele et al. (2011) in Venice, Italy and by Caceres et al. (2018) for intercity roads in Spain, although the former focuses on temporal pattern identification and does not take into account other characteristics. Finally, Das and Tsapakis (2019) also apply RF, Support Vector Regression (SVR) and K nearest neighbour (knn) in Vermont.

2.2. Drivers of road traffic flows

In order to estimate AADT several attributes have been explored by various studies so that the explanatory variables (predictors) identified in the literature can be classified in four major categories: roadway characteristics, socioeconomic, land use, public transport and parking facilities.

Roadway attributes, are related to various characteristics of the road at the location where the traffic count point is placed. For example, Zhao and Park (2004), Doustmohammadi and Anderson (2016) and Shojaeshafiei et al. (2017) have incorporated road class and number of lanes, while speed limits for street segments have been used by Selby and Kockelman (2013) and Fu et al. (2017). Apronti et al. (2016) incorporate type of road surface to distinguish between paved and unpaved roads in low volume roads in Wyoming, taking into account the large number of unpaved roads in the state. The same study, also incorporates a highway accessibility variable applied to low volume roads which is found in a number of previous studies as well (e.g. Mohamad et al., 1998; Selby and Kockelman, 2013; Xia et al., 1999; Zhao and Park, 2004), although applied to capture connectivity of higher class roads with motorways. In addition, Sarlas and Axhausen (2014) take into account road density in the vicinity of traffic count points. Other studies have also introduced topological roadway characteristics for traffic flow analysis, such as the degree (e.g. Jiang and Liu, 2009; Pun et al., 2019) and several centrality measures (e.g. Gao et al., 2013; Jayasinghe et al., 2015; Zhao et al., 2017).

Socioeconomic characteristics are the most common attributes used in the literature, being taken into account in almost all studies. Population of local settlements is considered by Fu et al. (2017) and Selby and Kockelman (2013) and number of households and household income is used by Eom et al. (2006). Apronti et al. (2016) used employment and per capita income, while other variables used in applied studies include age of population, gender balance in the population and car ownership (e.g. Cervero and Kockelman, 1997; Stead, 2001; Zhao and Chung, 2001; Zhang, 2007; Aditjandra et al., 2012). Jahanshahi and Jin (2016) state that car ownership influences traffic volumes, although its influence varies across areas. However, one has to bear in mind that car ownership is strongly connected to household income (Silva et al., 2012).

Land use variables indicate the surrounding environment at location where the count point is placed, with the majority of studies mainly distinguish the count points at either being on urban or rural areas (e.g. Eom et al., 2006; Fu et al., 2017; Zhao and Chung, 2001; Zhao and Park, 2004). However, other studies have introduced more detailed land use classification. For example, Xia et al. (1999) classified land use by introducing business, residential and fringe areas while Kim et al. (2016) used commercial, residential, industrial and miscellaneous classification. Apronti et al. (2016) refined this approach by introducing a more detailed classification, considering several types of agricultural land use, forest and recreational sites among others.

Public transport variables are almost entirely absent from AADT studies with only Sarlas and Axhausen (2014) incorporating density of public transport stops in the vicinity of traffic count points. On the other hand, behavioural studies have investigated the impact of public transport supply on mode choice and road traffic. Cervero (1994) finds that residents near rail stations are more likely to use public transport, which leads, to a reduction in private road traffic. Stead (2001) discovered that bus frequencies impact travelled distances by individuals and mode choice, albeit findings differ across geographic area. Aditjandra et al. (2012) also conclude that public transport accessibility reduces individual driving. The influence of parking availability and costs in AADT is also absent from all studies we have seen in the

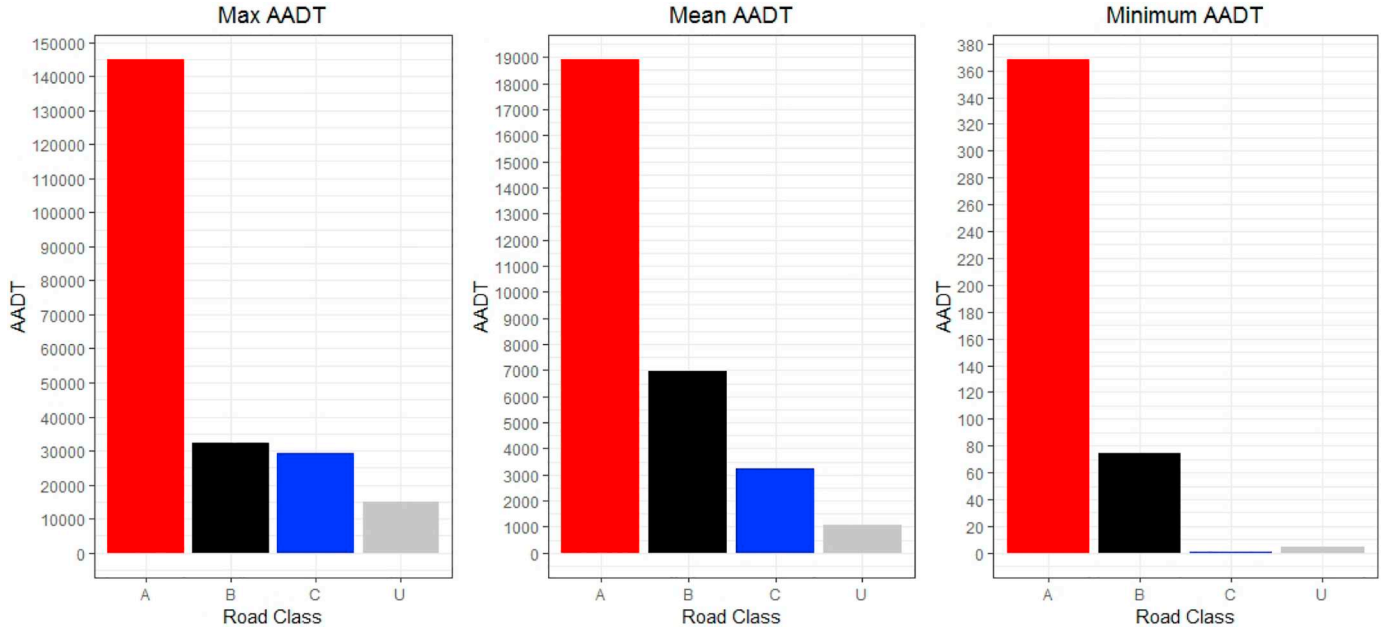

Fig. 1. Maximum (left), average (centre) and minimum (right) AADT values for all road types.

literature. On the other hand, parking impact on mode choice and therefore road traffic, is an established area of research in behavioural studies, where Hess (2001) and Zhang (2007) conclude that availability of parking encourages individual car use. In addition, there is a considerable literature about the impact of parking as a pulling factor for traffic, especially in those cases where free or low cost parking is available, as these factors can generate traffic in the area where it is provided and the surrounding areas (e.g. Arnott and Inci, 2006; Arnott and Williams, 2017; Inci et al., 2017; Kelly and Clinch, 2009; Shoup, 2006).

3. Data

For the purposes of our research, a number of spatial and non-spatial datasets have been extracted from different sources and manipulated within a GIS environment to facilitate feature design described in section 4.1. Selection of specific datasets is based on the identified factors in section 2.2, although not all variables used in the literature are available for the UK. For example, number of lanes and speed limits are not provided in the spatial road network dataset. As incorporating more variables into a Machine Learning algorithm has the potential to improve performance (Domingos, 2012), we use variables, potentially affecting AADT, additional to those considered in the literature so far.

Traffic count points were derived from the UK's Department for Transport (DfT) and consist of approximately 19,000 geocoded count points in England and Wales, classified as Major (Motorways and A roads) and Minor (B, C and U roads). The count points provide information about the number of vehicles (AADT) driving at that particular point over the course of 2016. It is important to mention that the counts further distinguish among vehicle types (Two-wheeled motor vehicles, Cars and Taxis, Buses and Coaches, Light and Heavy Goods Vehicles). This information is currently not used in this paper although it is useful as part of further research discussed below. For this dataset, we further check for potential missing information and exclude faulty counters where identified. In Fig. 1, we show the average and the range of AADT (total number of motorized vehicles) for 'A', 'B', 'C' and 'U' roads.

The Integrated Transport Network (ITN) and ITN Urban Paths (ITNUP) spatial datasets have been extracted from Ordnance Survey (OS) and consist of the entire road network in Great Britain (GB). The ITN dataset contains information for all roads and road junctions, while the ITNUP dataset contains all man-made footpaths, subways, steps and footbridges as well as cycle paths in all urban areas of Britain.

Socioeconomic characteristics are derived from the Office for National Statistics (ONS), including information about population, population density, workplace population, workplace density, number of households and median income at the Lower Super Output Areas (LSOAs)[2] level. LSOAs are spatial datasets derived from OS where the socioeconomic characteristics and the number of registered cars and vans – derived from the Office for Low Emission Vehicles (OLEV) – are matched.

Urban area polygons are spatial datasets derived from OS that designate urban area's boundaries as defined by Ministry of Housing Communities and Local Government and Defra report (Bibby and Brindley, 2014) and used to indicate whether a point is located in an urban or rural environment. Geolocated bus stops and bus stations as well as Train and Light Rail stations[3] have been derived from the National Public Transport Access Nodes (NaPTAN) database. Finally, land use data have been extracted from various sources. First, a list of rateable values for non-domestic properties in England and Wales is provided by the Valuation Office Agency (VOA). The VOA dataset contains approximately 2.5 million records classified in over 100 classes based on a coding system, while addresses and postcodes for most of the properties are also included. This dataset had to be geocoded and existing categories were reclassified to 17 new classes, so as to reduce complexity. The new classes are shown in Table A2 in the Appendix. Moreover, considering that ports and airports have an impact on the transport network of their surrounding area (Hesse, 2013), their locations are derived from the British Port Association and the Civil Aviation Authority respectively. Finally, electric vehicle charging

[2] Approximately 35,000 areas designed by the Office for National Statistics (ONS) for England and Wales, with population minimum of 1000.

[3] This dataset includes all National Rail as well as all local metro, tram and light rail system stations such as London Underground, London Overground, Docklands Light Railway (DLR), London Tramlink, Tyne and Wear Metro (Newcastle), Merseyrail (Liverpool), Manchester Metrolink, NET (Nottingham), Supertram (Sheffield), West Midlands Metro (Birmingham - Wolverhampton), West Yorkshire Metro (Leeds) and Blackpool Tramway.

point locations are taken from OLEV. Considering location availability, we were able to map the land use datasets. A summary of all used datasets is shown in Table A1 in the Appendix.

4. Methodology

To estimate AADT, three major steps are considered. First, we use the data described in section 3 to design the variables to be used as model inputs. All variables are designed using GIS. Second, we feed the variables into the selected algorithms and use validation metrics to assess model's performance. Finally, we calculate the weighted average errors for each road type and across the road network.

4.1. Feature design

The initial step in our approach is to consider each point's spatial position and incorporate characteristics of the point's environment and location. We first want to take into account that urban areas generate and attract more activity and that the larger the area the more transport is generated (Caceres et al., 2018). Second, we need to bear in mind that the urban areas dataset contains all build up areas whether they are large urban centres or small towns, likely to exhibit different traffic. Finally, we also want to account for points marginally contained within or marginally excluded from the urban area polygons. In order to address these three issues, for each point we determine whether it is either urban or rural and also calculate four distance measures. (i) distance from urban area (ii) distance from major urban area[4] (iii) distance from urban area centroid[5] and (iv) distance from major urban area centroid[4,5]. Distances to urban areas are calculated as straight lines (Euclidean distances) from each point to the nearest edge of the nearest urban/major urban area polygon, while for centroids, distances are calculated as straight lines from each point to the centroids.

In terms of roadway characteristics, we introduce two indicators for toll roads[6] and ring roads and also take into account the "road nature" related to each count point, which demonstrates whether a point is located on a single carriageway, double carriageway, slip road, roundabout as indicated by OS and either Trunk[7] or Principal road as indicated by the Department for Transport (2014).

In terms of variables reflecting the characteristics of an area, rather than a single point, we follow the work of Koperski et al. (1998) based on the concept of service areas[8] which are created around each point. The service areas are of six different sizes (500 m, 800 m, 1000 m, 1600 m, 2000 m and 3200 m) for all road types as shown in Fig. 2. We use the concept of service area in the case of land use, accessibility to motorways and some of the public transport characteristics. Service areas are overlaid with the VOA and charging points datasets as well as with the ports and airports datasets, so as to assess land use within each area. Accessibility to motorways which is associated with higher traffic volumes (Apronti et al., 2016; Zhao and Park, 2004), is also assessed by overlaying service areas with motorway junctions. Bus stops and bus stations are treated the same way.

Finally, we take into account the socioeconomic characteristics – already available at LSOA level – as well as train and light rail stations. For the latter we first utilise the ITNUP[9] dataset and create 800 m service areas around each train station[10] and then we calculate the proportion of each LSOA covered by station service areas, so as a station accessibility attribute is also available at LSOA level. Lastly, we overlay count point service areas with LSOAs and introduce socioeconomic and station accessibility characteristics for each count point. Specifically, we incorporate the mean values for station accessibility, population density, workplace density, income and workplace plus population density, the last variable being used in Fu et al. (2017). In addition, the summed values of population, workplace population, households and registered vehicles are also calculated.

The feature design process generates 41 independent (33 numerical - 8 categorical) and the dependent variable (AADT). The variables are summarised in Table 1.

4.2. AADT estimation

Considering the large geographic extent and mixed characteristics, we expect AADT values and other variables to exhibit large variations across our study area. For example large differences are expected between urban and rural areas (Morley and Gulliver, 2016). For this reason, we first apply a clustering algorithm to take into account (dis) similarities among count points and their surroundings and group points with similar characteristics. Then, we apply three models, namely standard multivariate linear regression, Random Forests (RF) and Support Vector Regression (SVR) within each cluster and validate the results. In order for the models to be comparable based on the validation metrics, we feed all designed features to all models without undertaking further statistical tests (e.g. checking for collinearity or feature importance). That is, if one model is able to automatically handle complexities within the dataset, we consider this as an asset of the particular model. The process is applied for each road class (A, B, C, U) and each service area individually. Finally, for each road type we select the service area where the algorithm resulted into the lowest errors and merge the selected points to construct the full dataset, so as we can detect the optimal service area size for each road type and point location. That allows us to identify the optimal distance where a particular road is influenced by its surroundings.

4.2.1. Clustering

We use the K-prototypes (Huang, 1998) algorithm, suggested by He et al. (2006), which integrates the K-means and K-modes processes for numeric and categorical data respectively (Huang, 1997a) to cluster mixed type data[11] – see list of variables in Table 1. K-prototypes,

[4] We take into account the six largest urban agglomerations in England and Wales as defined by Pointer (2005). These areas are: The Greater London, West Midlands (Birmingham), Greater Manchester, West Yorkshire (Leeds and Bradford), Tyneside (Newcastle and Sunderland) and Liverpool Urban Areas.

[5] Where centroid is defined as the geometric centre of each urban area polygon.

[6] The "toll road" feature also includes the London Congestion Charge Zone, where all count points within the zone are considered to be toll roads.

[7] Trunk roads indicate long distance roads, usually connecting cities and having heavy traffic flows (Department for Transport, 2014)

[8] This is an improvement on the work of Zhao and Chung (2001), Zhao and Park (2004), Sarlas and Axhausen (2014) and Doustmohammadi and Anderson (2016) which uses buffers of different radii. Service areas construct buffers by taking into account the street network instead of Euclidean distances. We consider this measure to be more suitable for our case study, since it can capture the actual predefined distance a vehicle has to cover from/to the traffic count point.

[9] Notice that the use of ITNUP indicates that access to train stations can be by foot as well, using the footpaths, subways, steps, footbridges and cycle paths, although this data is available for urban areas only. If a train station is located at any rural area, we use the ITN instead.

[10] This threshold is considered as the standard distance one would consider walking to reach a station in most research (e.g. Cardozo et al., 2012; Gutiérrez et al., 2011)

[11] Our choice has been dictated by the fact that most clustering algorithms, for example the K-means, do not take into account categorical data, as based on Euclidean distance. Alternatives such as the chi-square (Greenacre & Primicerio, 2015) have been found to perform poorly (Faith et al., 1987; McCune and Grace, 2002). Kaufman and Rousseeuw (1990) advocates the use of the K-medoids algorithm incorporating the Gower's similarity coefficient (Gower, 1971) although the computational cost when using this type of similarity metric increases and it is therefore unsuitable for large datasets (Huang, 1998).

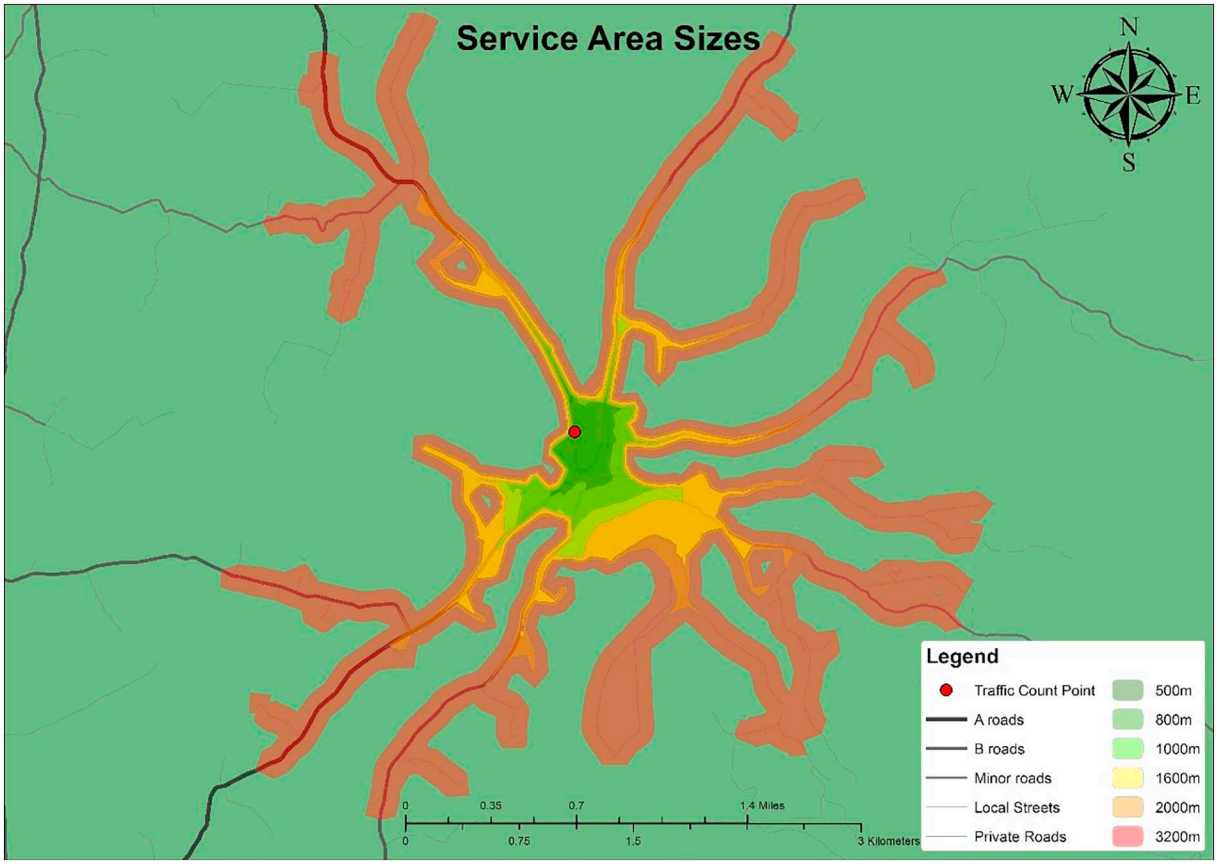

Fig. 2. Count point service areas.

Table 1
Independent variables.

Variable	Description	Type
1. Urban/Rural	A count point's surrounding environment	Categorical
2. Distance to Urban Area	The straight Euclidean distance from a count point to an urban area polygon edge	Numerical
3. Distance to Major Urban Area	The straight Euclidean distance from a count point to a Major urban area polygon edge	Numerical
4. Distance to Urban Area Centroid	The straight Euclidean distance from a count point to the geometrical centre of an urban area polygon	Numerical
5. Distance to Major Urban Area Centroid	The straight Euclidean distance from a count point to the geometrical centre of a major urban area polygon	Numerical
6. Toll Road	Whether or not the count point is located at a toll road	Categorical
7. Ring Road	Whether or not the count point is located on a ring road	Categorical
8. Road Nature	Whether the count point lies on a single or dual carriageway, slip road or roundabout	Categorical
9. Road Category	Whether the count point lies on a Primary or Trunk road	Categorical
10. Junction Accessibility	Whether the road where the count point is located has access to a motorway based on the specified service area	Categorical
11. Charging Points	The number of charging points within each service area	Numerical
12. Ports	Whether the road where the count point is located has access to a port based on the specified service area	Categorical
13. Airports	Whether the road where the count point is located has access to an airport based on the specified service area	Categorical
14. Bus Stops	The number of bus stops within each service area	Numerical
15. Bus Stations	The number of bus stations within each service area	Numerical
16. Train Accessibility	The adjacent LSOAs' average train station coverage	Numerical
17. Population	The total population of a count point's adjacent LSOAs	Numerical
18. Population Density	The average population density of a count point's adjacent LSOAs	Numerical
19. Workplace Population	The total workplace population of a count point's adjacent LSOAs	Numerical
20. Workplace Population Density	The average workplace population density of a count point's adjacent LSOAs	Numerical
21. Workplace plus Population Density	The average workplace plus population density of a count point's adjacent LSOAs	Numerical
22. Income	The average median income of a count point's adjacent LSOAs	Numerical
23. Households	The total number of households of a count point's adjacent LSOAs	Numerical
24. Registered Vehicles	The total number of registered cars and vans of a count point's adjacent LSOAs	Numerical
25–41. VOA (17 features – Table A2)	The total number of VOA elements in the predefined count point's service area	Numerical

instead of taking samples from the dataset, uses the whole dataset and thus it does not suffer from sampling bias and it is less computationally intensive compared to K-medoids or various Hierarchical Clustering algorithms that can handle mixed variable types. For numerical variables, K-prototypes uses squared Euclidean distances as in K-means, while for categorical variables, the dissimilarity measure is defined by the total mismatches of the attribute categories of two objects (Huang, 1998) so that the overall distance metric is equal to the squared

Euclidean distance to measure (dis)similarity for numerical variables and the matching (dis)similarity for the categorical variables,

$$d(X,Y) = \sum_{j=1}^{m_r} (x_j - y_j)^2 + \gamma \sum_{j=1}^{m_c} \delta(x_j, y_j) \quad (1)$$

where X and Y are the two mixed-type objects, m_r and m_c are the numbers of numeric and categorical attributes respectively and γ is a weight to avoid favouring either type of attribute (Huang, 1997b). δ indicates the dissimilarity (mismatches) for the categorical variables, where

$$\delta(x_j, y_j) = \begin{cases} 0\,(x_j = y_j) \\ 1\,(x_j \neq y_j) \end{cases} \quad (2)$$

We also transform the data to address the problem of different measurement units and ranges. Data transformations make features dimensionless to overcome the problems resulting from the dependence on different measurement units and the deviations among variable variances that affect cluster quality and formations (Rokach and Maimon, 2015; Zhang et al., 2019) so that each variable can play an equal role in the analysis (Greenacre and Primicerio, 2013; Han et al., 2012; Mohamad and Usman, 2013).[12]

In terms of the specific transformation, we apply the min-max normalisation, the most common form of normalisation, which sets all variables within the range of 0 to 1 based on:

$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (3)$$

where min(x) and max(x) are the minimum and maximum values of each variable respectively. We also bear in mind that the parameters thought to be more relevant in separating the groups should be assigned a higher influence factor (Hastie et al., 2011), i.e. weights[13] to raise their importance of certain variables which are considered more critical in cluster formation (Gebotys and Elmasry, 1989; Hummel et al., 2017). Weights can be assigned by multiplying the variables with a constant (Akhanli and Hennig, 2017; Hammah and Curran, 1999). In our case, we want to form clusters where AADT values are similar and independent variables are relatively correlated with the dependent within the same cluster, so as to have accurate predictions. We want the dependent variable to have a high enough weight (range) to influence the formation of the cluster, although without dominating it.[14] Hence, we follow Bacher et al. (2004) who apply random lower weights to variables separating the clusters to achieve equal influence and similarly, Opsahl and Panzarasa (2009) who also assign random weights between 1 and 10 to links (edges) on their work on clustering networks. That is, we change variable ranges and assign weights of 1 and 10 to the independent and dependent variables respectively.[15] We achieve that by implementing a generalised version of the min-max standardisation above which can be used to transform a range of values into another [α, β], i.e[16].

$$x' = (\beta - \alpha)\frac{x - \min(x)}{\max(x) - \min(x)} + a \quad (4)$$

with $\alpha = 1$ and $\beta = 10$ we can get the required range. In the case of the AADT, we set values within the range of 1 to 10 and do not consider the value of 0 for the dependent variable, since there are no observations with zero value.

As the K-prototypes algorithm requires to define the number of clusters (K) beforehand clustering is implemented, we employ the "elbow" method which is considered as the optimal since it is the only one considering mixed data types.[17] The elbow method examines the percentage of variance as a function of the number of clusters (Bholowalia and Kumar, 2014), the idea being that starting with K = 2 and increasing the number of clusters, at some point the marginal gain drops dramatically and gives an angle in the graph (Kodinariya and Makwana, 2013) indicating the optimal K. When testing 20 clustering processes, related to 4 road types and 6 service area sizes, the optimal number for K ranged between 4 and 6 depending on the case examined each time. For clarity reasons, we select five clusters for all cases, e.g. Fig. A1 in the Appendix.

4.2.2. AADT estimation

We first randomly split the dataset into two groups, 80% of the observations for training and 20% for testing and we use the training dataset to implement three different models, namely 1) standard multivariate linear regression (OLS), 2) Random Forest (RF) and 3) Support Vector Regression (SVR).

The multivariate linear regression model is as follows:

$$AADT_i = \beta_0 + \beta_1 x_{i1} + ..+\beta_j x_{ij} + \varepsilon_i \quad (5)$$

where: $AADT_i$ is the dependent variable at the *ith* observation, $i = 1, \ldots, n$, x_{ij} is the value of the *jth* independent variable in the *ith* observation, $j = 1, \ldots, m$, β_0 is a constant term, β_j is the regression coefficient for the *jth* independent variable, ε_i is the error term and m is the number of independent variables. Random Forest (RF) is a machine learning technique, used for classification and regression, introduced by Breiman (2001). RF are a collection of decision trees, an example of so-called ensemble methods, based on bootstrapping (Efron, 1979) and bootstrap aggregation (Breiman, 1996). The RF regression prediction is given by:

$$\widehat{f_{rf}^B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (6)$$

where: B is the number of trees and $T_b(x)$ is the b^{th} random forest tree grown from b bootstrapped data. Here, we use 500 trees and 5 variables for the forest to sample at each split. Finally, support Vector Regression (SVR) is the extension of Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995) proposed by Drucker et al. (1997). SVR aims

[12] Large variable range tend to have large effect on the resulting clustering structure (Kaufman and Rousseeuw, 1990; Mohamad and Usman, 2013). As variable measurement units and their respective ranges play a significant role in the cluster formations, methodological guidance on the use of transformation is very clear-cut in the literature, as applying data transformation is considered essential for most practical applications to enhance performance (Bishop, 2013). In particular, numerical variables should be transformed to scale their effect on the results (Larose, 2005) and conventional distance measures (e.g. Euclidean) should not be used without applying transformations on the data (Mohamad and Usman, 2013).

[13] The weights can be unequal among the variables to define their influence (Friedman and Meulman, 2004) and can also be zero if they do not possess any important information (Hammah and Curran, 1999).

[14] Considering we have 41 independent and 1 dependent features of different types and ranges, applying the K-prototypes algorithm without transforming the data, results into clusters dominated by the independent variables only, while the same output is observed when all variables have equal influence. On the other hand, transforming only the predictors, results into clusters dominated by AADT values, since the ranges are extremely different.

[15] We acknowledge that some variables do not directly affect all vehicles; hence their contribution to AADT may be questionable (e.g. charging points are only useful for electric vehicles). However, we do not examine the contribution of each variable to different types of vehicles, but to AADT for all motorised vehicles. Moreover, we focus on the accurate estimation and model comparison, thus we have chosen to give all independent variables the same weight, so as to be able to draw rational inferences when comparing models.

[16] As it can be seen, the required ranges are set by applying data transformations and consequently weighting is achieved without multiplying by a constant.

[17] Other methods include the "Silhouette" method, the Calinsky – Herabasz Criterion, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) among others.

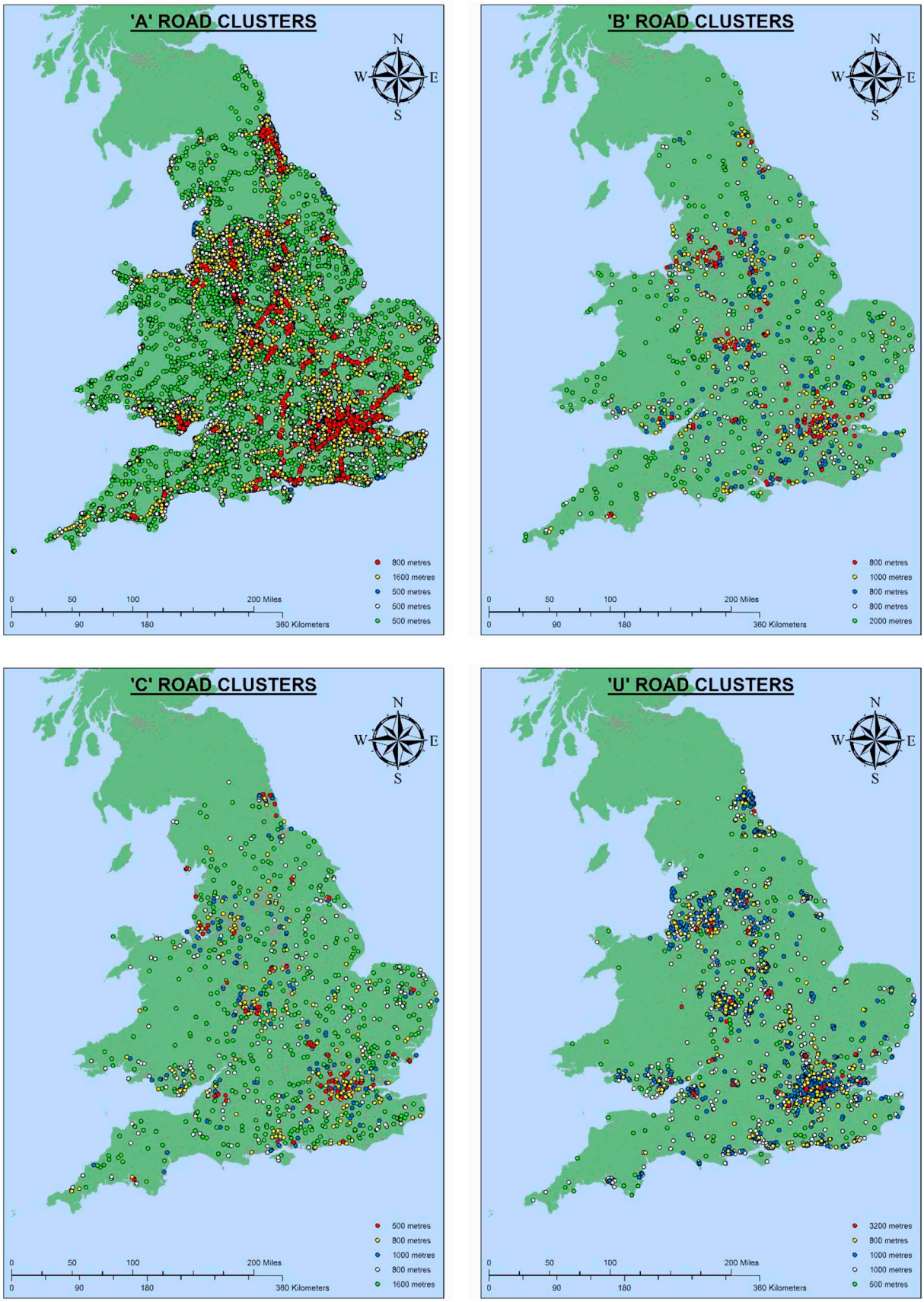

Fig. 3. Clusters for 'A' (top left), 'B' (top right), 'C' (bottom left) and 'U' (bottom right) roads.

to find a function $f(x)$ where predicted values are at most ε from the observed ones. The general SVR equation for non-linear predictions is given by (Basak et al., 2007):

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) kx_i, x + b \tag{7}$$

where: α_i, α_i^* are the Lagrange multipliers, $k\langle x_i, x\rangle$ is the kernel[18] and b is the bias. We use the radial basis Kernel and by replacing in (7) we have:

$$y = \sum_{i=1}^{N} (a_i - a_i^*) * \exp(-\gamma \, ||x_i - x_j||^2) \tag{8}$$

where: $\gamma = \frac{1}{2\sigma^2}$ and set to 0.1.

4.2.3. Validation

Prediction accuracy is validated using the test set comprising 20% of the dataset. We use two validation measures, the Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_i - F_i}{A_i} \right| \tag{9}$$

where: A_i is the observed value, F_i is the predicted value and n is the number of observations, and the Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (F_i - A_i)^2}{n}} \tag{10}$$

4.2.4. Weighted average

Weighted average is calculated on the lowest identified MAPEs for each model j across clusters i for each road class c,

$$WMAPE_{j,c} = \sum_{i=1}^{K} \left(\frac{AADT_{i,c}}{\sum AADT_{i,c}} \right) * MAPE_{i,j,c} \tag{11}$$

where: $WMAPE_{j,\,c}$ is the weighted average MAPE for model j in road class c, $AADT_{i,\,c}$ is the total traffic for cluster i at road class c, $MAPE_{i,\,j,\,c}$ is the MAPE for cluster i, model j and road class c and K is the number of clusters. Then, we similarly calculate the overall weighted average MAPE across road types for the whole road network for each model j,

$$OWMAPE_j = \sum \left(\frac{AADT_c}{\sum AADT_c} \right) * WMAPE_{j,c} \tag{12}$$

where: $AADT_c$ is the total traffic counted for road class c. Similarly, we also calculate the weighted values for the corresponding RMSEs.

5. Results

As the first result from this study, it is interesting to comment on the estimated clusters, as they exhibit similar patterns across road types. This is shown in Fig. 3, where the clusters and related optimal service area sizes are colour-coded. In particular, for each of the four road types, i.e. A, B, C and U

- cluster 1 (red) contains points located on roads where traffic counts tend to be higher, such as ring and trunk roads in the case of 'A' road class and evenly split between urban and rural areas. For 'B', 'C' and 'U' roads points are placed at locations of higher transport significance, almost exclusively located in urban areas;
- cluster 2 (yellow) includes relatively high traffic values with points in 'A' roads located both in urban and rural environments, while for other road types this cluster is mainly formed by urban locations;
- cluster 3 (blue) consist of medium AADT values with points for all road types located within urban areas, significantly concentrated in city centres. In particular, 'A' road points are observed within designated major urban areas as well as the city centres of some medium and small urban centres;
- cluster 4 (white) also contains medium AADT, although usually smaller than the values in cluster 3. These points are mainly located in suburban areas of large urban areas as well, but also in the centre of smaller settlements. Some of the points are also observed in rural areas, especially in the case of lower class roads;
- finally, cluster 5 (green) contains the lowest AADT values which are normally located in rural areas and the outskirts of urban centres.

As an additional result, our work casts light on the performance of different methods across the five clusters formed for each of the four road types. Fig. 4 displays both the MAPEs and RMSEs for the 120 combinations of clusters and road types for each of the three methods implemented in this study.

As one can see in the Fig. 4 and Table 2, the two Machine Learning methods are fairly equivalent and outperform the regression method. In the case of the SVR, the MAPE ranges between 2% (cluster 3 in C roads) and 276.9% (cluster 5 in C roads) while the MAPEs achieved by RF range between 2.2% (cluster 3 in B roads) and 288% (cluster 5 in C roads). Among the three methodologies we implemented, linear Regression exhibits the highest MAPEs in almost all cases, with values falling between 2.1% (cluster 3 in C roads) and 324.8% (cluster 5 in C roads). Linear Regression also produce a very big error in cluster 1 of U roads, probably due to the very small number of observations in this cluster (31 points – 25 for training and 6 for testing). Considering this result unreliable, we exclude it from Fig. 4, although it is interesting to see that the SVR performs well also in this case despite the very small number of observations. Similar conclusion emerges when assessing the performance of the methodologies implemented in this study based on the RMSEs, therefore adding robustness to the conclusion that the two Machine Learning methods are fairly equivalent and outperform the regression method (Table 2). It is worth mentioning however that RF produces lower RMSE than SVR in the case of cluster 1 – 'A' roads which is by far the combination with the highest level of traffic (Table A3). Linear Regression continues to produce the largest errors and again results into very large error in cluster 1 at U roads (not shown in Fig. 4).

One can also appreciate from Fig. 4 that the predicting patterns, as measured by the MAPE and RMSE are similar for all models, with higher MAPEs usually observed in cluster 5 and higher RMSEs in cluster 1 across road types. This is however a simple reflection of the fact that cluster 5 comprises observations with relatively low traffic which translates in higher MAPE, while cluster 1 contains cases with high level of traffic so that the RMSE (which tends to be influenced by the level of the observations) is corresponding high. The range of the RMSE values across clusters make comparison difficult – as an example it goes from about 5000 in the case of cluster 1 to as a low as 55 in the case of cluster 5 in C roads. The relatively small values for most combinations of clusters and roads in Fig. 4 shows that the 3 methods produce similar results when measured in terms of vehicles per day, in a way that they may all satisfy users' needs unless they focus on specific types of roads and traffic flows for which a specific method can work better than another.

This is confirmed by Table 2, which presents MAPEs and RMSEs, first averaged across clusters and eventually across road types to obtain an overall MAPE and an overall RMSE for each model. Traffic volumes – presented in Table A3 – are used throughout as weights in the averaging process. One can see that MAPE is highest for 'C' and 'U' roads and smaller for 'A' and 'B' roads with the lowest ones observed at 'B' roads. Moreover, one can see more clearly that SVR is the best performing model when measured based on the weighted MAPE, while regression has the highest MAPE for all road types. SVR again outperforms RF in 'B', 'C' and 'U' roads, with the MAPE of SVR being 0.2% lower than RF in

[18] The kernel refers to a function that maps data from one space to another higher dimensional feature space (Hastie et al., 2011)

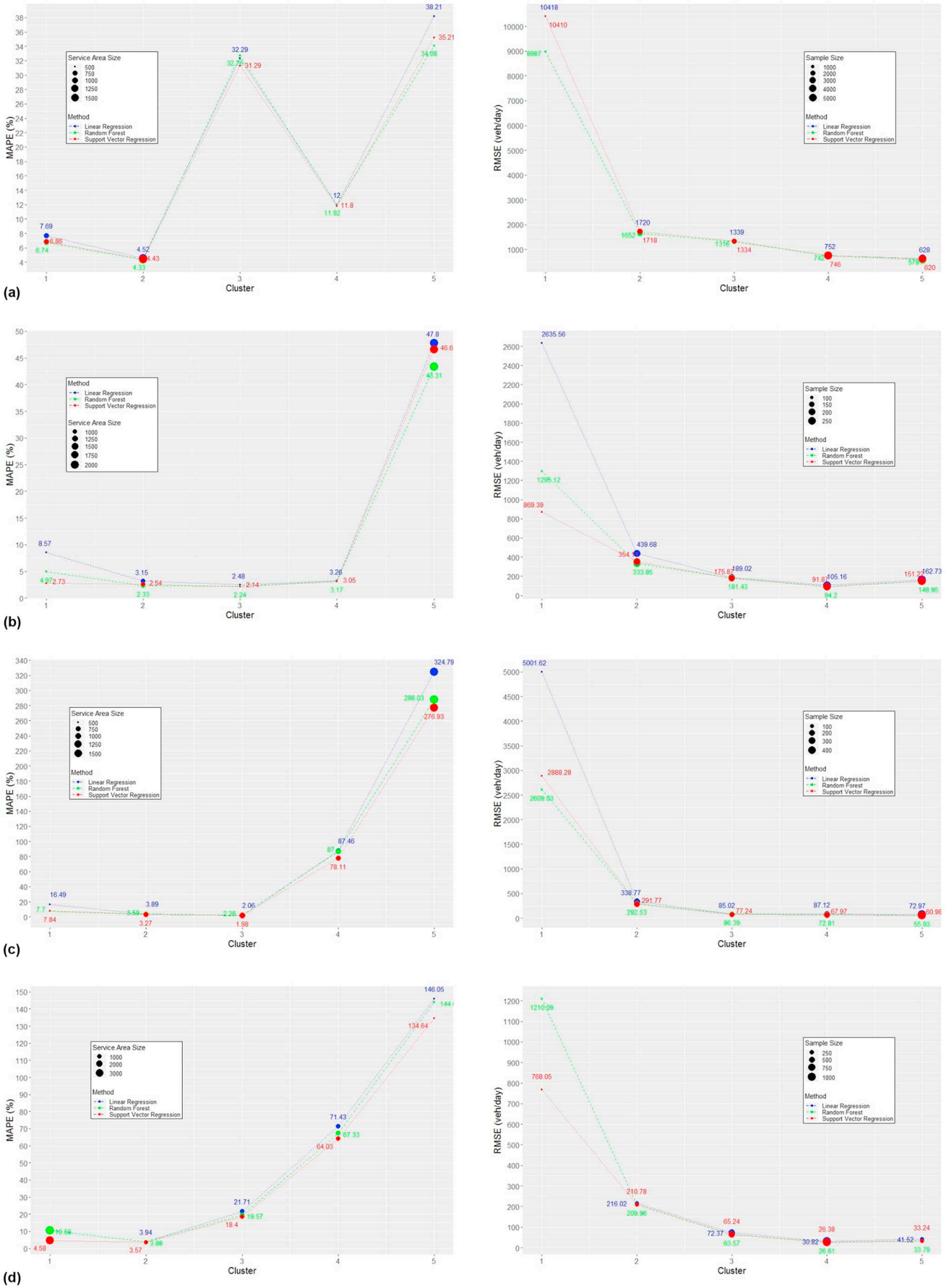

Fig. 4. (a) MAPE (left) and RMSE (right) for 'A' roads.
(4b) MAPE (left) and RMSE (right) for 'B' roads.
(4c) MAPE (left) and RMSE (right) for 'C' roads.
(4d) MAPE (left) and RMSE (right) for 'U' roads.

'B' roads and increasing at 'C' and 'U' roads respectively, e.g. 27% versus 29.3% in the case of 'U' roads. For 'A' roads, however, the performance of SVR and RF is essentially identical and the gap of these two methods with the linear regression shrinks to 0.8 percentage, as is the overall MAPE with SVR performing slightly better than RF but only by 0.01.

In terms of RMSEs, it can be seen that the errors are higher for higher class roads and decrease for the lower class roads as expected. This same expected pattern is also observed within the clusters of each road class for the unweighted errors as shown in Fig. 4 and Table 2. However, RMSE values are lower in 'B' compared to 'C' roads where AADT values are usually lower as shown in Fig. 1. The averaged RMSEs show that errors are again higher for Linear Regression and are also balanced between SVR and RF, with RF resulting to lower errors half of the time. However, observed differences are small and the mean difference between RF and SVR is 217 vehicles across all road types in favour of RF.

As a final result, we are able to elaborate on pattern of the optimal service areas, i.e. the area producing the lowest MAPE, across clusters and road types, as shown in Fig. 5. Here, a clear pattern is evident for road classes 'B' and 'C', where the service areas are small and similar for clusters 1 to 4 and increase at cluster 5. On the contrary, the optimal service areas for 'U' roads follow the opposite pattern starting at large service area for cluster 1 and gradually decreasing to reach the minimum (500 m) for cluster 5. Service areas for 'A' roads are of medium size and also minimise for clusters 3, 4 and 5. In addition it can be observed that small to medium service areas dominate the figures, with only two large service areas observed at 'U' roads cluster 1 (3200 m) and 'B' roads cluster 5 (2000 m).

In addition, one can observe that clusters 3 and 4 fluctuate around small to medium service area sizes and tend to increase as the road class decreases in significance (from 'A' to 'U'), in contrast with cluster 2 where the service area decreases together with the respective road class. Cluster 1 exhibits an increasing pattern and cluster 5 – representing the "rural" areas – is optimised at small service areas for road classes 'A' and 'U' and at higher ones for 'B' and 'C' roads.

6. Discussion

This study has focused on the development of a methodology to accurately and effectively predict AADT. The procedure has been applied to AADT figures collected for all roads in England and Wales, therefore providing a rigorous and comprehensive test of the process outlined in this article. We started by including several variables postulated to affect traffic flows, based on results from previous studies in the literature, as inputs into predictive models. These variables portray a detailed representation of roadway, land use, socioeconomic and public transport characteristics. Specifically, utilisation and manipulation of spatial data within GIS, facilitated feature design and the analysis, so as to incorporate related socioeconomic, land use and roadway attributes – used as AADT predictors – which are directly associated with the count points' spatial locations.

The output from our models has been assessed using statistical validation metrics normally employed in the literature, in particular MAPE and RMSE. As the focus of our approach is on prediction, the metrics above were computed on the test dataset, i.e. 20% of the sample, we had available. The fact that our choices conform to the standard in the literature both in terms of inputs and metrics to assess the output make our results even more compelling, as we are able to deliver a remarkable accuracy of the AADT predictions obtaining out-of-sample MAPEs as low as 2%. This contrast markedly with the results arising from the applications in the literature, where in some cases lowest errors are 50% for similar road types (e.g. Selby and Kockelman, 2013) or ranging between 39% and 400% in others (Wang et al., 2013).

We attribute the significant improvement in accuracy to two interrelated aspects of our approach: data transformation and clustering. First of all, the clustering algorithm revealed groups where data exhibit similar characteristics while the application of data transformations allowed the clustering algorithm to create groups where both similar AADT values and related characteristics have been taken into account. This can be concluded both from section 5 where the clusters are presented and even more so from Fig. 6 where points in city centres are

Table 2
Original and weighted MAPEs and RMSEs.

Road class	Cluster	Service area	MAPE (%)			RMSE (vehicles per day)		
			OLS	RF	SVR	OLS	RF	SVR
A Roads	1 – red	800	7.7%	6.7%	6.9%	10,418	8987	10,410
	2 – yellow	1600	4.5%	4.3%	4.4%	1720	1652	1718
	3 – blue	500	32.3%	32.8%	31.3%	1339	1316	1334
	4 – white	500	12.0%	11.9%	11.8%	752	742	746
	5 – green	500	38.2%	34.1%	35.2%	628	578	620
	Weighted Average		15.2%	14.4%	14.4%	2429	2193	2423
B Roads	1 – red	800	8.6%	5.0%	2.7%	2636	1295	869
	2 – yellow	1000	3.2%	2.3%	2.5%	440	334	376
	3 – blue	800	2.5%	2.2%	2.1%	207	166	177
	4 – white	800	3.3%	3.2%	3.1%	105	94	92
	5 – green	2000	47.8%	43.3%	46.6%	163	149	151
	Weighted Average		7.4%	5.9%	5. 7%	815	471	381
C Roads	1 – red	500	16.5%	7.7%	7.8%	5002	2609	2888
	2 – yellow	800	3.9%	3.6%	3.3%	339	293	292
	3 – blue	1000	2.1%	2.3%	2.0%	85	86	77
	4 – white	800	87.5%	87.0%	78.1%	87	73	68
	5 – green	1600	324.8%	288.0%	276.9%	73	56	61
	Weighted Average		37.1%	31.8%	30.3%	1408	796	863
U Roads	1 – red	3200	368.7%	10.6%	4.6%	58,650	1210	768
	2 – yellow	800	3.9%	3.9%	3.6%	216	210	211
	3 – blue	1000	21.7%	19.6%	18.4%	72	64	65
	4 – white	1000	71.4%	67.3%	64.0%	31	27	26
	5 – green	500	146.1%	144.0%	134.6%	42	34	33
	Weighted Average		75.2%	29.3%	27.0%	7327	235	181
ALL ROADS	Overall Weighted Average		15.69%	14.48%	14.47%	2413	2119	2336

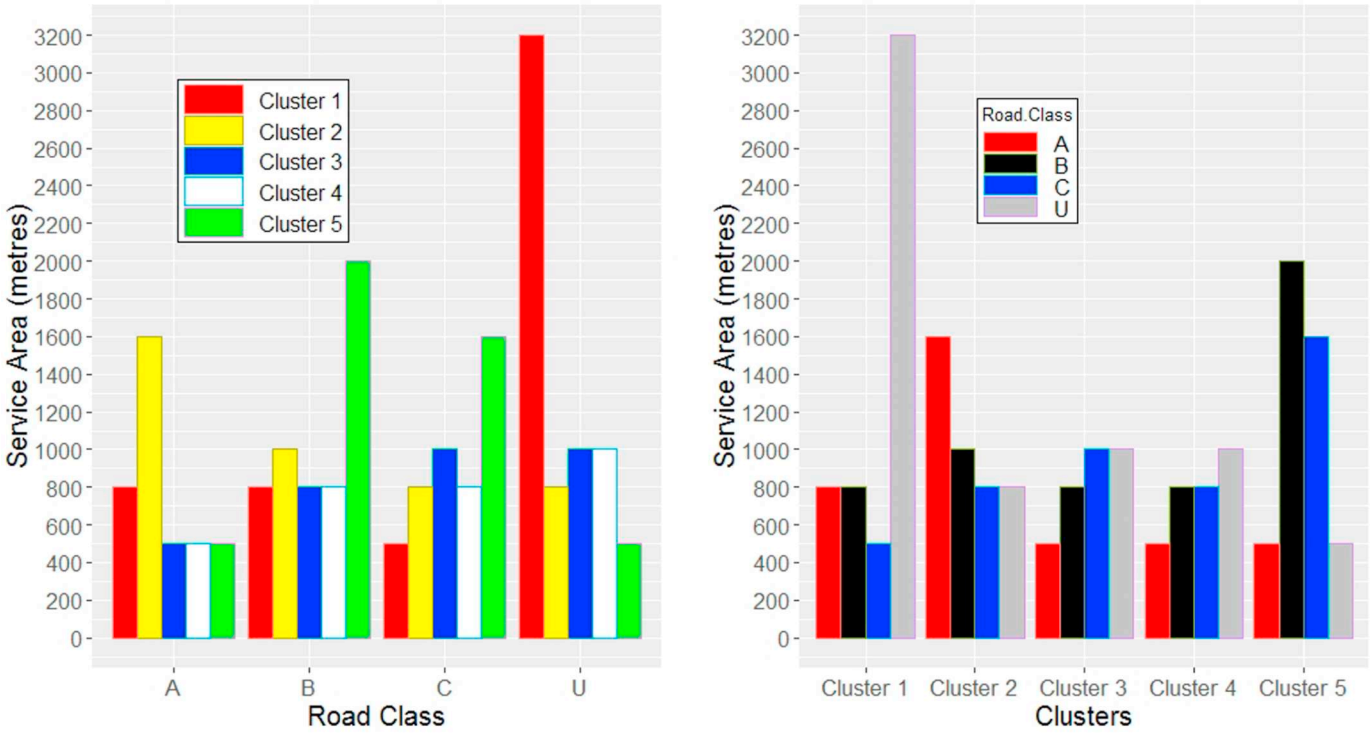

Fig. 5. Optimal service areas by road class (left) and cluster (right).

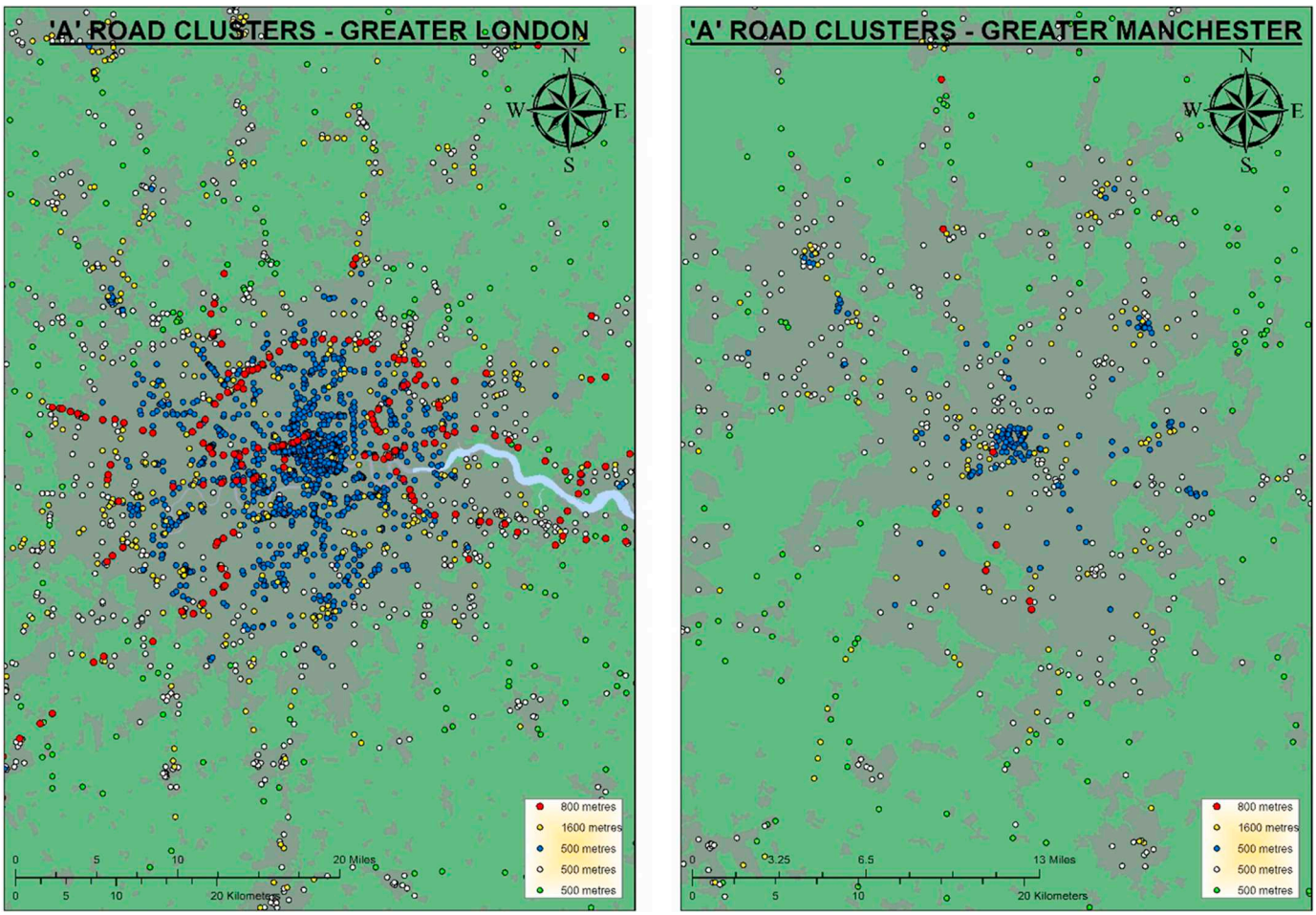

Fig. 6. A road clusters for Greater London (left) and Greater Manchester (right).

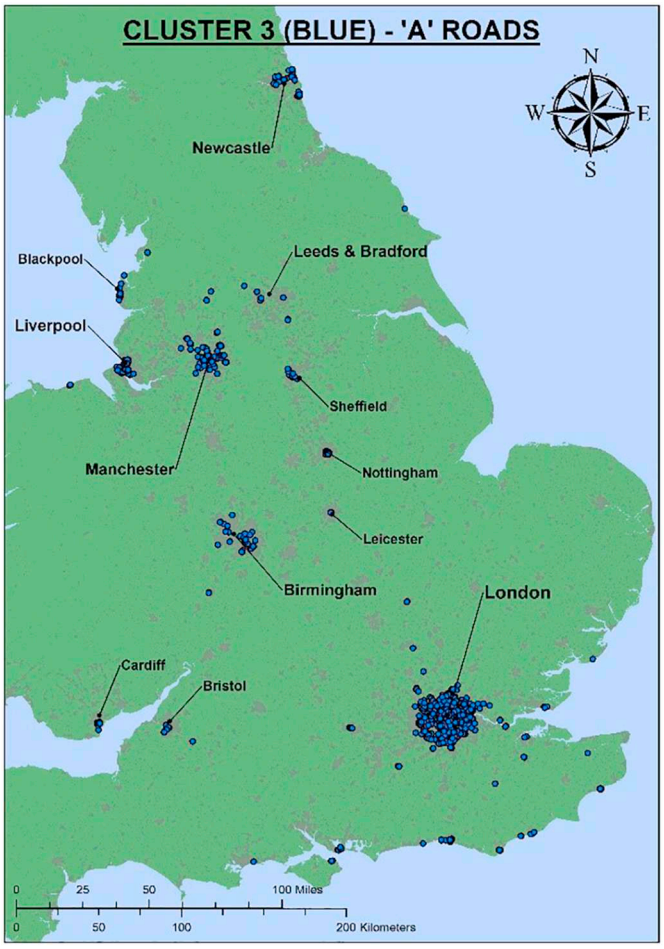

Fig. 7. Cluster 3 - 'A' Roads.

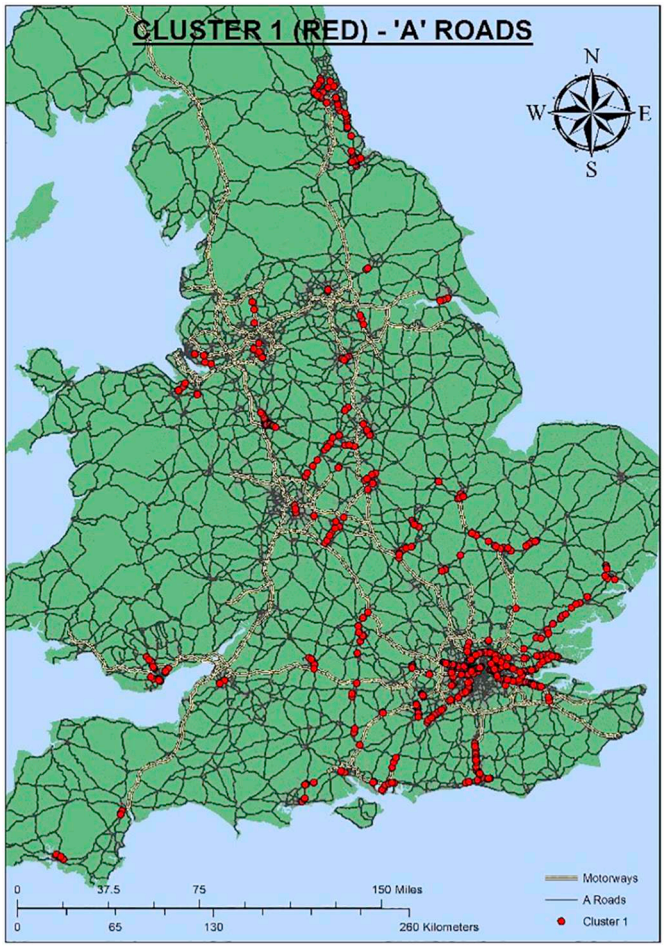

Fig. 8. Cluster 1 - 'A' Roads.

clustered together, indicating areas with similar characteristics (e.g. a large number of shops and businesses) and picking up underlying roads.[19]

However, error deviations among the models, clusters and road types presented in Table 2, show that the models' performance – in terms of MAPE – is dependent on two conditions. First is the value of the dependent variable (i.e. the amount of traffic per traffic count point) within each cluster, where high MAPEs are observed for clusters with low AADT values – usually clusters 4 and 5 – in most cases. Nonetheless, this expected outcome is due to the fact that the estimated variable can have values very close to zero (Caceres et al., 2018) and consequently even slight deviations would exaggerate the error. For example, a misprediction of 10 vehicles would have a different impact on MAPE for an observed value of 100 compared to an observed value of 100,000. However, the exception of the unexpectedly high MAPE in cluster 3 for 'A' roads, can be due to the characteristics of the areas where the points are located. As it is shown in more detail in Figs. 6 and 7 the points are located at city centres of large and major urban areas, usually associated with diverse land use and complexity. Thus, this cluster can be further disaggregated to improve accuracy (Greenacre and Primicerio, 2013).

Second condition to affect models' performance is the number of data instances (i.e. sample) within each cluster, particularly in the case of Linear Regression. Specifically, Linear Regression results into very high MAPE for the smallest sample across the data set (31 points at cluster 1 – 'U' roads) and also is over 9% higher compared to RF and SVR for the second smallest sample (59 points at cluster 1 – 'C' roads) and 3.5% - 6% higher for the third smallest sample (86 points at cluster 1 – 'B' roads). Sample effect is also noticeable at the RMSEs, where Linear regression again produces very large error in cluster 1 at 'U' roads, while all models also result into high RMSEs at cluster 1 at 'C' roads even compared with cluster 1 – 'B' roads where traffic counts are higher.

It is important to mention that sample size affects overall model performance. For example, models perform similarly – and potentially more accurately – in the case of 'A' roads, including most of the traffic count points comprised in our sample (i.e. approximately 15,000 out of 19,000 points). As 'A' roads account for over 95% of the total traffic in our database – see Table A3 – it turns out that the overall MAPE is fairly similar to the MAPE for 'A' roads, to the great benefit of linear regression in terms of comparison across methods. That is, if one is to take into account the traffic values estimated by DfT (Table A4) it appears that 'A' roads account for 57%, while 'C' and 'U' roads combined – where errors are higher – account for 34% of the total traffic. Consequently, MAPEs weighted based on Table A4 results in only 8% higher error for OLS compared to RF (27.2% versus 19.2%) and 8.6% higher error compared to SVR (27.2% versus 18.6%). This leads to the conclusion that again SVR performs better than RF and Linear Regression performance is overstated by the sample.

As a final remark, we look into cluster 1 at 'A' roads. From Fig. 8 it can be observed that the points clustered here are mainly associated with ring roads (e.g. north circular in London – also in Fig. 6), motorway extensions (e.g. part of A23 from Crawley to Brighton – Fig. 9) as well as roads connecting urban centres – usually trunk roads – such

[19] In the case of road, the north circular (the ring road in north London) is clearly visible among the red dots.

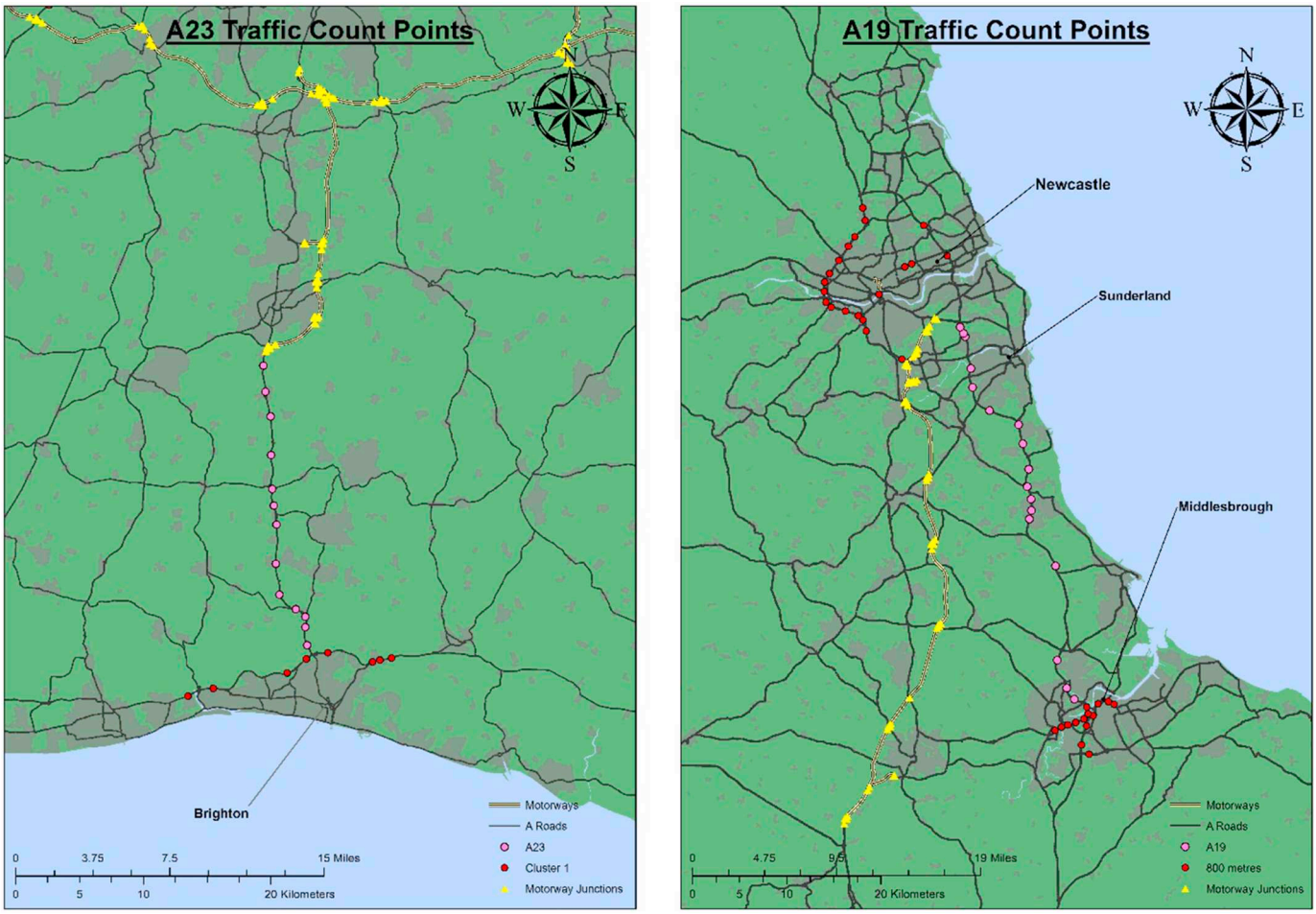

Fig. 9. A23 (left) and A19 (right) roads in cluster 1.

as the A19 (Fig. 9). Moreover, 96% of these points in this cluster are double carriageways, 15% ring roads and over 10% have an access to motorways within 800 m. In addition, from Table A3 it can be seen that points in this cluster have an average of approximately 75,000 vehicles per count point, while for motorways (not included in our analysis) there are approximately 74,000 vehicles per point. This leads to the conclusion that count points included in this cluster can potentially have strong similarities with motorways, indicating that traffic on these roads is not necessarily related with the road's surrounding environment. With regards to this point it can be concluded that road classes can be confused specifically when data from DfT and OS – or other sources – are combined. For example, busy roads based on the one source may be classified as minor according to the other. Consequently, roads should be classified based on the traffic and not ownership as it has also been pointed out by Xia et al. (1999) who faced similar issues.

7. Conclusions and further research

Based on the results presented here, further research requirements have become evident. First of all, the extension of the modelling approach proposed in this article require the identification of ways to classify points with unknown traffic (AADT) measurement to existing clusters. Considering that AADT values are not available, methods to classify data with missing variables need to be explored – perhaps by identifying the shortest distance from the new points to the centre of cluster centroids. One can also identify the variables with the smallest degree of overlap across clusters, and use this limited set of variables in the process of allocating new points to existing clusters. For K-prototypes – and similar clustering algorithms such as the K-means – clustering centroids and the corresponding variables can be identified and consequently, distances of new points to existing centroids can easily be computed. The identification of distinctive variables, implies of course further investigation of the clusters, so as to order the factors influencing cluster formation according to their importance. This would also enable better interpretation of results in a way that can be beneficial to transport, city planning as well as environmental studies. In fact, preliminary analysis on the impact of several variables to cluster formation, has revealed that there are variables taking a much more distinct set of value for each group and road type, while the contribution of AADT was found to have an impact on cluster formation, as it should be expected, although not dominant.

These aforementioned factors can also be further explored with statistical models and methods, so as to identify and explain the degree of influence on traffic flow variations across the road network, while also addressing the limitations identified in similar studies. The collection of additional data and incorporation of explanatory variables that have the potential to improve performance (Domingos, 2012; Junqué de Fortuny et al., 2013) can also be explored in future studies. It would also be interesting to implement models like ours for individual vehicle types so as to reveal patterns peculiar to specific traffic flows, as

well as to further expand this research by estimating mileage for the corresponding street segments where traffic counts are located. The significance of mileage cannot be overstated since it is extensively used in numerous studies in road transport, such as estimating air pollutant emissions (e.g. Labib et al., 2018; Patarasuk et al., 2016), and there can be no accurate calculation of mileage without precise AADT (Leduc, 2008). Moreover, implementation of other clustering and validation techniques – e.g. automated weighting clustering algorithms proposed in other studies (e.g. Chen and Wang, 2013; Huang et al., 2005) and k-fold cross validation (Koul et al., 2018) – could reveal different patterns of traffic flows and affecting factors as well as potentially provide more accurate error measurement. Finally, considering data availability and computational capacity a spatio-temporal modelling approach for a detailed and comprehensive assessment of AADT and changes in AADT across space and time should be considered in future studies.

Acknowledgments

The authors would like to thank the Valuation Office Agency for providing us with access to the dataset. Information provided by the Valuation Office Agency contains in this article is provided under the Open Government Licence. This work was supported by the Natural Environment Research Council (NE/M019799/1) and by the UK Energy Research Centre (Grant Number: EP/L024756/1).

Appendix

Table A1
Outline of used datasets.

Dataset	Source	Description	Spatial
1. Traffic count points	Department for Transport (DfT)	Geocoded count points in England and Wales	N
2. Integrated Transport Network (ITN)	Ordnance Survey (OS)	road network in Great Britain (GB) – roads and road junctions	Y
3. Urban Paths (ITNUP)	Ordnance Survey (OS)	man-made footpaths, subways, steps, footbridges and cycle paths in Britain's urban areas	Y
4. Lower Super Output Areas (LSOAs)	Ordnance Survey (OS)	Designated areas for England and Wales with minimum 1000 population	Y
5. Socioeconomic Characteristics	Office for National Statistics (ONS)	population, population density, workplace population, workplace density, number of households and median income at each LSOA	N
6. Number of registered vehicles	Office for Low Emission Vehicles (OLEV)	Number of registered cars and vans for each LSOA	N
7. Urban Area polygons	Ordnance Survey (OS)	Urban Areas boundaries	Y
8. Bus stops and stations	National Public Transport Access Nodes (NaPTAN) database	Geolocated bus stops and stations in Britain	N
9. Train and light train stations	National Public Transport Access Nodes (NaPTAN) database	Geolocated train and light rail stations in Britain	N
10. Ports	British Port Association	Geolocated passenger and commercial ports	N
11. Airports	Civil Aviation Authority	Geolocated passenger airports	N
12. Charging points	Office for Low Emission Vehicles (OLEV)	Geolocated charging points for electric vehicles	N
13. Non-domestic properties	Valuation Office Agency (VOA)	Geolotagged and classified properties in England and Wales	N

Table A2
VOA re-classified variables.

Class	Elements
Research, Education and Training	Schools, Colleges, Libraries, Universities, Language and Music Schools, etc.
Factories, Workshops and Industrial Activity	Energy Production Facilities, Factories, Workshops, Mines, Oil Fields, Recycling Plants, Shipyards, Scrap Yards
Healthcare	Hospitals, GPs, Surgeries, Clinics
Leisure	Public Houses, Bars, Nightclubs, Restaurants, Art Galleries, Cinemas and Theatres, Coffee shops
Office and Business Space	Offices, Banks, Business Units
Public Services, Infrastructure and Buildings	Post Offices, Community Centres, Police and Fire Stations, Prisons, Courts
Shops, Stalls, Kiosks and Markets	Shops, Kiosks, Showrooms, Stores
Super/Hyper Stores	Superstores, Malls
Sport	Stadia, Sport Centres, Golf Courses, Tennis Centres, Football Grounds
Vacation Sites, Accommodation and Facilities	Campsites, Caravan Sites, Hotels, Guest Houses, Holiday Units, Hostels, Motels, Beach Houses
Petrol Stations	Petrol Stations
Vehicle Infrastructure	Vehicle Repair Workshop, Garages, Car Wash
Warehouse and Storage	Warehouses, Depots, Storage Depots, Land Used for Storage
Parking Space	Car/Vehicle Park Sites and Park Spaces, Motorcycle Bays
Animal Husbandry, Farming and Agriculture	Aviaries, Farms, Animal Shelters, Stud Farms
Marine Infrastructure	Mooring Sites, Quays, Wharfs, Lifeboat Stations, Marine Control Centres
Under (re)construction	Properties and Premises Undergoing (re)Construction

Table A3
Traffic values by road class and cluster.

Road Class	Cluster	Number of points	Traffic Sum	Traffic per point	Share
A Roads	1 – red	521	39,224,718	75,287.37	14.2%
	2 – yellow	2170	76,883,890	35,430.36	27.8%
	3 – blue	1672	25,533,749	15,271.38	9.2%
	4 – white	5627	92,841,986	16,499.38	33.6%
	5 – green	4680	41,880,043	8948.73	15.2%
	Total	14,670	276,364,386	Share of traffic in sample	95.5%
B roads	1 – red	86	1,613,758	18,764.63	22.6%
	2 – yellow	216	2,516,417	11,650.08	35.2%
	3 – blue	194	1,429,658	7369.37	20.0%
	4 – white	252	1,090,089	4325.75	15.2%
	5 – green	284	501,520	1765.92	7.0%
	Total	1032	7,151,442	Share of traffic in sample	2.5%
C roads	1 – red	59	886,950	15,033.05	24.7%
	2 – yellow	207	1,561,951	7545.66	43.5%
	3 – blue	147	660,926	4496.10	18.4%
	4 – white	218	192,124	881.30	5.3%
	5 – green	427	290,185	679.59	8.1%
	Total	1058	3,592,136	Share of traffic in sample	1.2%
U Roads	1 – red	31	269,195	8683.71	12.3%
	2 – yellow	187	594,486	3179.07	27.2%
	3 – blue	557	739,595	1327.82	33.9%
	4 – white	1070	509,433	476.11	23.3%
	5 – green	196	69,499	354.59	3.2%
	Total	2041	2,182,208	Share of traffic in sample	0.8%
ALL ROADS	Overall Total	18,801	289,290,172	Share of all road traffic	100%

Table A4
Total traffic share by road class.

Road Class	Traffic Volume – Vehicle Miles Travelled (VMT) in billions	Share
A	144.9	57%
B	22.9	9%
C	52.5	20%
U	35	14%
Total	255.3	100%

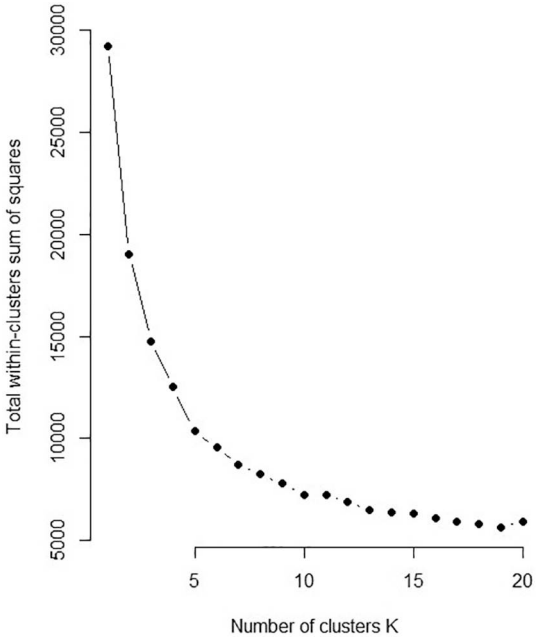

Fig. A1. "Elbow" method indicating K = 5.

References

Aditjandra, P.T., Cao, X., Mulley, C., 2012. Understanding neighbourhood design impact on travel behaviour: an application of structural equations model to a British metropolitan data. Transp. Res. A Elsevier Ltd 46 (1), 22–32.

Akhanli, S.E., Hennig, C., 2017. Some issues in distance construction for football players performance data. Arch. Data Sci. 2 (1), 1–17.

Apronti, D., Ksaibati, K., Gerow, K., Hepner, J.J., 2016. Estimating traffic volume on Wyoming low volume roads using linear and logistic regression methods. J. Traffic Transp. Eng. (English Edition) 3 (6), 493–506 Elsevier Ltd.

Arnott, R., Inci, E., 2006. An integrated model of downtown parking and traffic congestion. J. Urban Econ. 60 (3), 418–442.

Arnott, R., Williams, P., 2017. Cruising for parking around a circle. Transp. Res. B, Elsevier Ltd 104, 357–375.

Bacher, J., Wenzig, K., Volger, M., 2004. SPSS TwoStep Cluster - a First Evaluation. Arbeits- Und Diskussionspapiere / Universität Erlangen-Nürnberg, *Sozialwissenschaftliches Institut, Lehrstuhl Für Soziologie*, Amsterdam available at: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-327153.

Basak, D., Pal, S., Patranabis, D.C., 2007. Support vector regression. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 10634. pp. 699–708 LNCS No. 10.

Bholowalia, P., Kumar, A., 2014. EBK-means: a clustering technique based on elbow method and K-means in WSN. 105 (9), 17–24.

Bibby, P., Brindley, P., 2014. 2011 Rural-Urban Classification of Local Authority Districts in England: User Guide. No. December, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/389780/RUCLAD2011_User_Guide.pdf.

Bishop, C.M., 2013. In: Jordan, M., Kleinberg, J., Scholkopf, B. (Eds.), Pattern Recognition and Machine Learning. Springer, Cambridge, UK available at. https://doi.org/10.1117/1.2819119.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Caceres, N., Romero, L.M., Benitez, F.G., 2012. Estimating traffic flow profiles according to a relative attractiveness factor. Procedia Soc. Behav. Sci. 54, 1115–1124.

Caceres, N., Romero, L.M., Morales, F.J., Reyes, A., Benitez, F.G., 2018. Estimating traffic volumes on intercity road locations using roadway attributes, socioeconomic features and other work-related activity characteristics. *Transportation*, Springer US 45 (5), 1449–1473.

Cardozo, O.D., García-Palomares, J.C., Gutiérrez, J., 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. Appl. Geogr., Elsevier Ltd 34, 548–558.

Castro-Neto, M., Jeong, Y., Jeong, M.K., Han, L.D., 2009. AADT prediction using support vector regression with data-dependent parameters. Exp. Syst. App. Elsevier Ltd 36 (2), 2979–2986 PART 2.

Cervero, R., 1994. Transit-based housing in California: evidence on ridership impacts. Transp. Policy 1 (3), 174–183.

Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. Transp. Res. Part D: Transp. Environ. 2 (3), 199–219.

Chen, L., Wang, S., 2013. Central clustering of categorical data with automated feature weighting. In: *IJCAI International Joint Conference on* Artif. Intell. pp. 1260–1266.

Çodur, M.Y., Tortum, A., 2015. An artificial neural network model for highway accident prediction: a case study of Erzurum, Turkey. PROMET - Traffic&Transportation 27 (3), 217–225.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Das, S., Tsapakis, I., 2019. Interpretable machine learning approach in estimating traffic volume on low-volume roadways. Int. J. Transp. Sci. Technol. https://doi.org/10.1016/j.ijtst.2019.09.004. Tongji University and Tongji University Press.

Department for Transport, 2013. Road Traffic Estimates Methodology Note. available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524848/annual-methodology-note.pdf.

Department for Transport, 2014. Road Traffic Estimates. available at: https://www.gov.uk/government/publications/road-traffic-estimates-great-britain-jan-to-mar-q1-2014%5Cnhttps://www.gov.uk/government/collections/road-traffic-statistics.

Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.

Doustmohammadi, M., Anderson, M., 2016. Developing direct demand AADT forecasting models for small and medium sized urban communities. Int. J. Traffic Transp. Eng. 5 (2), 27–31.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. In: 9th International Conference on Neural Information Processing Systems. MIT Press, Cambridge, MA, USA, Denver, Colorado, pp. 155–161.

Efron, B., 1979. Bootstrap methods: another look at the Jackknife. Ann. Stat. 7 (1), 1–26.

Eom, J.I.N.K.I., Man, S.I.K.P., Heo, T.-Y., Huntsinger, L.F., 2006. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. Transp. Res. Rec. 1968, 22–29.

Faith, D.P., Minchin, P.R., Belbin, L., 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69 (1–3), 57–68.

Fricker, J., Saha, S., 1987. Traffic Volume Forecasting Methods for Rural State Highways. West Lafayette, Indiana.

Friedman, J.H., Meulman, J.J., 2004. Clustering objects on subsets of attributes. J. R Stat. Soc. B 66 (4), 815–839.

Fu, M., Kelly, J.A., Clinch, J.P., 2017. Estimating annual average daily traffic and transport emissions for a national road network: a bottom-up methodology for both nationally-aggregated and spatially-disaggregated results. J. Transp. Geogr. Elsevier Ltd 58, 186–195.

Gao, S., Wang, Y., Gao, Y., Liu, Y., 2013. Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. Environ. Planning B 40 (1), 135–153.

Gebotys, C.H., Elmasry, M.I., 1989. Integration of algorithmic VLSI synthesis with testability incorporation. IEEE J. Solid State Circuits 24 (2), 409–416.

Gecchele, G., Rossi, R., Gastaldi, M., Caprini, A., 2011. Data mining methods for traffic monitoring data analysis: a case study. Procedia Soc. Behav. Sci. 20, 455–464.

Gower, J.T., 1971. A general coefficient of similarity and some of its properties. Int. Biom. Soc. 27 (4), 857–871.

Greenacre, M., Primicerio, R., 2013. Michael Greenacre Raul Primicerio Multivariate Analysis of Ecological Data. Fundacion BBVA, Bilbao, Spain.

Gutiérrez, J., Cardozo, O.D., García-Palomares, J.C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. J. Transp. Geogr. 19 (6), 1081–1092.

Hammah, R.E., Curran, J.H., 1999. On distance measures for the fuzzy K-means algorithm for joint data. Rock Mech. Rock. Eng. 32 (1), 1–27.

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques, 3rd Editio. Morgan Kaufmannhttps://doi.org/10.1016/b978-0-12-381479-1.00001-0. available at:.

Hastie, T., Tibshirani, R., Friedman, J., 2011. The Elements of Statistical Learning, 2nd ed. https://doi.org/10.1007/978-0-387-84858-7. Elements, available at:.

He, Z., Deng, S., Xu, X., 2006. Approximation algorithms for K-modes clustering. In: Huang, D.-S., Li, K., Irwin, G.W. (Eds.), Computational Intelligence. Kumming, China, pp. 296–302.

Hess, D., 2001. Effect of free parking on commuter mode choice: evidence from travel diary data. Transp. Res. Rec. 1753 (1), 35–42.

Hesse, M., 2013. Cities and flows: re-asserting a relationship as fundamental as it is delicate. J. Transp. Geogr. Elsevier Ltd 29, 33–42.

Huang, Z., 1997a. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data. (DMKD).

Huang, Z., 1997b. Clustering large data sets with mixed numeric and categorical values. Vol. 3, 2303–2307.

Huang, Z., 1998. Extensions to the k -means algorithm for clustering large data sets with categorical values. Vol. 304, 283–304.

Huang, J.Z., Ng, M.K., Rong, H., Li, Z., 2005. Automated variable weighting in k-means type clustering. IEEE Trans. Pattern Anal. Mach. Intell. 27 (5), 657–668.

Hummel, M., Edelmann, D., Kopp-Schneider, A., 2017. Clustering of samples and variables with mixed-type data. PLoS One 12 (11), 1–23.

Inci, E., van Ommeren, J., Kobus, M., 2017. The external cruising costs of parking. J. Econ. Geogr. 17 (6), 1301–1323.

Jahanshahi, K., Jin, Y., 2016. The built environment typologies in the UK and their influences on travel behaviour: new evidence through latent categorisation in structural equation modelling. Transp. Plan. Technol. 39 (1), 59–77.

Jayasinghe, A., Sano, K., Nishiuchi, H., 2015. Explaining traffic flow patterns using centrality measures. Int. J. Traffic Transp. Eng. 5 (2), 134–149.

Jiang, B., Liu, C., 2009. Street-based topological representations and analyses for predicting traffic flow in GIS. Int. J. Geogr. Inf. Sci. 23 (9), 1119–1137.

Junqué de Fortuny, E., Martens, D., Provost, F., 2013. Predictive modeling with big data: is bigger really better? Big Data 1 (4), 215–226.

Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data - An Introduction to Cluster Analysis. John Wiley & Sons, Hoboken, New Jersey. https://doi.org/10.1002/9780470316801. available at:.

Kelly, J.A., Clinch, J.P., 2009. Temporal variance of revealed preference on-street parking price elasticity. Transp. Policy, Elsevier 16 (4), 193–199.

Kim, S., Park, D., Heo, T.-Y., Kim, H., Hong, D., 2016. Estimating vehicle miles traveled (VMT) in urban areas using regression kriging. J. Adv. Transp. https://doi.org/10.1002/atr. available at.

Kodinariya, T., Makwana, P., 2013. Review on determining number of cluster in K-means clustering. 1 (6), 90–95.

Koperski, K., Han, J., Stefanovic, N., 1998. An efficient two-step method for classification of spatial data. In: The Eighth International Symposium on Spatial Data Handling (SDH'98), https://doi.org/10.1007/s13398-014-0173-7.2. available at:.

Koul, A., Becchio, C., Cavallo, A., 2018. Cross-validation approaches for replicability in psychology. Front. Psychol. 9 (JUL), 1–4.

Labib, S.M., Neema, M.N., Rahaman, Z., Patwary, S.H., Shakil, S.H., 2018. Carbon dioxide emission and bio-capacity indexing for transportation activities: a methodological development in determining the sustainability of vehicular transportation systems. J. Environ. Manag. Elsevier 223 (May), 57–73.

Larose, D.T., 2005. Discovering knowledge in data: an introduction to data mining. Discov. Knowl. Data. https://doi.org/10.1002/0471687545. available at:.

Leduc, G., 2008. Road traffic data: collection methods and applications. In: EUR Number: Technical Note: JRC 47967. Vol. JRC 47967. pp. 55 No. January 2008.

Loyd, C.D., 2007. Local Models for Spatial Analysis. CRC Press, Boca Raton, Florida.

McCune, B., Grace, J., 2002. Distance measures. Anal. Ecol. Commun. 45–57.

Mohamad, I. Bin, Usman, D., 2013. Standardization and its effects on K-means clustering algorithm. Res. J. Appl. Sci. Eng. Technol. 6 (17), 3299–3303.

Mohamad, D., Sinha, K., Kuczek, T., Scholer, C., 1998. Annual average daily traffic prediction model for county roads. Transp. Res. Rec. 1617 (98), 69–77.

Morley, D.W., Gulliver, J., 2016. Methods to improve traffic flow and noise exposure estimation on minor roads. Environ. Pollut. Elsevier Ltd 216, 746–754.

Neveu, A.J., 1983. Quick-response procedures to forecast rural traffic. Transp. Res. Rec. 944, 47–53.

Opsahl, T., Panzarasa, P., 2009. Clustering in weighted networks. Soc. Networks 31 (2), 155–163.

Patarasuk, R., Gurney, K.R., O'Keeffe, D., Song, Y., Huang, J., Rao, P., Buchert, M., et al., 2016. Urban high-resolution fossil fuel CO2emissions quantification and exploration

of emission drivers for potential policy applications. Urban Ecosyst. Urban Ecosyst 19 (3), 1013–1039.
Pointer, G., 2005. The UK's major urban areas. In: Chappell, R. (Ed.), *Focus on People and* Migration, pp. 45–60.
Puliafito, S.E., Allende, D., Pinto, S., Castesana, P., 2015. High resolution inventory of GHG emissions of the road transport sector in Argentina. Atmos. Environ. 101, 303–311.
Pun, L., Zhao, P., Liu, X., 2019. A multiple regression approach for traffic flow estimation. *IEEE Access*, IEEE 7, 35998–36009.
Roess, R., Prassas, E., McShane, W., 2011. Traffic Engineering. *Pearson*https://doi.org/10.1017/CBO9781107415324.004. available at:.
Rokach, L., Maimon, O., 2015. Data mining with decison trees: theory & application. 91 (479), 399–404.
Sarlas, G., Axhausen, K.W., 2014. Towards a direct demand modeling approach. In: 14th Swiss Transport Research Conference, available at: http://www.strc.ch/conferences/2014/Sarlas_Axhausen.pdf%5Cnhttp://www.ivt.ethz.ch/vpl/publications/#997.
Selby, B., Kockelman, K.M., 2013. Spatial prediction of traffic levels in unmeasured locations: Applications of universal kriging and geographically weighted regression. J. Transp. Geogr. Elsevier Ltd 29, 24–32.
Shojaeshafiei, M., Doustmohammadi, M., Subedi, S., Anderson, M., 2017. Comparison of estimation methodologies for daily traffic count pediction in small and medium sized communities. Int. J. Traffic Transp. Eng. 6 (4), 71–75.
Shoup, D.C., 2006. Cruising for parking. Transp. Policy 13 (6), 479–486.
Silva, J. de A., Morency, C., Goulias, K.G., 2012. Using structural equations modeling to unravel the influence of land use patterns on travel behavior of workers in Montreal. Transp. Res. A Policy Pract. 46 (8), 1252–1264.
Stead, D., 2001. Relationships between land use, socioeconomic factors, and travel patterns in Britain. Environ. Planning B 28 (4), 499–528.
Wang, X., Kockelman, K., 2009. Forecasting network data. Transp. Res. Rec. 2105 (2105), 100–108.
Wang, T., Gan, A., Alluri, P., 2013. Estimating annual average daily traffic for local roads for highway safety analysis. Transp. Res. Rec. 5 (2398), 60–66.
Xia, Q., Zhao, F., Chen, Z., Shen, D., Ospina, D., 1999. Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County. 99. pp. 32–40.
Zhang, M., 2007. The role of land use in travel mode choice: evidence from Boston and Hong Kong. J. Am. Plan. Assoc. 70 (3), 344–360.
Zhang, Y., Cheng, T., Aslam, N.S., 2019. Exploring the relationship between travel pattern and social-demographics using smart card data and household survey. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-2/W13 No. June, pp. 1375–1382.
Zhao, F., Chung, S., 2001. Contributing factors of annual average daily traffic in a florida county: exploration with geographic information system and regression models. Transp. Res. Rec. 1769 (01), 113–122.
Zhao, F., Park, N., 2004. Using geograhically weighted regression models to estimate annual average daily traffic. Transp. Res. 99–107.
Zhao, S., Zhao, P., Cui, Y., 2017. A network centrality measure framework for analyzing urban traffic flow: a case study of Wuhan, China. Physica A Elsevier B.V. 478, 143–157.