Ersatzwertbildung für stationäre Verkehrsdetektoren

A Replacement Value Procedure for Stationary Traffic Detectors

Lukas Graf, Gernot Pucher, Sebastian Gatscha

Trafficon – Traffic Consultants GmbH · info@trafficon.eu

Zusammenfassung: Die korrekte und vollständige Erfassung von Verkehrsdaten stellt eine wichtige Grundlage zur Überwachung und Steuerung von Verkehrsströmen dar. Damit es auch bei Fehlfunktionen und Ausfällen einer Verkehrsmesstechnik nicht zu Störungen im Betrieb abhängiger Systeme kommt, werden Ersatzwertverfahren eingesetzt. In diesem Beitrag wird ein clusterbasiertes Verfahren für die Ersatzwertbildung von Verkehrszählungen stationärer Verkehrsdetektoren vorgeschlagen. Dabei werden sowohl statistische Merkmalsausprägungen als auch räumliche Aspekte berücksichtigt. Die absolute Höhe sowie die zeitliche Dynamik der Verkehrszählungen konnten mit hoher Genauigkeit aus Ersatzwerten gebildet werden.

Schlüsselwörter: Verkehrsdetektion, Ersatzwertbildung, Datenqualität

Abstract: *The correct and complete recording of traffic counts constitutes an important basis for controlling and monitoring traffic flows. To avoid disruptions of dependent systems during malfunctions or failures of measuring equipment, replacement value procedures are applied. In this paper, a cluster-based approach for identifying adequate replacement values for traffic counts from stationary traffic detectors is introduced. Statistical and spatial aspects are considered in the approach. Both absolute values as well as temporal dynamics of traffic counts could be determined with high accuracy.*

Keywords: *Traffic detection, replacement values, traffic data quality*

1 Motivation und Stand der Technik

Zunehmende Verkehrsmengen führen immer häufiger zur Überlastung der vorhandenen Verkehrsinfrastruktur. Dies ist insbesondere zu Spitzenzeiten entlang wichtiger Verkehrsachsen und in dicht verbauten, urbanen Räumen der Fall. Dadurch entstehen Zeitverluste, verstärkte verkehrsbedingte Emissionen und hohe volkswirtschaftliche Kosten (Schrank et al., 2012).

Die Überwachung und Steuerung von Verkehrsströmen können dazu beitragen, den Verkehr effizienter zu lenken und Überlastungen zu verhindern oder abzuschwächen (Otterstätter, 2013). Eine entsprechende Datengrundlage wird häufig aus stationären Detektormessungen erzeugt. Dabei wird die Anzahl der Fahrzeuge, die über einen Messzeitraum hinweg einen Straßenquerschnitt durchfahren, gemessen. Abhängig vom Messgerät werden unter Umständen auch weitere Daten erfasst, wie etwa Fahrzeugklassen oder Geschwindigkeiten (Fröhlich et al., 2018).

Eine zumindest temporär falsche oder fehlende Datenerfassung bzw. -übertragung von Verkehrsdetektoren kann jedoch zu fehlerhaften Informationen führen. Dadurch kann beispielsweise die korrekte Funktion von Verkehrsbeeinflussungsanlagen beeinträchtigt sein. In solchen Fällen wird daher häufig auf Ersatzwerte zurückgegriffen (Lehnhoff, 2005).

AGIT – Journal für Angewandte Geoinformatik, 5-2019, S. 184-191. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-669-7, ISSN 2364-9283, eISSN 2509-713X, doi:10.14627/537669017.

Für die Bildung von Ersatzwerten in Zeitreihen steht eine große Bandbreite statistischer Verfahren zur Verfügung. Diese lassen sich unterscheiden in Algorithmen, welche alleine die zeitliche Dimension berücksichtigen (Chang & Ge, 2011) oder raumzeitliche Aspekte mit einbeziehen (Bae et al., 2018). Des Weiteren kann eine Untergliederung in selbstlernende Verfahren – wie die in diesem Artikel vorgestellte Cluster-Analyse – und Algorithmen zur überwachten Ersatzwertbildung vorgenommen werden. Grundannahme bei allen Verfahren ist, dass Ersatzwerte unter Verwendung einer stochastischen Verteilung aus der Grundgesamtheit aller vorliegenden Werte der betrachteten Größe abgeleitet werden können (Donders et al., 2006). Exemplarisch für solcherlei Verfahren sind Autoregressions- und Moving Average Modelle (Park et al., 2007), sowie nicht-parametrische Verfahren wie die Cluster-Analyse (Zhang et al., 2008).

Tang et al. (2015) konnten zeigen, dass clusterbasierte Algorithmen gegenüber anderen Verfahren – etwa Autoregressions-Ansätzen – zu einer höheren Genauigkeit bei der Ersatzwertbildung führen können, da nichtlineare Eigenschaften von Verkehrsdaten abgebildet werden. Der in dieser Arbeit verwendete Fuzzy-C-Means Clustering (FCM) Algorithmus (Bezdek et al., 1981; Bezdek et al., 1984; Cannon et al., 1986) unterscheidet sich von deterministischen Cluster-Verfahren – etwa dem k-Means-Clustering – durch die Verwendung von Zugehörigkeitswahrscheinlichkeiten, die Rückschlüsse auf die Stabilität der Clusterung und deren Unsicherheiten geben. Somit kann neben dem Erzielen einer hohen Genauigkeit die statistische Stabilität der Ersatzwertbildung erfasst werden. Aufgrund dessen erscheint der FCM-Ansatz als geeignete Methode der Ersatzwertbildung.

2 Methode

2.1 Ersatzwertbildung für stationäre Verkehrsdetektoren

Die von den einzelnen Detektoren gemessenen Rohdaten werden zunächst auf Vollständigkeit und Plausibilität geprüft, um Datenlücken zu erkennen und unrealistische Werte zu identifizieren. Bei der Plausibilisierung werden die Rohdaten sowohl automatisiert gegen einen Konfidenzbereich getestet als auch manuell überprüft. Unplausible Werte werden aus der Gesamtheit der Messungen entfernt und folglich wie Fehlwerte behandelt. Die daraus resultierenden unvollständigen Zeitserien bilden die Ausgangsdatenbasis für die Clusterung.

Vor der eigentlichen Clusterung erfolgt eine Vorgruppierung der Daten durch die Verwendung von Vorwissen und unter Berücksichtigung der verwendeten Detektortechnik, da je nach Messgerät die Daten in unterschiedlicher Granularität vorliegen. Im Falle von Verkehrsdaten ist bekannt, dass sich das Verkehrsaufkommen an Werktagen von dem an Wochenenden unterscheidet. Selbiges gilt für Urlaubszeiten. Zudem können Unterscheidungen in einzelne Fahrzeuggruppen (beispielsweise PKW und LKW) getroffen werden, um die Diversität des Verkehrsaufkommens in die Analyse aufzunehmen. Die Berechnung von Ersatzwerten erfolgt folglich nur auf Messdaten einer Gruppe.

Da die Anzahl der Cluster die Qualität des Ergebnisses maßgeblich beeinflusst, wird die „optimale“ Anzahl an Clustern mit der sogenannten Ellbow-Methode (Bholowalia & Kumar, 2014) grafisch ermittelt, bei der die durch die Clusterung erklärte Varianz gegen die Anzahl der Cluster aufgetragen wird. Ziel ist es, einerseits die Anzahl der Cluster so gering wie möglich zu halten und andererseits den Anteil der durch die Clusterung erklärte Varianz an der

Gesamtvarianz des Datensatzes zu maximieren. Das Optimum findet sich an dem Punkt im Ellbow-Plot, an dem die erklärte Varianz in die Sättigung übergeht.

Mit der so ermittelten Cluster-Zahl wird die eigentliche Clusterung mit dem Fuzzy-C-Means(FCM)-Algorithmus durchgeführt. Dabei wird der Ausgangsdatensatz (Zeitreihen aller Detektoren in einem bestimmten Zeitraum) in *x* Cluster aufgeteilt, wobei die Zuordnung einer Zeitreihe anhand von Zugehörigkeitswahrscheinlichkeiten erfolgt (Bezdek et al., 1984). Als Ähnlichkeitskriterium wird die euklidische Distanz der Zeitreihen berechnet. Der Prototyp eines Clusters ist repräsentativ für alle Zeitreihen, die dem Cluster angehören (Tan et al., 2013). Cluster-Prototypen können somit als Ersatzwerte verwendet werden.

Der FCM-Algorithmus kann allerdings nicht direkt angewandt werden, da dieser als Input vollständige Datensätze ohne Fehlwerte erwartet (Hathaway & Bezdek, 2001). Um dieses Problem zu umgehen wird die Clusterung mit einem iterativen Expectation Maximation (EM) Ansatz kombiniert, welcher eine statistische Methode zur Abschätzung der Maximum Likelihood in Grundgesamtheiten darstellt (Nasser et al., 2006). Die Fehlwerte werden in der ersten Iteration mittels einem parametrischen Imputationsverfahren (gleitender Mittelwert) ersetzt und die Clusterung durchgeführt. In den folgenden Runden werden die Fehlwerte mit den Werte des entsprechenden Clusterprototypen imputiert und die Clusterung mit den Prototypen der vorherigen Iteration neu initialisiert. Dies wird *n*-mal wiederholt, um eine Konvergenz der Clusterung gegen ein globales Optimum zu gewährleisten. Die finalen Prototypen werden folglich als Ersatzwerte für Datenlücken eingesetzt.

Die Fuzzy-Wahrscheinlichkeiten bieten darüber hinaus die Möglichkeit, die Stabilität der Clusterung zu bewerten. Ist beispielsweise bei einer hohen Anzahl von Zeitreihen keine eindeutige Zuweisung zu einem Cluster möglich, so kann die Clusterung nochmals neu durchgeführt und etwa die Anzahl der Iterationen bei der EM-Methode angepasst werden.

Ein schematischer Überblick über die obigen Schritte zur Ersatzwertbildung ist in Abbildung 1 dargestellt.

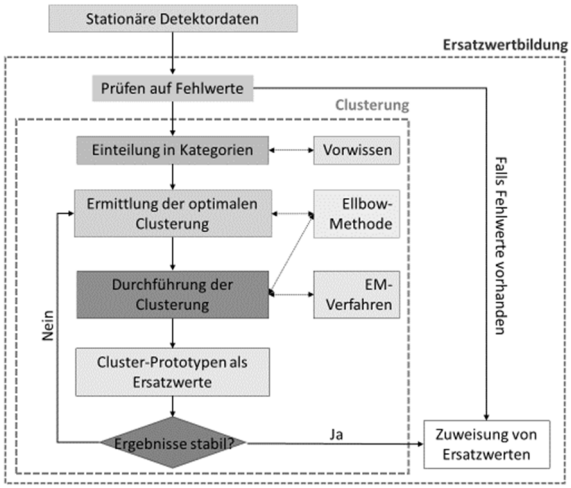

Abb. 1:
Schematische Darstellung des Algorithmus zur Ersatzwertbildung

2.2 Implementierung in R

Der in 2.1 beschriebene Ablauf wurde in R (Version 3.5.1) implementiert. Die Detektordaten werden unter Berücksichtigung der getroffenen Vorgruppierungen aus einer relationalen Datenbank abgefragt. Die FCM-Clusterung erfolgt mittels des „ppclust"-Packages (Cebeci et al., 2018).

2.3 Datengrundlage

Zur Methodenentwicklung wurde jeweils Messdaten aus einem Induktionsschleifengeräte und einem Seitenradargerät herangezogen, welche entlang von Bundes- und Landesstraßen in Österreich lokalisiert sind. Ersatzwerte wurden dabei für stündliche Messintervalle der Fahrzeuganzahl je Straßenquerschnitt ermittelt. Die Messdaten sind dabei nach Zählstelle, Fahrtrichtung, Fahrspur und Fahrzeuggruppen gemäß RVS 02.01.12 „Straßenverkehrszählungen" differenziert (FSV, 2015). Je Zählstelle wurden Daten von Juni 2017 bis Juni 2018 betrachtet.

2.4 Evaluierung der Ergebnisse

Die Evaluierung der Ersatzwertbildung erfolgt mittels einer Fehlerstatistik. Hierbei wird in einem Zeitraum der Länge n für jeden Messzeitpunkt i ($1 \leq i \leq n$) ein Abgleich zwischen den Messwerten y_i und den durch die Clusterung vorgeschlagenen Ersatzwerten $\dot{y}_l$ durchgeführt. Ermittelt wird jeweils die Wurzel des mittleren quadratischen Fehlers (Root Mean Square Error, *RMSE*), dessen Formel in (1) aufgeführt ist:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \dot{y}_l)^2} \tag{1}$$

Aus (1) wird darüber hinaus mittels Division des RMSE durch das arithmetische Mittel $\bar{y}$ aller Messwerte im Zeitraum der normalisierte RMSE (*NRMSE*) abgeleitet, wie in (2) notiert:

$$NRMSE = \frac{RMSE}{\bar{y}} \tag{2}$$

Zusätzlich erfolgt eine visuelle Beurteilung der Ergebnisse, um zu überprüfen, ob die Ersatzwerte die temporale Charakteristik des Verkehrsaufkommens adäquat abbilden können.

3 Ergebnisse

3.1 Qualität der Ersatzwertbildung

Aufgrund der Größe der zugrunde liegenden Verkehrsdatenbank erfolgt die Bewertung der Qualität der Ersatzwertbildung anhand zweier zufällig ausgewählter Detektoren a und b, für welche eine Tagesganglinie mit gemessenen und durch die Clusterung vorgeschlagenen Werten vorliegt (siehe Abb. 2). Der RMSE liegt bei 32 Fahrzeugen pro Stunde bei Detektor a (NRMSE: 27 %) und bei 44 Fahrzeugen pro Stunde bei Detektor b (NRMSE: 31 %).

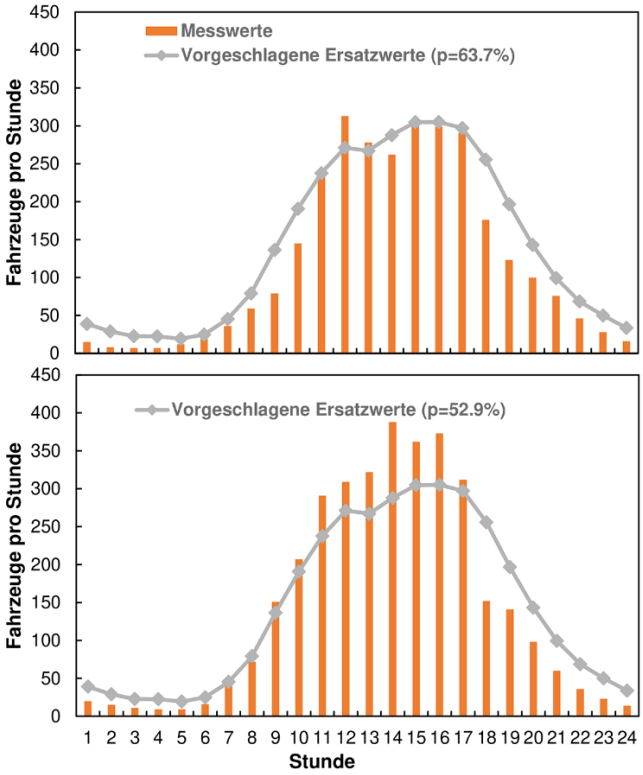

Abb. 2: Darstellung der gemessenen Anzahl der Fahrzeuge pro Stunde (Säulen) und die vorgeschlagenen Ersatzwerte (Linien) mit Fuzzy-Zugehörigkeitswahrscheinlichkeit zu einem Cluster in Klammern für zwei Detektoren a und b für einen zufällig ausgewählten Tag

Die ermittelten Fuzzy-Zugehörigkeitswahrscheinlichkeiten von 63.7 % bei Detektor a und 52.9 % bei Detektor b lassen auf eine stabile Zuweisung zu dem entsprechenden Cluster schließen. Wie Abb. 2 zeigt, folgt die Form der Ersatzwertkurve im Falle beider Detektoren dem Tagesverlauf der Messwerte. Zudem sind die Beträge der beobachteten Abweichungen zwischen Mess- und Ersatzwerten in einem vergleichsweisen geringen Bereich. Verkehrsspitzen am Mittag und frühen Nachmittag werden bei b unterschätzt, sowie der Verkehr in den Abendstunden bei beiden Detektoren überschätzt.

3.2 Räumliche Aspekte

Es zeigt sich, dass Detektoren, die entlang einer Straße angeordnet sind, häufig demselben Cluster angehören. Das Clusterverfahren spiegelt somit zu erwartende räumliche Autokorrelationen wider. Adjazenz im Straßengraphen kann zudem genutzt werden, die Clusterung durch Ausnutzung räumlicher Beziehung weiter zu verbessern. Durch die Verwendung räumlicher Cluster als initiale Prototypen konnte in weiterführenden Untersuchungen der Fehler (RMSE und NRMSE) zwischen den vorgeschlagenen Ersatz- und den Messwerten weiter verringert werden.

3.3 Laufzeit

Durch die große Datenmenge und die Vielzahl an möglichen Untergruppen steigt die Laufzeit des Algorithmus stark an. Dies kann jedoch durch die parallele Berechnung von Ersatzwerten verschiedener Untergruppen in betriebsarmen Randzeiten (Wochenenden) vermindert werden. Zudem muss eine Neuberechnung der Ersatzwerte nur in vergleichsweise langen Zeitintervallen erfolgen, sodass einer einmalig hohen Laufzeit eine lange Verwendungsperiode der Ergebnisse gegenübersteht.

4 Diskussion

Wie die Ergebnisse aus 3.1 zeigen, ist das vorgestellte Ersatzwertverfahren in der Lage die Charakteristik des Verkehrsaufkommens aufgeschlüsselt nach Fahrzeuggruppen und Tageskategorien sowohl in absoluter Höhe der Werte als auch in der temporalen Dynamik darzustellen. Die Verwendung des Cluster-Prototypen als Ersatz für fehlende Werte in stationären Detektordaten erscheint daher möglich. Durch die Ermittlung der Clusteranzahl mit der Ellbow-Methode und die Verwendung des EM-Verfahrens (2.1) kann der Algorithmus zudem flexibel mit unterschiedlichen Datengrundlagen umgehen und bedarf keiner Parametrisierung durch den Nutzer.

Problematisch an dem Verfahren ist die je nach Größe der zu clusternden Daten im Vergleich zu anderen Imputationsverfahren (etwa Moving Average Imputation) die hohe Laufzeit. Zudem bewirkt die Verwendung von Cluster-Prototypen zwangsläufig eine starke Generalisierung der Ersatzwerte, sodass nur selten auftretende Verkehrsspitzen oder Minima unter Umständen nicht adäquat rekonstruiert werden können. Weiterhin ist die anfängliche Imputation von Fehlwerten mittels gleitender Mittelwertbildung ein kritischer Punkt, da Extrema das Ergebnis der Clusterung beeinflussen können. Alternativ kann stattdessen der Median verwendet werden.

Das Ergebnis der Clusterung reflektiert topologische Beziehungen der Detektoren (3.2). Wie in 3.2 beschrieben wurden in einem weiteren Schritt die initialen Clusterprototypen auf Basis von Adjazenz im Straßengraphen ermittelt. Neben einer systematischen Evaluierung der Veränderungen in den Ersatzwerten durch diesen Ansatz ist zu klären, wie die Granularität der Cluster optimiert werden kann, da durch die Verwendung von Adjazenz-Matrizen in Abhängigkeit vom gewählten Nachbarschaftskriterium die Zahl der Cluster sehr schnell ansteigen kann.

Weiterer Forschungsbedarf besteht des Weiteren in der Frage, welchen Einfluss das verwendete EM-Verfahren auf die Qualität der Ersatzwertbildung hat und ob in diesem Kontext andere Ansätze zur Clusterung unvollständiger Datensätze wie etwa das Partial-Distance-Verfahren (Hathaway et al., 2001) zu einer höheren Qualität der Ersatzwertbildung führen können. Zu klären ist ebenfalls, ob die Einbindung weiterer topologischer Information aus dem Straßengraphen Verbesserungen bewirkt.

Literatur

Bae, B., Kim, H., Lim, H., Liu, Y., Han, L., & Freeze, P. (2018). Missing data imputation for traffic flow speed using spatio-temporal cokriging. *Transportation Research Part C, 88*(February), 124–139. Retrieved Jan 18, 2019, from: doi:10.1016/j.trc.2018.01.015.

Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981). Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines. *SIAM Journal on Applied Mathematics*, *40*(2), 339–357.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM : The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2), 191–203.

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 17-24.

Cannon, R., Dave, J., & Bezdek, J. (1986). Efficient Implementation of the Fuzzy c-Means Clustering Algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, *PAMI-8*(2), 248-255.

Cebeci, Z., Yildiz, F., A., K., Cebeci, C., & Onder, H. (2018). *{ppclust}: Probabilistic and Possibilistic Cluster. CRAN-R-project*. Retrieved Jan 18, 2019, from https://cran.r-project.org/web/packages/ppclust/index.html.

Chang, G., & Ge, T. (2011). Comparison of missing data imputation methods for traffic flow. *Proceedings 2011 International Conference on Transportation, Mechanical and Electrical Electrical Engineering (TMEE)*, 639-642. Retrieved Jan 18, 2019, from: doi:10.1109/TMEE.2011.6199284.

Donders, A., van der Heijdenc, G., Stijnend, T., & Moons, K. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*, 1087–1091. Retrieved Jan 17, 2019, from: doi:10.1016/j.jclinepi.2006.01.014.

Forschungsgesellschaft Straße – Schiene – Verkehr (2015). Richtlinien und Vorschriften für das Straßenwesen. *(RVS) 02.01.12 Straßenverkehrszählungen.* Retrieved Oct 14, 2018, from: http://www.fsv.at/shop/produktdetail.aspx?IDProdukt=e05921e7-fab4-4cb6-8378-95bd039fdff3.

Fröhlich S., Pape S., Gassel C., & Krimmling J. (2018). Nutzung von Verkehrsdaten zur Verkehrsflussoptimierung. In: M. Wiesche, P. Sauer, J. Krimmling J., & H. Krcmar (Eds.), *Management digitaler Plattformen*. *Informationsmanagement und digitale Transformation*. Wiesbaden: Springer Gabler. Retrieved Jan 17, 2019, from: doi.org/10.1007/978-3-658-21214-8_7

Hathaway, R. J., & Bezdek, J. C. (2001). Fuzzy c-Means Clustering of Incomplete Data. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, *31*(5), 735–744.

Lehnhoff, N. (2005). *Überprüfung und Verbesserung der Qualität von automatisch erhobenen Daten an Lichtsignalanlagen* (Dissertation). Fakultät für Bauingenieurwesen und Geodäsie der Universität Hannover. Retrieved Jan 18, 2019, from: d-nb.info/976146703/34.

Nasser, S., Alkhaldi, R., & Vert, G. (2006). A Modified Fuzzy K-means Clustering using Expectation Maximization. *2006 IEEE International Conference on Fuzzy Systems*, 231–235. Retrieved Jan 18, 2019, from: doi:10.1109/FUZZY.2006.1681719.

Otterstätter, T. (2013). *Methoden zur Erfassung von Verkehrsströmen und Fahrzeiten mit stationären fahrzeugwiedererkennenden Detektoren* (Dissertation). Universität Stuttgart, Institut für Straßen- und Verkehrswesen. Retrieved Jan 17, 2019, from: https://elib.uni-stuttgart.de/bitstream/11682/525/1/Dissertation_Otterstaetter_BiB.pdf.

Park, J. W., Genton, M. G., & Ghosh, S. K. (2007). Censored time series analysis with autoregressive moving average models. *Canadian Journal of Statistics, 35*(1), 151–168.

Schrank, D., Eisele, B., & Lomax, T. (2012). *TTI's 2012 Urban Mobility Report. Powered by INRIX Traffic Data*. College Station, TX, Texas Transportation Institute. Retrieved Aug. 20, 2015, from: http://tti.tamu.edu/documents/mobility-report-2012.pdf.

Tan, P. N., Steinbach, M., & Kumar, V. (2013). Cluster Analysis: Basic Concepts and Algorithms. In: P. N. Tan, M. Steinbach, & V. Kumar (Eds.), *Introduction to Data Mining* (525–612). Retrieved Jan 18, 2019, from: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php.

Tang, J., Zhang, G., Wanga, Y., Wang, H., & Liu, F. (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transportation Research Part C, 51*, 29–40. Retrieved Jan 16, 2019, from: doi:10.1016/j.trc.2014.11.003.

Zhang, S., Zhang, J., Zhu, X., Qin, Y., & Zhang, C. (2008). Missing value imputation based on data clustering. *Transactions on computational science*, 128–138.