Clustering of Urban Road Paths; Identifying the Optimal Set of Linear and Nonlinear Clustering Features

Glykeria Myrovali[1](✉), Theodoros Karakasidis[2], Maria Morfoulaki[1], and Georgia Ayfantopoulou[1]

[1] Hellenic Institute of Transport/Centre for Research and Technology Hellas (HIT/CERTH), 6th Km Charilaou-Thermi Road, 57001 Thessaloniki, Greece
{myrovali, marmor, gea}@certh.gr

[2] Laboratory of Hydromechanics and Environmental Engineering, Department of Civil Engineering, University of Thessaly, 38334 Volos, Greece
thkarak@uth.gr

Abstract. Urban traffic is undoubtedly a dynamic phenomenon presenting variations over both time and space, that in the majority of cases are the result of a mixture of, either well known (i.e. weather, seasonality) or not easily predictable (i.e. events, accidents) external factors. Identification of similarities in the performance of different urban road paths under different traffic states (different travel demand conditions) is the main subject of the current paper. Floating taxi travel time data (timeseries per road path) collected in the framework of Thessaloniki Smart Mobility Living Lab (initiated and operated by CERTH/HIT) consist the basic input for the hierarchical clustering that is applied. Clustering applies upon different combinations of road paths' features (data points of travel time timeseries, descriptive statistics and mutual information of timeseries). The comparison of the clustering results based on average weekdays travel times per road path (from a six months period) with the respective results of a typical and an atypical day adds on the interpretability of underlying relations among paths under different states. The analysis reveals that resulting clusters can be a building block for the spatiotemporal understanding of urban traffic. Furthermore, it is shown that adding as clustering feature the criterion of mutual information of timeseries, therefore taking into account also non-linear dependences of the different road paths, the clustering interpretability is differentiated.

Keywords: Urban traffic state · Timeseries · Clustering · Mutual information criterion

1 Introduction

1.1 Urban Road Traffic; an Undoubtedly Dynamic System

Urban road traffic is undoubtedly a dynamic, therefore complex, system which becomes even more complex with the increase of city size and of the number of intersections (nodes) of the road network. The gradual (structured or even unstructured)

E. G. Nathanail et al. (Eds.): CSUM2020, AISC 1278, pp. 1107–1116, 2021.
https://doi.org/10.1007/978-3-030-61075-3_106

cities' evolution has increased the complexity of urban structures and functionalities, part of which is the urban transport network and urban road traffic. Road network structure (topology), socio-economic factors, weather effects, traffic management policies and mobility options, purpose of daily trips, mode choice, trips distribution during the day and drivers' behavior and choices are some of the aspects which contribute in system's complexity. The multiplicity of decisions (personal decisions of many discrete individuals, which are nevertheless strongly related) adds also in the complexity of urban traffic [1–8].

Urban traffic, being a non-static phenomenon, is subject to a series of variations related to both space and time. The former refers to the different traffic patterns per case (road path, sub-area, flow direction) while the latter relates to traffic variations over different time-periods:

- Short term variations caused by signaling (delays)
- Daytime variation- peak and non-peak hours
- Variation between days - weekdays, Saturday and Sunday
- Long-term variations, fluctuations over the years due to demographic, economic and geographical changes [9].

Under normal situations (recurrent and known patterns), when no special event has occurred, urban road systems operate in the expected way. When this balance is disturbed, with demand to exceed capacity ('*the maximum number of vehicles that can reasonably be expected to be served in the given time period*' [10]) or with the appearance of a special event, congestion phenomena are noticed [11, 12] while ex ante flexible and quickly adaptable traffic management schemes should have been mobilized. Spatiotemporal dynamics understanding is a main challenge for transport systems management. In this paper, complementing other spatiotemporal analysis, clustering analysis is estimated to be able to help identify operation scenarios that can support transportation system management. The key goal of the paper is not to compare the effectiveness of different clustering methodologies but to compare resulting clusters at different traffic states.

1.2 Aim and Structure

During the last decades, the large penetration of Intelligent Transport Systems in daily city's operations has changed the traditional way of collecting and analyzing data for transport planning and traffic management; large databases of traffic data have been built and are updated at real time. Floating car data, Bluetooth readers with application on mobility, cameras/radars and traffic related data mined from social media are jointly offering a broader picture of the traffic situation for the entire road network [13, 14].

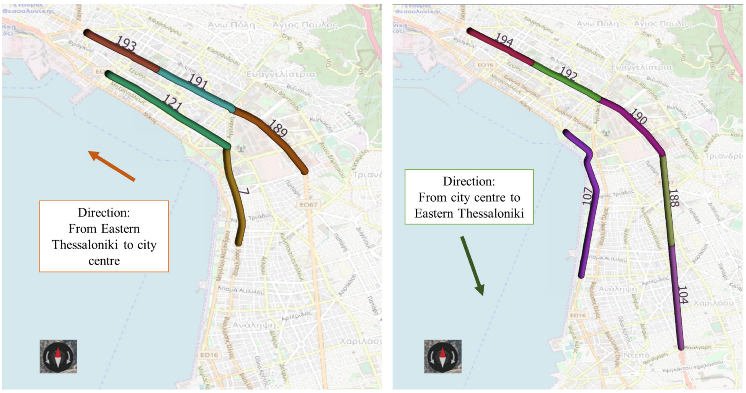
Taking advantage of the existence of such a large traffic databases (time series) in the framework of Thessaloniki Smart Mobility Living Lab (https://smartmlab.imet.gr/), the current paper aims at presenting two practical approaches in clustering urban road paths; clustering based on point values and clustering based on global features extracted from the time series. The scope per research orientation defines the effectiveness of the different approaches while the interpretability of the clusters (injection of prior knowledge under an intuitive way) acts as a measure of effectiveness and accuracy.

The remainder of the paper is organized as follows. Section 2 presents the reference area where analysis took place, the main urban road paths and the timeseries used for the clustering of the road paths. Section 3 presents in brief the methodological approach while results are presented in Sect. 4. Concluding the paper, Sect. 5 highlights key advantages and opportunities of the proposed approach.

2 The Reference Research Area

2.1 Thessaloniki's Urban Road Network

The reference area for the current work consists of basic road paths of the wider area of Thessaloniki, a medium size Greek city with a population of around 800.000 inhabitants (2011). High dependence of private vehicles use (50%) is characterizing the city transport (SUMP of Thessaloniki, www.svakthess.imet.gr). The selection of road paths follows wider research scopes (availability of different traffic datasets and already existing geo-reference from CERTH/HIT research activity [15–22]). The examined paths (with coding, direction, location) are presented in Fig. 1.


193
191
121
189
7
Direction:
From Eastern Thessaloniki to city centre
194
192
190
107
188
104
Direction:
From city centre to Eastern Thessaloniki

	path7	path104	path107	path121	path188	path190	path191	path192	path193	path194
Matlab coding	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Path length (km)	1.3	1.6	2.3	1.5	1.4	1.1	0.95	0.95	0.85	0.85
Direction	west	east	east	west	east	east	west	east	west	east
Number of lanes	Bus lane +2 operational		Bus lane +4 operational	Bus lane + 2 operational						
Subpart of greater axis*	V. Olgas	K. Karamanli	M. Alexandrou	Tsimiski	K. Karamanli	Egnatia	Egnatia	Egnatia	Egnatia	Egnatia
Daily Average Flow / Passenger Car Unitσ (PCU)	35.000	35.000	45.000		35.000 – 40.000					

* subparts oflarger horizontal road axis are highlighted in the same color

Fig. 1. Examined urban road paths, Thessaloniki (GR), and respective coding for the analysis in MATLAB software.

Figure 1 contains also information on the operational characteristics of the paths; as a general comment we can say that all road paths examined are parts of the three large horizontal road axis serving a large percentage of daily trips. With the exception of Path 107 all the other paths are of similar length (1–1.6 km) while consist of one bus lane plus two more operational lanes.

2.2 The Traffic Data

After verifying the relatively high accuracy and validity of the information contained in traffic data collected via 1200 floating taxi (GDPR respecting) in the study area, travel time taxi GPS based timeseries were used for the current analysis. Travel time timeseries per road path in 15 min intervals (reference period: first semester of 2017, hereinafter A2017) consisted the basis of the analysis– 13.966.878 timestamps were used for calculating average travel times based on CERTH/HIT methodology [20, 21].

3 The Hierarchical Clustering Approach

3.1 Methodological Steps

Data retrieval from taxi floating data and travel time data per road path were the input data received from CERTH/HIT – additional elaboration of the data took place for the current work:

1a. Travel time timeseries per examined road path in 15-min intervals for a long period were extracted, interpolation took place for completing the timeseries where necessary (10560 data points for A2017).
1b. Travel time timeseries of specific days in 15-min intervals (one typical – expected traffic variations and one atypical, day with extreme weather conditions) were elicited.
1c. Data aggregation for calculating average travel times per road path in 15-min intervals took place developing the average travel time timeseries;
2. Timeseries features were also calculated for the average travel time timeseries (1c).

Finally, clustering applied and comparison among the resulting clusters based on point data and on timeseries feature took place.

3.2 Hierarchical Clustering

Time series clustering remains a challenging issue partially attributed to its high sensitivity to input data. In its generic notion, cluster analysis or clustering is the grouping of objects that present similarities. The objects are placed into groups, based on specific measures and similarity criteria according to the following principles:

- In each created group, the close observations resemble each other in the characteristics examined
- The objects belonging in the same group differ significantly from those of the rest groups.

The ultimate scope of clusters analysis is to draw useful conclusions about the timeseries while the efficiency of the methods used is related to the type and size of the timeseries but also to the use of similarity criteria (ex-ante detailed definition of research scope i.e. describing objects or describing relations?).

In the current paper, hierarchical clustering is applied given its principle advantage of not requiring a priori knowledge of the number of clusters and its visualization strength (dendograms generation) [23]. The agglomerative bottom-up clustering approach is exploited – starting with each observation as a cluster, the two closest clusters are joined into one while this repetitive process ends when there is only one cluster with all the data [24]. For a more detailed description on the hierarchical clustering, the interested reader is redirected to [25–27].

3.3 The Input Features for Clustering Trials

Following the distinction among 'shape level' and 'structure level' clustering of Wang et al. (2006), the current paper presents a clustering of urban road paths (timeseries) based on point values as well as on their structural characteristics. Structural characteristics in the current analysis include distribution characteristics and underlying relations: max, min, mean, skewness, kurtosis, autocorrelation (linear measure) and mutual information (non linear measure) [28].

4 Results

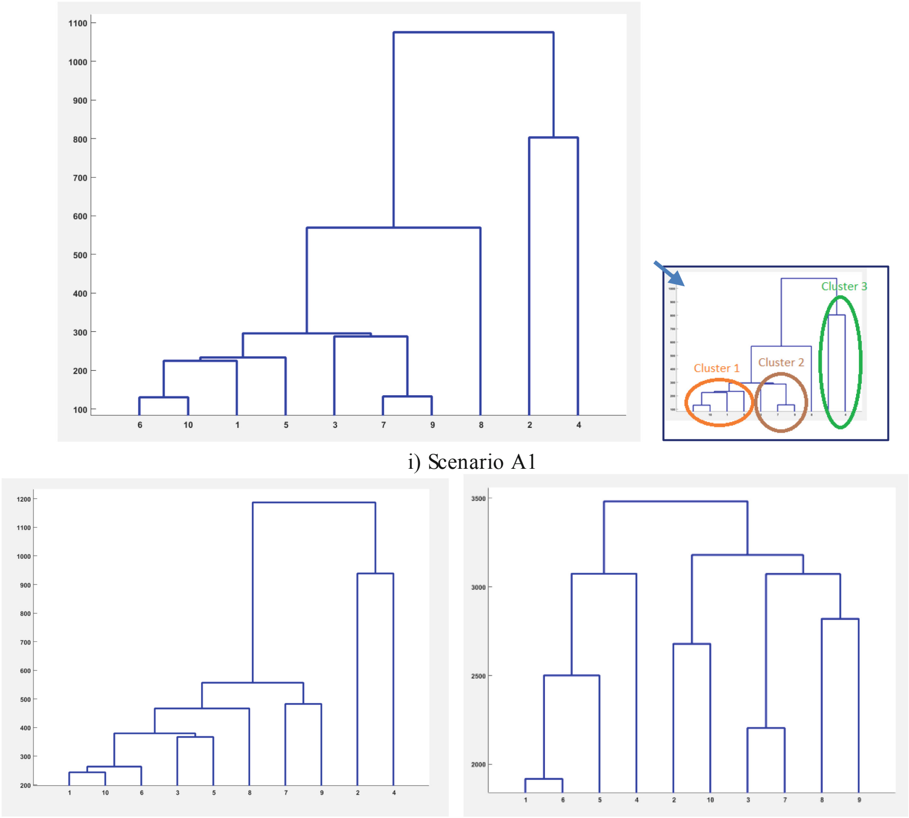
4.1 Long Length Timeseries vs Timeseries of a Typical and an Atypical Day

Aiming to identify similarities at the performance (travel time variations within the day) of road paths under typical (averaged travel time data of a long period) network operation and under non recurrent events and to explore whether abstract similar patterns among road paths exist, hierarchical clustering applied at different scenarios:

i. Scen.A1: Raw travel time timeseries (x-axis: 15 min intervals) of a large period, namely of one semester (weekdays) – 150 days, 96 values per day.
ii. Scen.A2: Raw travel time timeseries of a typical day (within the same semester, a day with typical and expected traffic conditions) – one day, 96 values.
iii. Scen.A3: Raw travel time timeseries of an atypical day (within the same semester, a day with extreme weather conditions) – one day, 96 values.

MATLAB software was used for applying clustering, dendrograms generation and cophenetic correlation coefficient calculation. Among the similarities/distances tested (Euclidean distance, squared Euclidean distance, single and centroid linkage), Euclidean distance (y-axis) presented the highest correlation coefficient for all three scenarios (varying between 0.6 and 0.7). Figure 2 shows the results of the hierarchical clustering procedure per scenario.

Running again the clustering methodology with average timeseries on weekdays of A2017 the clusters are the same as in A1 with slightly differentiated distances.



i) Scenario A1

ii) Scenario A2

iii) Scenario A3

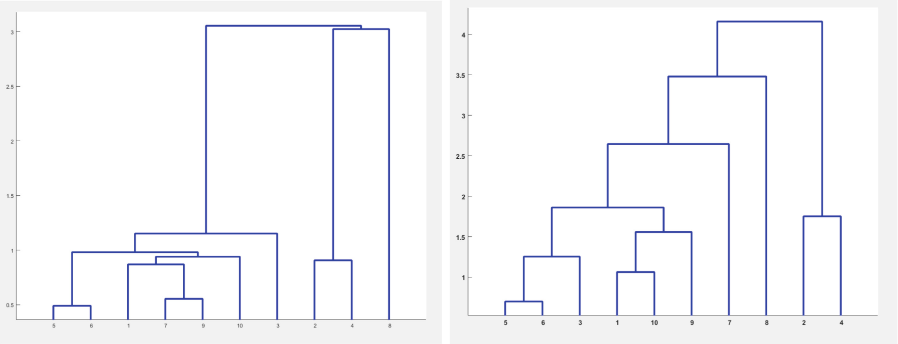
Fig. 2. Hierarchical clusters dendrograms per scenario Ax.

The dendrograms of scenarios 1 and 2 presents high similarity (branches) fact that can be used as an indication that objects (road paths) similarity at a typical weekday is high enough to high dimensionality clustering such as travel time timeseries of a large period (approach incorporating the notion of averaging). The prior knowledge for the reference paths under expected traffic state verifies the 3 obvious clusters created; Cluster 1 consists of paths 190 and 194 that are paths of same length, consisting part of a greater axis crossing the city center (direction to the east) and from path 7 that has similar length and operational characteristics (1 bus lane + 2 lanes, average daily flows of similar levels) and serves the opposite direction, Cluster 2 consists also of the opposite direction paths to 190 and 194, namely of 191 and 193 (direction to the West) and 107 with direction to East, Cluster 3 consists of the last legs (with the opposite direction) of the two large horizontal axis that are examined. Path 192 seems to differentiate from the other paths, which is indeed expected (it serves also turnover trips around the city center) – correlation analysis undertaken as an extension to the above analysis, verifies also the above understanding and interpretation.

From the other side, when clustering road paths using travel times based on raw data of an atypical day (extreme weather conditions with severe congestion in the whole city) the clusters derived are, as expected, totally different from these of the first two scenarios. Distances are higher while also similarities change significantly; the developed Clusters are difficult to be interpreted given also the breakdown that was noticed the specific date in the city (snowcovered roads, widespread traffic jam).

4.2 Raw Data Clustering vs Clustering Based on Linear and a Combination of Non-linear Timeseries Characteristics

Serving the second goal of the current work, namely to compare the clustering results based on raw data and on features in time domain extracted from the raw datasets, hierarchical clustering was conducted based on distribution related features – max, min, mean, skewness, kurtosis – (linear measure of) autocorrelation and (non linear measure of) mutual information of the timeseries. Figure 3 shows the results of the structural characteristics based clustering (following a standardization procedure due to the existence of big differentiations in the values - subtracting of the mean of the attribute values and dividing by the standard deviation).



i) Scenario B1 – without mutual information

ii) Scenario B2 – including mutual information

Fig. 3. Hierarchical clustering dendrograms based on structural characteristics.

The resulting clusters in the two scenarios B1 and B2 are differentiated in a some degree; distances among road paths and positioning among clusters are not exactly the same which shows a relevant impact of taking into consideration also non linear features while conducting clustering of objects belonging to a dynamic system. Although the clusters do not differ to such a large degree, the morphology of the B2 dendrogram seems more clear – it is however an indication which should be further investigated however given the fact that traffic is a dynamic phenomenon it is well noted that including non linear features can enhance results interpretability.

5 Conclusions and Discussion

Many studies have dealt with clustering traffic data variables, however in the framework of the current work we focused on exploring the efficiency of dimensionality reduction of large datasets rather than selecting the most efficient clustering algorithm. In line with previous research results [26–29], the current paper showed that using dimensionality reduction methods can provide high accuracy clustering results with a very low computational cost; road paths clustering based on long length travel time timeseries is similarly enough with the clustering based on average travel time timeseries (averaging the 110 weekdays of A2017). Furthermore, the results of the clustering of average travel time timeseries of A2017 do not differ much from the clustering of road paths in a typical weekday. From the other side and as expected, the road paths clustering is totally different under unusual circumstances (a day with extreme weather conditions). Correlation and causality analysis [22] shows also the different underlying relations between timeseries under different traffic states, therefore a two steps analysis joining the results of correlation-causality and clustering is estimated to add in the interpretability of outputs –clusters can be only a part of the total necessary input for a decision support system. From the other side and when concept-based clustering, i.e. relationships interpretation is the scope of the research, clustering can be based on characteristics as distribution features (max, min & mean values, skewness, kurtosis) and correlation information (autocorrelation and mutual information). This advantage can be exploited not only when missing data in timeseries exist and when the length of the timeseries is not the same but also when the underlying relations and not just the shape of the timeseries is the scope of the clustering. And since we are talking for dynamic systems, as urban road traffic, the authors strongly supports that when combining non linear measures the results of the clustering present higher accuracy.

References

1. Barthelemy, M.: Spatial networks. Phys. Rep. **499**, 1–101 (2011)
2. Strano, E., Nicosia, V., Latora, V., Porta, S., Barthelemy, M.: Elementary processes governing the evolution of road networks. Sci. Rep. **2**, 296 (2012)
3. Van Ommeren, J., Rietveld, P., Nijkamp, P.: Job mobility, residential mobility and commuting: a theoretical analysis using search theory. Ann. Reg. Sci. **34**(2), 213–232 (2000)
4. Polyzos, S.: Urban Development, p. 541. Kritiki Publications, Athens (2015)
5. Polyzos, S., Tsiotas, D., Minetos, D.: Determining the driving factors of commuting: an empirical analysis from greece. J. Eng. Sci. Technol. Rev. **6**(3), 46–55 (2013)
6. Chowell, G., Hyman, J.M., Eubank, S., Castillo-Chavez, C.: Scaling laws for the movement of people between locations in a large city. Phys. Rev. **68**, 066102 (2003)
7. Kwon, J., Mauch, M., Varaiya, P.: Components of congestion: delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. Transp. Res. Rec. J. Transp. Res. Board **1959**(1), 84–91 (2006)
8. Wen, T.H., Chin, W.C.B., Lai, P.C.: Understanding the topological characteristics and flow complexity of urban traffic congestion. Phys. A Stat. Mech. Appl. **473**, 166–177 (2017)
9. Weijermars, W.A.: Analysis of urban traffic patterns using clustering (2007)

10. Roess, R.P., McShane, W.R., Prassas, E.S.: Traffic Engineering, 2nd edn. Pretence Hall, USA (1998)
11. Rehborn, H., Klenov, S.L., Palmer, J.: An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. Phys. A Stat. Mech. Appl. **390**(23–24), 4466–4485 (2011)
12. Wright, C., Roberg, P.: The conceptual structure of traffic jams. Transp. Policy **5**(1), 23–35 (1998)
13. Leduc, G.: Road traffic data: collection methods and applications. Working Pap. Energ. Transp. Clim. Change **1**(55), 1–55 (2008)
14. Efthymiou, D., Antoniou, C.: Use of social media for transport data collection. Procedia Soc. Behav. Sci. **48**, 775–785 (2012)
15. Salanova Grau, J.M., Toumpalidis, I., Chaniotakis, E., Karanikolas, N., Aifadopoulou, G.: Correlation between digital and physical world, case study in Thessaloniki. J. Location Based Serv. **11**(2), 118–132 (2018)
16. Salanova Grau, J.M., Mitsakis, E., Tzenos, P., Stamos, I., Selmi, L., Aifadopoulou, G.: Multisource data framework for road traffic state estimation. J. Adv. Transp. (2018)
17. Aifadopoulou, G., Salanova Grau, J.M., Tzenos, P., Stamos, I., Mitsakis, E.: Big and open data supporting sustainable mobility in smart cities – the case of Thessaloniki. In: Proceedings of 4th Conference on Sustainable Urban Mobility (CSUM2018), pp. 24–25. May, Skiathos Island, Greece (2019)
18. Mitsakis, E., Chrysohoou, E., Salanova Grau, J.M., Iordanopoulos, P., Aifadopoulou, G.: The sensor location problem: Methodological approach and application. Transport **32**(2), 113–119 (2017)
19. Mitsakis, E., Salanova Grau, J.M., Chrysohoou, E., Aifadopoulou, G.: A robust method for real time estimation of travel times for dense urban road networks using point-to-point detectors. Transport **30**(3), 1648–4142 (2015)
20. Stamos, I., Salanova Grau, J.M., Mitsakis, E., Aifadopoulou, G.: Modeling effects of precipitation on vehicle speed: floating car data approach. Transp. Res. Rec. J. Transp. Res. Board **2551**(1), 100–110 (2015)
21. Salanova Grau, J.M., Maciejewski, M., Bischoff, J., Estrada, M., Tzenos, P., Stamos, I.: Use of probe data generated by taxis. Big Data for Regional Science. Routledge Advances in Regional Economics, Science and Policy. Taylor & Francis Group, Abingdon (2017)
22. Myrovali, G., Karakasidis, T., Charakopoulos, A., Tzenos, P., Morfoulaki, M., Aifadopoulou, G.: Exploiting the knowledge of dynamics, correlations and causalities in the performance of different road paths for enhancing urban transport management. In: Freitas, P., Dargam, F., Moreno, J. (eds) Decision Support Systems IX: Main Developments and Future Trends. EmC-ICDSST 2019. Lecture Notes in Business Information Processing, vol 348. Springer, Cham (2019)
23. Keogh, E., Lin, J., Truppel, W.: Clustering of time series subsequences is meaningless: implications for past and future research. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, USA, pp. 115–122 (2003)
24. An introduction to data science, Dr. Saed Sayad. https://www.saedsayad.com/clustering_hierarchical.htm. Accessed 06 2020
25. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. KDD Workshop Text Min. **400**(1), 525–526 (2000)
26. Kraskov, A., Stögbauer, H., Andrzejak, R.G., Grassberger, P.: Hierarchical clustering using mutual information. EPL Europhys. Lett. **70**(2), 278 (2005)
27. Kojadinovic, I.: Agglomerative hierarchical clustering of continuous variables based on mutual information. Comput. Stat. Data Anal. **46**(2), 269–294 (2004)

28. Wang, X., Smith-Miles, K., Hyndman, R.: Characteristic-based clustering for time series data. Data Min. Knowl. Discov. **13**, 335–364 (2006)
29. Steinbach, M., Ertoz, L., Kumar, V.: Challenges of clustering high dimensional data. In L.T. Wille, editor, New Vistas in Statistical Physics - Applications in Econophysics, Bioinformatics, and Pattern Recognition. Springer-Verlag (2003)