Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Complex network enabled time series analysis of nightlight data

Taha Hossein Rashidi[1], Seyedehsan Seyedabrishami[2]

[1]rCITI, UNSW

[2]University of Sydney, and Tarbiat Modares University

Email for correspondence (presenting author): rashidi@unsw.edu.au

1. Introduction

As new data sources continue to expand, data analytics is adapting to accommodate the distinctive characteristics of emerging data types. It is important to note that the size of the data is not the sole factor that piques the interest of researchers when it comes to investigating the attributes of big data. Big data encompasses a range of data types, including multimodal, crowdsourced, imagery, and text-based data. However, these data types also introduce new challenges when it comes to extracting valuable insights for purposes such as analysis, planning, and prediction.

In the field of transport, there has been a growing interest in exploring the potential of emerging data sources like social media. Traditionally, methods such as sentiment analysis, semantic analysis, text mining, and natural language processing were somewhat separate from the techniques employed by transport experts in modelling transportation data, demand analysis, accident analysis, and network design. Nevertheless, these methods have now become part of the toolkit that transport analysts utilise to model travel-related attributes, such as trip generation, and mode of travel.

The US Air Force owns nighttime light data, which is gathered through the Defence Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) sensors. This data is managed by the United States National Oceanic and Atmospheric Administration (NOAA) and has been available digitally since 1992, following the satellite's launch in 1972. Daily images are collected, but they require adjustments for quality due to atmospheric and weather conditions (Li et al., 2020).

In 2012, a new generation of nighttime light data known as VIIPS was introduced, provided by the Suomi National Polar-orbiting Partnership (NPP) Satellite. VIIPS offers higher resolution and fewer over-glow effects compared to its predecessor (Li et al., 2020).

A novel approach to analysing time series data, rooted in complex network theory, involves converting time-series data into network structures while preserving their inherent properties. This transformation unlocks a rich toolbox of tools and techniques within the well-established field of complex networks. These methods have already found utility in modelling time series data across diverse domains like finance and industrial production (Silva et al., 2022). Silva et al.'s work notably demonstrates that advanced complex network features are particularly valuable for characterizing time series data, especially when identifying clusters within spatially distributed data. This integration of complex network techniques into time series analysis holds promise for revealing intricate data patterns and enhancing our understanding of various applications.

This research introduces three innovative dimensions to the field of transport modelling. Firstly, it highlights the untapped potential of satellite imagery data for addressing transport-related

issues, particularly in the context of urban growth analysis from a strategic planning perspective, which has been relatively underexplored. Secondly, the study delves into the temporal aspect of imagery data, emphasizing the transformation of such data into complex networks. While temporal analyses of nightlight data exist in the literature, this research underscores the novel opportunities that emerge when treating such data as time series, opening up new avenues in data mining. Lastly, the research outlines methodologies for the analysis of spatially distributed complex network data and the clustering of this information. The study also addresses considerations regarding spatial resolution and the impact of zoning methods on the analysis across these three countries. These clustering methods offer flexibility in summarizing the spatial and temporal variations within multidimensional time series data, which are otherwise challenging to analyse comprehensively.

2. Approach

Time series data, which records information collected sequentially over time for various entities, is invaluable in fields like health monitoring, traffic management, finance, and weather forecasting. Analyzing such data requires accounting for correlations among adjacent observations. A new approach called feature-based analysis, inspired by data mining, involves classifying, clustering, or recognising patterns in data. Features reflect specific data properties, including conventional time series attributes like seasonality and autocorrelation. When combined with network concepts, a broader set of features can be generated.

Mapping time series data into a network is a crucial step. This network-based approach, using methods like proximity, visibility, and transition, has consistently shown promise for data classification and clustering. The paper adopts mapping methods from Silva et al. (2022), with networks defined as ordered pairs of nodes and edges, which can be weighted and directed. This overview outlines the sequence of methods applied to time series data: mapping to a network, generating network-based features, and extracting valuable information to describe the data. Detailed mapping methods are available in Silva et al. (2022).

The literature encompasses node-based, path-based, and community-based attributes, providing insights into network structure. Transport modelling has already explored such properties. This study, following Silva et al. (2022), considers 15 network properties available in the NetF package. These properties include centrality measures (average weighted degree, in-degree, out-degree), path-based measures (average path length), and community-based measures (global clustering coefficient, number of communities, modularity). These features are calculated across three graphs (WHVG, WNVG, QG) as discussed earlier.

Key considerations for these measures based on the NetF package are as follows: average weighted degree (k_bar) is calculated as the average of node degrees, while average path length (d_bar) considers shortest path lengths between nodes, regardless of edge weights. The clustering coefficient (C) is designed for undirected networks. The number of communities (S) uses methods to identify densely connected subgraphs through short random walks. Modularity (Q) assesses the separability of nodes belonging to different communities within the network.

This study addresses the challenge of dealing with large-scale multivariate time series data. Unlike traditional univariate time series analysis, where single observations predict the future, this study focuses on multivariate time series data. It emphasises the need to reduce dimensionality when applying a feature-based approach to extract meaningful information from the data. The methodology, inspired by Silva et al. (2022), involves three key steps. First, it normalises the features to comparable ranges using the min-max technique. Second, it employs Principal Component Analysis to combine features into components that capture the most variation in the data. Finally, it clusters the time series into groups with similar features and

characteristics, using the k-means approach. The clustering results are evaluated using metrics like Average Silhouette, Adjusted Rand Index, and Normalised Mutual Information.

In summary, this study transforms spatially correlated time series data into networks, extracts and groups features, and performs clustering. This process yields insights into the data's characteristics, which can be valuable for forecasting and analysis purposes. The study's methodology is illustrated in a flowchart for clarity.

3. Data

This article introduces novel data sources that offer valuable insights for urban planners and transport modellers regarding the spatial and temporal distribution of urban activities. It also presents specialised data analysis tools to handle the complexity of this data. The nightlight data, accessible through platforms like R, provides imagery of Earth's activities but requires thorough preprocessing to remove noise from various sources like weather, atmosphere, over-glow, vessel lights, and pollution. Data from DMSP (processed by NOAA) and VIIRS is collected by the US Air Force Weather Agency. DMSP data is available yearly as global images, while VIIRS offers monthly data. Integrating these two sources has been attempted in various regions worldwide, with efforts to generate DMSP-like data from VIIRS data. Li et al. (2021) presented a harmonised approach for the entire world, covering data from 1992 to 2018.

DMSP data originally served as a cloud cover observation tool but has been adapted for stable nightlight data analysis. This stable data is annually-averaged, with outliers controlled. VIIRS, designed to standardise low-light imagery data, overcomes DMSP-OLS limitations and broadens the applications of nightlight data. After data processing, it must be aggregated into zones, which can significantly impact results. Various shapefiles linked to nightlight data and different aggregation levels are available online. Processing time varies with zone size. The article later illustrates how zoning assumptions can affect clustering analysis using Australian nightlight data.

4. Results

This section presents the results of the NetF model applied to Nighttime Light (NTL) data spanning from 1993 to 2018. Additionally, data from the transport network during the same period, with matching zoning at precise geographical resolutions, is modelled. The objective is to illustrate how complex network methods combined with time series modelling techniques can extract valuable insights from spatially distributed time series data, which is traditionally challenging to analyse. The analysis begins by applying the NetF model to NTL data across 31 provinces in Iran. To assess the NetF approach's effectiveness, two other feature-based methods, tsfeatures and catch22, as recommended by Silva et al. (2022), are benchmarked. The data comprises 31 time series grouped based on predefined classes, determined by factors such as province size, population, or geographical location.

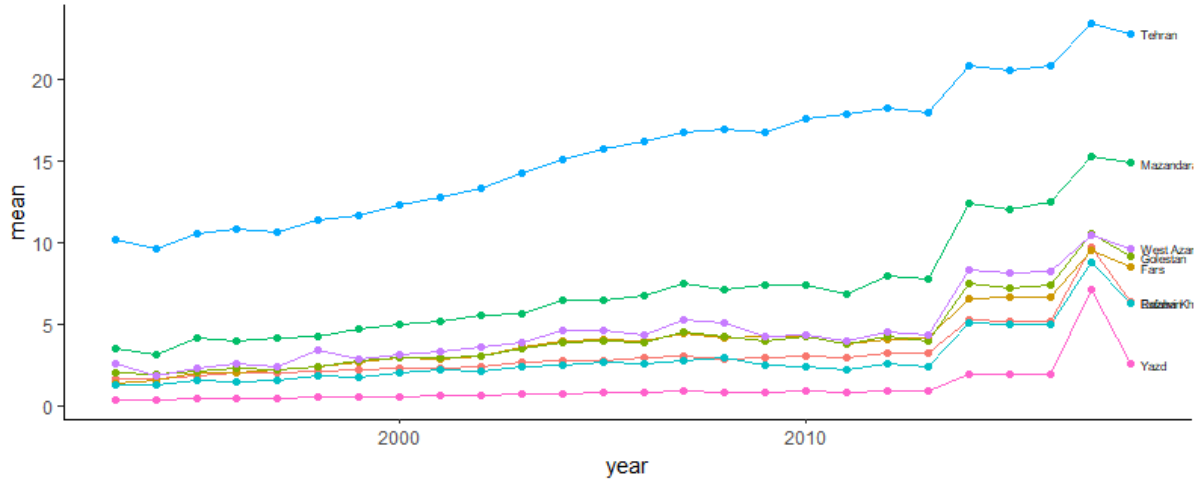
Each province's NTL time series is transformed into a complex network structure using WNVG, WHVG, and QG methods. Five features are generated for each network, resulting in 15 features for each time series. The principal component analysis combines these 15 features into several components. This summarised information is then utilised for clustering the time series data into a predefined number of clusters via a k-means algorithm.

Since there is no ground truth available for the clustering exercise, the average Silhouette is used to assess the clustering method's performance. This analysis demonstrates how advanced techniques can uncover meaningful patterns and structures in complex, spatially distributed time series data.

In this section, the study demonstrates the practical application of the discussed methods and algorithms through an analysis of nightlight data for Iran. Additionally, various time series data related to transportation are analysed using the same techniques to compare the findings and identify potential correlations between these two distinct data sources. This comparative analysis serves as a significant contribution, showcasing how clustering methods can offer insights into the relationship between spatially and temporally distributed time series data.

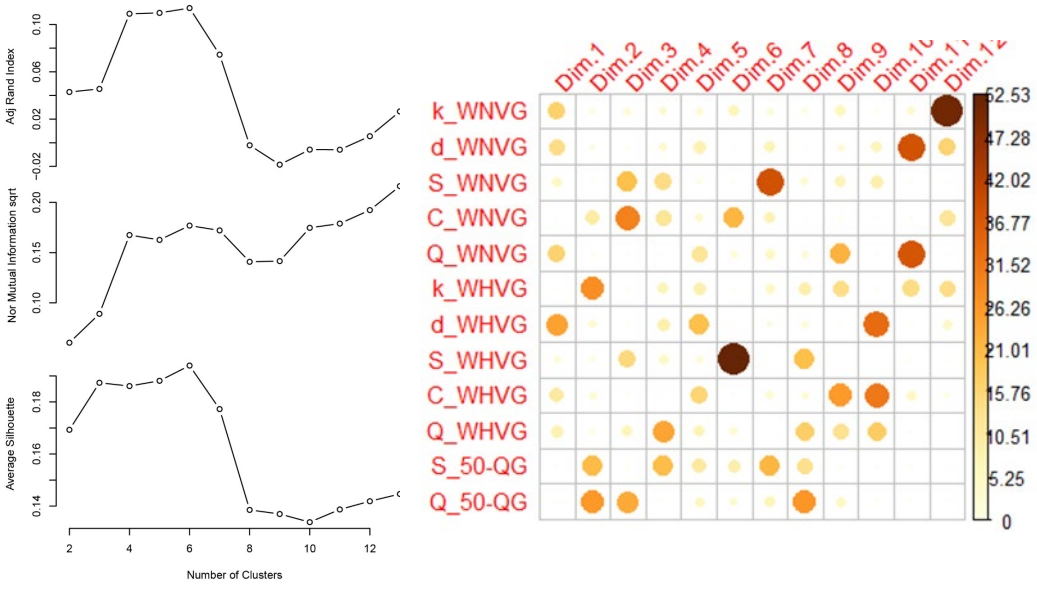
The initial analysis involves gaining a comprehensive understanding of the raw nightlight data, aggregated by different provinces within the region. Figure 1 displays the time series input data for the major provinces in Iran. Upon initial observation, a discernible trend emerges – a gradual increase in the strength of nightlight data across all provinces, with Tehran exhibiting the highest intensity, as anticipated. This analysis sets the stage for further exploration and insights into the patterns and dynamics of these data sources.

Figure 1: Raw time series nightlight data for large provinces in Iran



The subsequent phase involves performing a mapping analysis to convert the time series data into a graph, facilitating the extraction of complex-network metrics. The nightlight time series data undergoes analysis utilising the previously introduced NetF method. The determination of the number of clusters is guided by metrics such as the Average Silhouette, normalised mutual information square root, and adjusted random index.

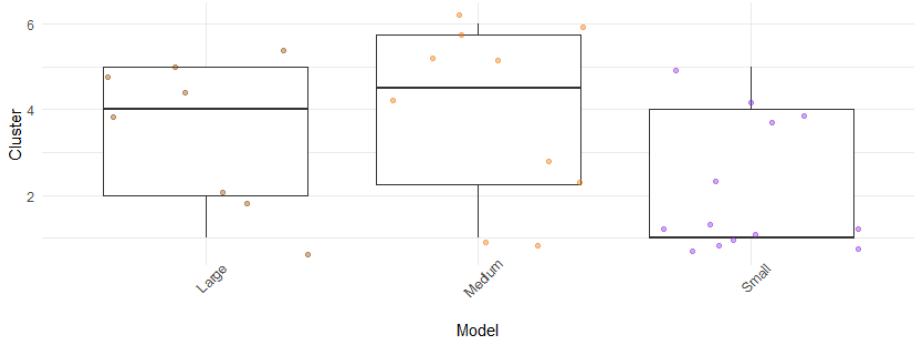
Figure 2: The PCA and the k-means cluster analysis of the nightlight data



Based on the average Silhouette metric, we opt for 6 clusters as the preferred number for further analysis, considering the absence of ground truth information. These clusters are formed through a principal component analysis (PCA) conducted on the 15 indicators, as depicted in Figure 2. Notably, the dominant explanatory factor in the last PCA dimension is the weighted average of nodes in the natural visibility graph. Within the NV Graph, the arithmetic shortest path and the Q of NVS indicators reflect distinctions among network communities. Among HVS indicators, the weighted number of communities (S) significantly impacts the sixth component, while the two QG indicators exert a lesser influence.

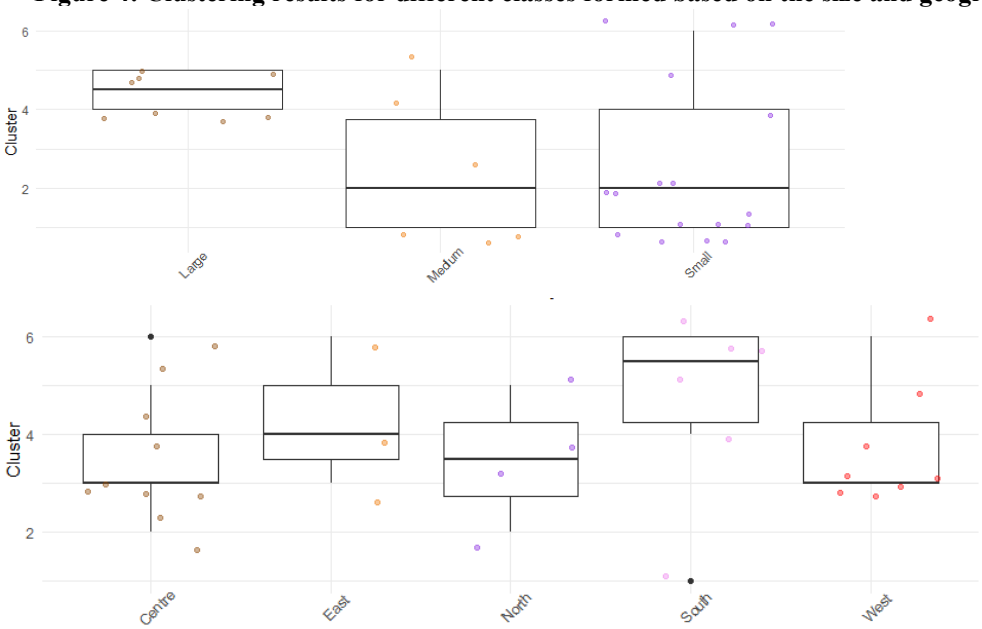
After estimating the network indicators, they are grouped based on the variance they capture in the data, as revealed by PCA analysis (Figure 2). The right diagram in the figure illustrates the indicators and their contributions to the PCA dimensions, focusing on those deemed critical. These dimensions (components) are then employed in the cluster analysis, taking into consideration externally determined classes. These classes can be based on province size or geographical location, and we explore how these labels can affect the clustering approach. Utilising them, clusters are formed based on the province population, as illustrated in Figure 3.

Figure 3: Clustering results for different classes formed based on the population of provinces



As depicted in this figure, where we have chosen to maintain a fixed number of six clusters, guided by the average Silhouette index, it becomes evident that provinces categorized into three distinct classes exhibit considerable variability in their assignment to different clusters. To delve deeper into the influence of province grouping, we conducted a parallel analysis by classifying provinces according to their size and geographical locations. The resulting clustering outcomes are presented in the subsequent diagrams.

Figure 4: Clustering results for different classes formed based on the size and geographical locations



As seen in the above figure, different classification settings can change the clustering outcomes, although six clusters are always determined to be the optimum number of clusters. It is worth noting that different classifications do not affect the estimation of NetF matrices. It is only the grouping of the class members into clusters that are affected by the various class settings. As demonstrated in the previous paragraphs, the clustering analysis can be affected by how classes of multidimensional time series are defined. In the case of this section, if the provinces are classified based on their size, population location or even jurisdiction zoning systems, different clustering arrangements are likely obtained.

4. Conclusion

This paper introduces the concept of utilising nighttime light data as a valuable resource for transport planners and modelers, highlighting its untapped potential in infrastructure modeling and planning. Given the unique characteristics of this data, including its size, spatial distribution, temporal nature, and imagery format, advanced analysis tools and algorithms are essential to harness its insights effectively. The study proposes advanced techniques that merge time series modeling with feature-based graph theory methods.

The featured-based approach, explored in this paper, involves mapping time series data into a different space where its transformed features can be analysed using unconventional methods. Complex networks play a central role in this analysis, offering a wide range of techniques for studying transformed time series data while preserving its core properties. Three popular methods, namely natural visibility graph, horizontal visibility graph, and quantile graph, are discussed, each retaining specific information from the original time series data. The NetF approach, introduced in this paper, combines the features of these three networks to efficiently extract valuable insights.

The study demonstrates the versatility and effectiveness of the NetF approach by applying it to nighttime light data from Iran. The data require advanced image processing methods to extract useful information while mitigating background noise sources like clouds and light glare. The NetF approach identifies communities within the time series data, revealing correlations between different provinces over time. The analysis is further cross-checked with transport attributes, such as the total number of trips and road lengths in Iran, providing insights into the correlation between nighttime light data, transport information, and demographic data. By employing complex network methods to analyse spatially correlated imagery data, the study introduces a novel approach for extracting meaningful insights from this data source within the context of transportation analysis.

Future avenues for this research encompass integrating the road network data with nighttime light data to gain a deeper understanding of their potential synergies. Exploring the capabilities of nighttime data in planning scenarios can be extended by incorporating additional transport infrastructure and transport demand data sources. A significant future focus lies in developing practical methodologies to leverage clustering results for forecasting purposes. This involves quantifying the spatial correlations identified during the clustering process.

References

Li, X., Zhou, Y., Zhao, M., & Zhao, X. (2020) A harmonised global nighttime light dataset 1992–2018. *Scientific data*, 7(1), 1-9

Silva, V. F., Silva, M. E., Ribeiro, P., & Silva, F. (2022) Novel features for time series analysis: a complex networks approach. *Data Mining and Knowledge Discovery*, 36(3), 1062-1101.