Node importance corresponds to passenger demand in public transport networks

Šfiligoj, Tina; Peperko, Aljoša; Bajec, Patricija; Cats, Oded

DOI
10.1016/j.physa.2025.130354

Publication date
2025

Document Version
Final published version

Published in
Physica A: Statistical Mechanics and its Applications

Physica A 659 (2025) 130354

Contents lists available at ScienceDirect

Physica A

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

Node importance corresponds to passenger demand in public transport networks

Tina Šfiligoj [a,*], Aljoša Peperko [b,c], Patricija Bajec [a], Oded Cats [d]

[a] Faculty of Maritime Studies and Transport, University of Ljubljana, Pot pomorščakov 4, Portorož, 6320, Slovenia
[b] Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva 6, Ljubljana, 1000, Slovenia
[c] Institute of Mathematics, Physics and Mechanics, Jadranska ulica 19, Ljubljana, 1000, Slovenia
[d] Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, Delft, 2600 GA, The Netherlands

ARTICLE INFO

Keywords:
Public transport
Graph theory
Network science
Node centrality
General transit feed specification (GTFS)
Smart card data

ABSTRACT

We investigate the correspondence between network-based public transport network (PTN) supply indicators and passenger demand at the node level, by systematically assessing correlations between node centrality measures and passenger boarding counts across different graph representations of PTNs. At the stop-level, undirected L- and P-space representations with three different edge weightings: unweighted, service-frequency-weighted, and in-vehicle-time-weighted are analysed. In each case, we calculate degree, closeness, betweenness and eigenvector centralities and examine the relation shapes. At the route level, we examine degree and eigenvector centrality for unweighted and weighted C-space representations. We introduce a modified C-space representation with self-loops, with service frequencies as self-loop weights, and propose eigenvector centrality as a route-level supply indicator. Stop- and route-level properties are integrated using the B-space representation. This methodology was applied to a case study for a bus PTN in Ljubljana, Slovenia. Results show strong correspondence between passenger demand and degree and eigenvector centrality scores in the frequency-weighted P-space (correlation $\approx 0.7 - 0.8$). Notably, the relationship between eigenvector centrality and passenger counts in the new C-space representation with self-loops exhibits logarithmic behaviour. Furthermore, the results suggest a minimum eigenvector centrality threshold ($\approx 10^{-3}$) for a route to start facilitating passenger use. The route-level results from the B-space analysis show exponential convergence of passenger counts to route eigenvector centrality. Results of the stop-level analysis are in line with previous research and deepen the understanding of centrality measures as supply indicators. Most significantly, the route-level analysis is novel, and the results open promising venues for further research.

1. Introduction

Graph theory and network science methods have become increasingly important in public transport network (PTN) research. Public transport (PT) systems can be represented as complex networks and their structure and dynamics can thus be conceptualised and studied using related notions and techniques. A seminal work in connecting network science and (public) transport research is the one by von Ferber et al. [1] where the four - by now - standard graph representations of PTNs - L, P, C and B-space - were introduced and a detailed analysis of centrality distributions, checking for small-world and scale-free phenomena, was performed.

* Corresponding author.
E-mail address: tina.sfiligoj@fpp.uni-lj.si (T. Šfiligoj).

https://doi.org/10.1016/j.physa.2025.130354
Received 13 June 2024; Received in revised form 28 November 2024
Available online 6 January 2025

In [2], the methods introduced in the previous paper, were used on a dataset of 14 real-world urban PTNs. Notable examples of stop-level network-science-based PTN research analysed topological properties, principally the degree distributions and looking for small-world or scale-free phenomena (e.g. [3–11]). Other studies included route-level and integrated stop and route representations (e.g. [7,12–14]), where power-law or exponential shapes of degree distributions were observed. Notably, similar behaviour of PTN topological characteristics was observed over numerous case studies of real-world PTNs in the cited studies, located in different countries and continents, and with different network sizes and modalities.

While network science-based research into PT systems provides new insights into their properties, a gap between using PTNs for applied network science on the one hand and using network science for transport research on the other, began to emerge. An early call to bridge this gap was made by Derrible and Kennedy in [15], who offered a comprehensive review of the literature and proposed a unified approach of PTN analysis based on a set of topological properties.

A shifting to transport research focus is evident in numerous (recent) studies where network science was used to study crucial aspects of PT systems, such as robustness and stability (e.g. [16–19]) and accessibility (e.g. [20,21]). Both robustness and accessibility are among the main determinants of PT use and deep understanding of their properties is essential for transport researchers and practitioners.

The emergence of a unified format for timetable data, General Transit Feed Specification (GTFS), that is mostly publicly available, makes research into network properties ever more accessible and reproducible. Based on network properties, obtainable from GTFS data, unified supply and quality-of-service indicators can be readily constructed and compared over a number of different networks. To strengthen the point, robustness and accessibility research has become easier with the availability of GTFS data, as evident from the above references. Both robustness and accessibility are intrinsic properties of the system, and the same holds for any supply indicators. Notwithstanding, the ultimate goal of designing and planning PT services is their usage. Due to the interplay between supply and demand, a degree of correlation is expected. This gives rise to an important question: to what extent do network-based supply indicators correspond to passenger demand at node level?

This question was posed and examined in a pioneering study [22] where it was shown that structural properties, as reflected in node centralities, indeed correlate positively with passenger flows as measured at node level, based on empirical smart card data from Dutch tram networks. Directed unweighted and weighted L-space and P-space representations of the tram networks were built, and degree, closeness and betweenness centrality scores were computed for nodes, and correlations with passenger flows were calculated. [23] presented a correlation analysis of node centralities (degree, closeness, betweenness) in a multi-modal bus and metro PT network with passenger flows for a PTN in Beijing, China, for L- and P- and C-space representations. [24] constructed a layered network with three layers representing the P-space, OD network and transfer network. The study developed metrics to calculate degree and strength centrality scores of nodes in the network, thus connecting network topology and passenger flows. [25] explored the relation between centrality measures and passenger boarding and alighting counts in the L-space representation of the Athens metro system and significant Pearson correlation coefficients were observed. [26] employed the C-space representation to identify clusters of routes based on passenger transfers. [27] studied the influence of network topology, together with urban density and socio-economic effects, on PT ridership on a global network level for 48 PTNs worldwide.

In this study, we further advance the idea of using node centrality measures as supply indicators. Centrality scores of nodes represent different estimates of node importance in the network. Primarily, they reflect the connectedness of a node to other nodes in the network at different levels. If weighted networks are used, with edge weights representing service frequency or in-vehicle time, this provides a natural integration of node connectedness (complex networks perspective) and quality-of-service parameters (transport research perspective). While network and service properties are modelled at edge-level, the topological properties are aggregated at node-level in the form of centrality measures. Aggregated numbers of boardings (or alightings) represents a qualitatively comparable measure on the demand side.

We have used the term “node-level” when discussing PTN properties. In the most common representations, nodes represent stops and edges represent infrastructural or operational connections between them. However, in other representations, nodes can correspond to routes, such as in C-space, or in the B-space the PTN is represented as a bipartite network with two disjunct sets of nodes representing stops and routes, respectively. We argue that these representations are underexplored and propose to use route-level properties in examining the relations between supply and demand in PTNs. Route-level representations are on a higher level of abstraction than stops, as the spatial properties are aggregated, based on service properties. This offers a less detailed spatial picture of passenger demand patterns than stop-level analysis. Route-level analysis offers arguably a complementary view. Where passenger flows are not known, aggregating data at stop level, such as total number of boardings, provides a 0-dimensional image of the network by describing only properties of static points in space. In the C-space, however, due to higher-level abstraction, node-level properties already represent (legs of) journeys and thereby offer a view in a 1-dimensional space with demand at nodes representing paths in the physical/operational network. Even in entry-only systems, route-level journeys can be captured fully from the data. Thus, there is a trade-off between the granularity of network description and completeness of passenger demand information. This trade-off can be balanced to some extent by analysing the combined-levels description that is captured by the B-space representation and is discussed in more detail in Section 2.2.3.

While the above literature review shows considerable progress in research into relations between network structure and passenger demand, we identify several research gaps. First, we see an opportunity to improve upon existing stop-level network representations (L- and P-space) in terms of appropriate choice of centrality measures which can yield more insightful interpretations. Second, we argue that there is uncovered potential in examining route-level properties by exploring the C-space representation. Furthermore, the interrelation of stop- and route-level properties would offer an integrated perspective from several PT system levels, which can be studied by examining the B-space representation. Importantly, in the latter two representations, the appropriate edge weights

need to be determined. Finally, we observe that the shapes of relations between structural properties and passenger flows have not been examined in detail and investigate which correlation measures are the most appropriate to study the strength of interaction. In an attempt to fill these gaps, we pose the following research questions:

- **RQ1** What structural properties of stop-level graph representations of PTNs best correspond to passenger demand at the node level?
- **RQ2** What representations and which of their structural properties best correspond to route-level passenger demand?
- **RQ3** How can stop- and route-level structural properties be integrated to offer a combined view of passenger demand?

The main contributions of this study are:

- Extending the analysis of P- and L-space representations, with a focus on:
  - interpretations of centrality measures,
  - analysis of the impact of different edge weights, and
  - choice of suitable correlation measures.
- An extensive analysis of route-level structural properties, by introducing a modified doubly weighted C-space representation with self-loops, and the relations to passenger demand.
- An examination of the relationship between route and node properties by studying the relations in the B-space representation.

The remainder of the paper is structured as follows. In Section 2, methodology is presented, divided into subsections on graph theory and node centrality preliminaries 2.1, graph representations of PTNs 2.2, and statistical methods 2.3. Results of the application to a case study are presented in Section 3. Section 4 concludes with a discussion and offers potential venues for future research.

2. Methodology

2.1. Preliminaries

A graph G is defined by the set of vertices $V = \{v_1, \dots, v_n\}$ and a set of edges E, where $e_{ij} = (v_i, v_j) \in E$ is an edge between vertices v_i and v_j. A graph is undirected when $(v_i, v_j) = (v_j, v_i)$ for all $e_{ij} \in E$. Number of vertices $|V| = n$ is called the dimension or order of G and number of edges $|E| = m$ is its size. The graph is represented by its adjacency matrix A of dimension $n \times n$ with elements $a_{ij} = [A]_{ij} = 1$ if an edge exists between vertices v_i and v_j, and $a_{ij} = 0$ otherwise. This can be generalised to graphs in which edges are assigned weights w_{ij}, where the corresponding entries of the weighted adjacency matrix A become $a_{ij} = w_{ij}$. In a bipartite graph B, the vertex set is a union of two disjunct subsets, or bipartitions, $V = V_1 \cup V_2$ and an edge is allowed only between two nodes from distinct subsets.

While graphs are mathematical objects, networks are graphs with certain properties representing real world systems. Often, when describing networks, the term node is used instead of vertex, and link instead of edge. Nodes represent objects or elements of the system, and edges represent connections or relations between them. In the case of PT networks, nodes most often represent stops in the system, and links the infrastructural or operational connections between them. On a different level, nodes can also represent routes and links the existence of shared stops. The two bipartitions in a bipartite network can represent stops and routes, respectively, and an edge exists between two vertices in distinct subsets if the corresponding stop lies on the corresponding route.

The underlying topology of PTNs is described with unweighted representations. To incorporate additional properties of the PT system in the description, weighted graph representations are often used. Two of the most important link-level properties of PT networks are travel time and frequencies. It is important to note that they represent two fundamentally different types of weights: travel times represent a generalised cost of the link, while frequencies indicate the strength of the connection.

2.1.1. Node centrality measures

There are numerous definitions of node centrality, each representing a different perspective on node importance. The four most commonly used centralities are degree, closeness, betweenness and eigenvector centrality, which will be also used in this analysis. Below, the definitions of the used centralities are given.

- **Degree centrality** (C^D): degree of node i is the number of its neighbours, which is equal to the number of incident edges. Degree centrality is defined as degree, divided by the number of all other nodes in the network. In terms of the adjacency matrix, it is written as:

$$C_i^D = \frac{\sum_{j \neq i} a_{ij}}{n-1}. \tag{1}$$

- **Closeness centrality** (C^C): the inverse sum of distances d_{ij} of shortest paths from node i to all other nodes in the network:

$$C_i^C = \frac{1}{\sum_{j \neq i} d_{ij}}. \tag{2}$$

Table 1
A summary of the centralities used and their properties, relevant to the analysis. The focus is on the scope of each measure inside the network (local vs. global), type of weight needed for the calculation in the weighted version (strength vs. cost), and a short description of the node importance interpretation.

Centrality	Notation	Scope	Weight type	Definition
Degree	C^D	Local	Strength	Number of neighbours, or direct connections
Closeness	C^C	Global	Cost	Inverse of the sum of shortest paths to all other nodes
Betweenness	C^B	Global	Cost	Fraction of the shorterst paths in the network passing through the node
Eigenvector	C^E	Global	Strength	Higher importance of nodes, connected to other more important nodes

- **Betweenness centrality** (C^B): the fraction of the number of shortest paths between all pairs of nodes that pass node i, denoted by $p_{jk}(i)$, in the numerator, and the number of all shortest paths between them (p_{jk}) in the denominator:

$$C_i^B = \frac{\sum_{j,k\neq i} p_{jk}(i)}{\sum_{j\neq k} p_{jk}}. \tag{3}$$

- **Eigenvector centrality** (C^E): we consider coordinates of the eigenvector x, computed as: $x_i = \frac{1}{\lambda}\sum_{j\neq i} a_{ij}x_j$, where λ is the largest eigenvalue of A. Since x is determined up to a multiplicative constant, only relative eigenvector centrality scores are sensible, e.g.:

$$C_i^E = \frac{x_i}{\max_j (x_j)}. \tag{4}$$

While these definitions were originally developed for unweighted networks, the generalisation to weighted networks is straightforward. However, there is one important caveat which we briefly discuss here. There are two fundamentally different types of weights, representing link strength or link cost, respectively. In the description of real-world networks, they stand for quantities that are inversely related to each other: a high edge cost implies weaker connections, while high strength implies relative ease of communication and thus a low cost.

When generalising degree and eigenvector centrality, edge weights of strength type will be used. Closeness and betweenness centrality definitions are based on shortest paths, the calculation of which takes edge cost as input in weighted networks. If one type of weight is given naturally, e.g. connection strength (resp. cost) and the aim is to compare different centrality measures, a transformation on edge weights must be used to convert it to the weight of cost (resp. strength) type. The most commonly used transformation is $f(x) = 1/x$; for obtaining edge cost c_{ij} (resp. edge strength s_{ij}) from edge strength s_{ij} (resp. edge cost c_{ij}), one applies $c_{ij} = 1/s_{ij}$ (resp. $s_{ij} = 1/c_{ij}$). In the analysis of PTNs, two natural edge weights are travel time, of cost-type, and frequency of service, of strength-type.

Different centrality measures are associated with different scales of influence in the network. Degree centrality is a local measure as it is only influenced by the immediate neighbourhood of a node. The other three measures are global and are affected by all other nodes in the network. Closeness and betweenness centrality are based on the shortest paths between all pairs of nodes. While closeness centrality offers an indicator of accessibility by measuring the inverse of the sum of all shortest paths from a node to all other nodes, betweenness centrality offers a measure of importance in terms of flow distribution by counting the shortest paths between pairs of all other nodes that pass a given node. Eigenvector centrality is also a global measure, but in contrast to C^C and C^B, where all nodes are counted equally, in C^E the importance of other nodes is considered. The main points of the discussion of types of weights, scope of influence and interpretation of each centrality measure are summarised in Table 1.

2.2. Graph representations of PTNs

There are four main graph representations of PTNs which we introduce in the following subsections.

2.2.1. Stop-level

In stop-level graph representations, nodes represent PT stops and edges represent infrastructural or operational relations between them. Two standard representations are the so-called L-space and P-space representations.

P-space

P-space, also called space-of-service, explicitly incorporates the operational properties of the PT system by connecting each pair of nodes, representing stops that are reachable from one another without transfer. Thus, there exists a connection between each pair of stops that lie on the same route. Paths in P-space reflect the underlying passenger behaviour, their length counting the number of journey legs. A large majority of journeys consist of one or two legs, and moreover, in P-space, the diameter of the network (the longest of the shortest paths between all pairs of nodes) is usually small. Thus, we do not expect closeness and betweenness centrality measures, based on shortest paths, to significantly correspond to passenger demand. Especially betweenness centrality, measuring the inclusion of a node on shortest paths between other nodes, is not expected to be a meaningful measure in P-space.

On the other hand, from the same arguments, degree and eigenvector centrality appear to more closely reflect passenger behaviour: degree centrality as a local measure should correspond to journeys of length 1, and eigenvector centrality including the importance of end or transit stops, to many-leg journeys. Frequency of service f_{ij} is a natural weight to include in P-space representations. As a weight of strength type, the weighted degree and eigenvector centrality measures in the frequency-augmented representation is expected to correspond even better to passenger demand. For closeness and betweenness centrality calculation, the $t_{ij} = 1/f_{ij}$ transformation should be used. In this case, the resulting cost-type weight t_{ij} has a ready interpretation as directly proportional to the waiting time. Since it is known from literature that waiting time is valued more negatively than in-vehicle time by passengers, also the closeness centrality might show closer correspondence than in the unweighted case.

In the literature, weighted P-space representations use frequencies as weights. In this paper, also travel time is used as a weight. The travel time t_{ij} for the edge between nodes i and j is obtained as the weight of the path between i and j in travel-time-weighted L-space. Although in calculating shortest paths this is a natural weight for determining closeness centrality, due to the passengers' valuation of time, in this representation poorer correspondence to passenger demand is expected than in the frequency-weighted, where the edge weight corresponds to waiting time.

L-space

In L-space, also called space-of-infrastructure, the edges between nodes represent route segments, so that there exists a link between two nodes if they represent consecutive stops on a route. In the weighted representations, the weights can reflect either the number of routes connecting the two stops or, more commonly, in-vehicle travel time t_{ij} between the stops i and j. In the latter case, travel time being the cost-type weight, closeness and betweenness centrality are the natural measures of node importance. If one wanted to determine degree and eigenvector centrality, a transformation of the weight must be used. In this paper, the transformation used was $1/t_{ij}$. However, contrary to P-space, there seems not to be a straightforward interpretation for the so-obtained frequency-like measure. Thus, we would expect low correspondence of these measures to passenger demand. In literature, to the best of our knowledge, only travel-time-weighted L-space representations have been studied. In this paper, also the frequency-weighted representation is used. In this case, degree and eigenvector centrality are readily calculated, while cost-type weights are again obtained by the $1/f$ transformation. However, since this measure is related to waiting time, and in L-space a passenger trip with one leg is a path with length $l \geq 1$, thus summing the waiting times in closeness and betweenness centrality calculations, we do not expect a significant correspondence for these measures. A similar argument holds for degree and eigenvector centrality, since in L-space only local neighbourhoods are described.

2.2.2. Route-level

C-space

On the route-level, the standard graph representation is the so-called C-space representation, where nodes represent routes in the network and an edge between two nodes exists if the corresponding routes share a stop. A weighted C-space version can be constructed by including the number of shared stops on the corresponding routes. Note that this weight is of strength type.

In terms of passenger behaviour, journeys occur along one or several routes. In contrast to stop-level representations, where journeys are described as paths in the network, in C-space the journeys are already captured at node-level. Each journey leg is completed within a node, and the edges in C-space represent transfers between routes. As transfers are done at stops, the C-space representation can in this sense be seen as dual to L- and P-space representations.

C-space with self-loops

Following the above reasoning and considering that most of the journeys have one or two legs, we propose a variation of the C-space representation with self-loops on all nodes. In this representation, the one-leg journey is a path on the self-loop. A two-leg journey on routes A and B that share a stop is represented by entering the self-loop at node v_A, travelling along the edge e_{AB}, and finishing in the self-loop at node v_B. Accounting for the service frequency being an important determinant for passenger demand, and high percentages of one-leg journeys, it is expected that frequency is a more important predictor of route-level passenger demand than route interconnectedness. To integrate this with the connectedness of routes, we propose a weighted C-space representation with self-loops with two different types of weights:

- Weights on connecting edges w_{ij}: the number of shared stops between the corresponding routes.
- Weights on self-loops w_{ii}: service frequency on route i.

Note that both weights are of strength-type. Calculating centrality measures for a network with two different quantities for weights is not straightforward for most measures. E.g., calculating the degree centrality is not sensible, as summing the route frequency and numbers of shared stops gives no meaningful measure. Of the four measures, the eigenvector centrality is the only suitable quantity in this case. In the basic C-space representation, the diagonal elements of the adjacency matrix a_{ii} are all zero, and the score of node i is recursively dependent on the scores of other nodes, being in term dependent on the score of node i:

$$\lambda x_i = \sum_{j \neq i} a_{ij} x_j. \tag{5}$$

In the self-loop C-space representation, the equations change so that node i is dependent on the strength of connection to other nodes, but also to its frequency-weighted self:

$$\lambda x_i = \sum_{j \neq i} a_{ij} x_j + a_{ii} x_i, \tag{6}$$

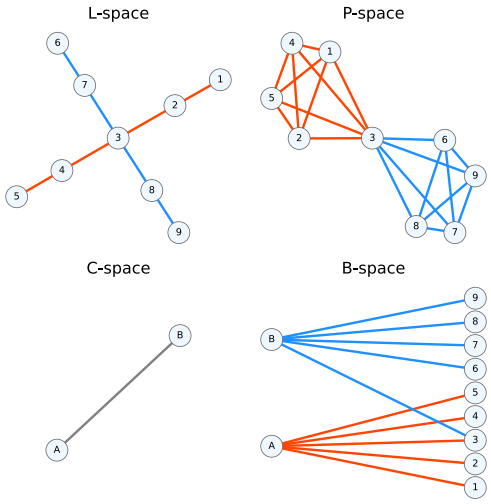


Fig. 1. The four graph representations of a simple toy PTN with nine stops and two routes. Route A (shown in red) services stops 1, 2, 3, 4, and 5; route B (shown in blue) services stops 6, 7, 3, 8, and 9. Top left: L-space, or the space-of-infrastructure. Top right: P-space, or the space-of-service. Each route becomes a clique, and they share a node (stop 3). Bottom left: C-space; the two nodes represent routes A and B, and there is an edge connecting them, meaning they share a stop (here, stop 3). Bottom right: B-space. In the left column are the route nodes and in the right column are the stop nodes. Edges $(A,1),(A,2),(A,3),(A,4),(A,5)$ indicate that stops 1, 2, 3, 4, and 5 lie on route A (and similarly for route B).
Source: Adapted from [2].

or,

$$(\lambda - a_{ii})x_i = \sum_{j \neq i} a_{ij} x_j. \tag{7}$$

This formulation incorporates route frequency as the leading factor in the determination of its importance, while taking into account also the connectedness to other routes. In line with the above reasoning, we expect this measure to exhibit a significantly better correspondence to route-level passenger demand than any centrality measures in the basic C-space representation.

2.2.3. Combined levels

B-space is a bipartite graph representation, where the two bipartition sets of nodes represent stops and routes, respectively. There is an edge between two nodes if the corresponding stop lies on the corresponding route. A weighted version can be constructed, where service frequencies are added as edge weights. In contrast to L-, P- and C-space representations, where route-stop relations are present implicitly, in B-space these interconnections are captured explicitly. Calculation of centrality scores can be applied also to bipartite networks. However, the results will have different meanings, and often different scales, for both bipartitions. This is due to the size difference between the two sets, e.g. in PTNs, the number of stops is typically an order of magnitude higher than the number of routes.

Here, eigenvector centrality was calculated for the weighted bipartite graph representation. Correlations to passenger demand at stop and route levels were calculated separately for the centrality scores of each bipartite set. The eigenvector centrality was chosen due to the importance of each stop being recursively dependent on the importance of the route and the importance of other stops on the route (and similarly for routes).

In entry-only systems, partial paths of passenger journeys can be reconstructed in B-space, i.e. the path from the first stop to the last route is known from smart-card data in entry-only systems. To complete the journey in B-space, the link from the last route to the last exit stop is missing. This suggests the B-space analysis as a potential candidate to balance the trade-off between granularity of network description and completeness of passenger information in entry-only systems.

The overview of the used representations, and the corresponding centrality measures, calculated in each case, is given in Table 2. For a graphical demonstration of the four representations on a simple toy network, see Fig. 1.

2.3. Statistical analysis

To assess the strength of relations between all pairs of variables, correlation coefficients are calculated. The most commonly used is the Pearson correlation coefficient r which measures the degree of linear dependence between two variables X and Y and is defined as:

$$r = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{8}$$

Table 2
An overview of the stop-level representations used in the analysis. Original weights of the representation are shown in column Representation weight. Each weighted representation is further divided into strength- and cost-like weights in the Tye of weight column. The value *'yes'* in the $1/w$ column indicates that the inverse weight transformation was used in the calculation. The last four columns show which centralities were calculated for which representation.

Representation	Weight	Type of weight	$1/w$	C^D	C^C	C^B	C^E
L-space	Unweighted	1		✓	✓	✓	✓
	Frequency	strength		✓			✓
		cost	yes		✓	✓	
	Time	cost			✓	✓	
		strength	yes	✓			✓
P-space	Unweighted	1		✓	✓	✓	✓
	Frequency	strength		✓			✓
		cost	yes		✓	✓	
	Time	cost			✓	✓	
		strength	yes	✓			✓
C-space	Unweighted	1		✓			✓
	No. of shared stops	strength		✓			✓
C-space with SL	No. of shared stops/Frequency	strength					✓
B-space	Frequency	strength					✓

Here, cov is the covariance and σ is the standard deviation. Values of r lie in the interval $[-1, 1]$, where $r = 1$ ($r = -1$) means perfect positive (negative) linear correlation, while $r = 0$ means that linear correlation is not observed, however, independence is not necessarily implied.

Since it was observed that most relations in our data do not appear linear and the measurements tend to cluster at lower values, the Pearson correlation coefficient is calculated also for logarithmic transformations of both variables. The Pearson correlation then examines linear relationship between logarithmic transformations of X and Y:

$$\log Y = a \log X + b. \tag{9}$$

If linear dependence is observed, this indicates a power-law relationship between the original variables, since after exponentiation on both sides, the relationship is of the form:

$$Y = cX^a, \tag{10}$$

where $c = e^b$.

In addition, we calculate the Spearman ranking correlation coefficient which measures the monotonicity of the relationship regardless of its shape and is used since relations in our data are observed to be monotonic. It is defined similarly to the Pearson correlation coefficient, but takes measurement ranks $R(X)$ and $R(Y)$ instead of values:

$$r_S = \frac{\mathrm{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}. \tag{11}$$

Similarly to r, values of r_S lie in the interval $[-1, 1]$ and the difference in the interpretation is the indication of monotonicity instead of linearity. Since logarithm is a positive monotonic transformation, the Spearman ranking coefficient is not affected by its application to the variables.

3. Application and results

3.1. Case study and data description

The proposed methodology was applied on a case study of a bus PTN in Ljubljana, Slovenia. Graph representations were constructed using the publicly available GTFS data.[1] Smart card data were obtained from the local operator (Ljubljanski potniški promet - LPP) for one representative workday with favourable weather conditions (Wednesday, May 11th 2022). The Ljubljana bus PTN is an entry-only fare collection system, so only boarding data are available.

After merging the two datasets and keeping stops present in both, the L-space (P-space) representation has 403 nodes and 494 (9113) edges. Stops on opposite sides of the road, serving as the same stop, were merged into one node. Route frequencies were calculated as the number of trips on a route per hour, averaged for the time period from 5AM to 12PM. Fig. 2 shows the L-space representation of the network, where nodes are positioned according to their geographical coordinates and the edges are presented schematically as straight lines. The size of the node is proportional to the total number of boardings at the corresponding stop.

[1] https://www.transit.land/operators/o-u24q-ljubljanskipotni%C5%A1kipromet

Fig. 2. L-space representation of the Ljubljana bus PTN. Nodes are plotted with their geographical coordinates and the edges are schematic. The size of a node is proportional to the total number of boardings at the respective stop.

After data pre-processing where routes in the opposite directions of the same line are merged into one node and also counting several versions (extensions) of the same route as one, the C-space representation has 23 nodes. 116650 passenger daily boardings are recorded. Passenger demand at stop-level and route-level was estimated by counting total passenger boardings at each stop and route, respectively, in the given 5AM–12PM time window.

To estimate the percentages of journey lengths in P- and C-space, the number of journey legs were determined using the system-defined window of transfer, which is 90 min in Ljubljana. Therefore, if subsequent boardings of the same card were recorded in the 90-minute time window, the boarding is classified as a transfer. If another boarding for the same card was recorded outside that range, it was classified as a separate journey. Following these definitions, it was found that about 56% of journeys consist of one leg, while about 27% consist of two legs (i.e. involve one transfer), and the remaining 17% were journeys with more than two legs.

The complete analysis software was written in Python. Network construction and analysis were performed using the Python NetworkX library [28].

3.2. Stop-level analysis

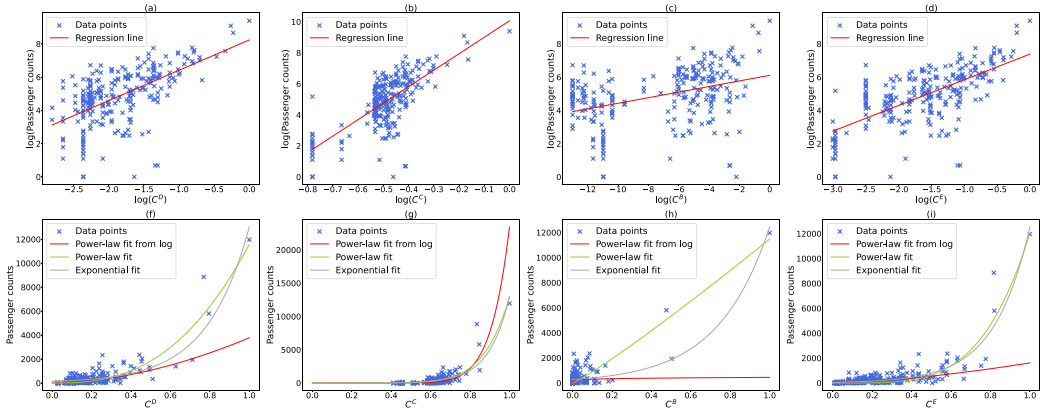
In this section, the relations between centrality measures and passenger counts for all representations are reported. Note that in both L- and P-space, as well as C-space, outliers are present. In the L- and P-space, there are three nodes with significantly higher passenger counts and (in many cases) centrality scores. Their boarding counts are: 11965, 8856, 5803 (while all other stops have boarding counts below 2400) and together they amount to almost a quarter of all boardings taking place in the system. Such stops are known as hubs in PT networks. Therefore, the outliers are exactly the most relevant stops and the implications for the analysis are described in the results, where applicable.

3.2.1. P-space

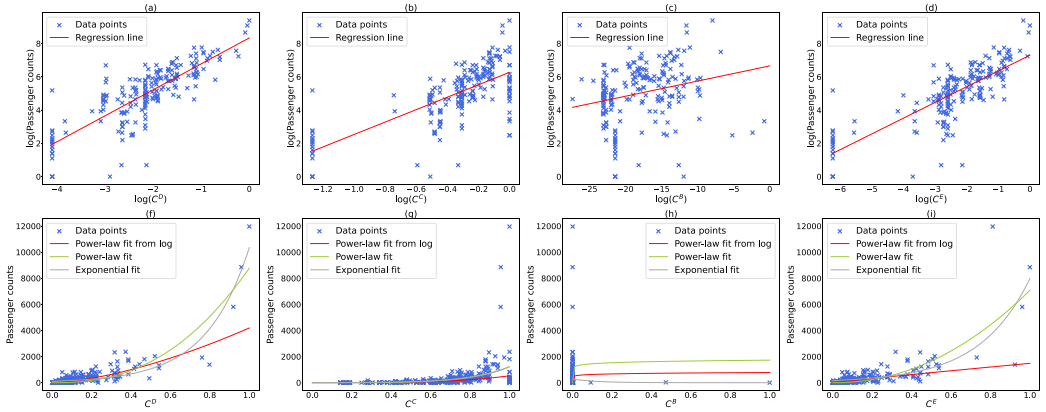
Centrality scores for all four measures (C^D, C^C, C^B, C^E) were calculated for all three graph representations (G_P, G_P^f, G_P^t) and were normalised to the interval $[0, 1]$ by dividing each entry with the maximum value of the variable. This gives $3 \times 4 = 12$ cases in the P-space analysis.

For the regression analysis, each of the twelve centrality measures was treated separately as an independent variable, and passenger counts (P.C.) were the dependent variable. For each case, exponential ($y = \alpha e^{cx}$) and power-law ($y = \beta x^b$) fits were calculated (bottom rows of Figs. 3(a), 3(b), and 3(c)). Additionally, logarithmic transforms of both the independent and dependent variable were used to produce linear fits and transformed to power-law fits ($y = \gamma x^a$) of the original variables according to Eq. (10) (top rows of Figs. 3(a), 3(b), and 3(c)). In the case of betweenness centrality, $C^B = 0$ for several nodes; these data points were excluded from the log-transform and the corresponding fits. The values of fit parameters are provided in Appendix A.1 in Table A.6.

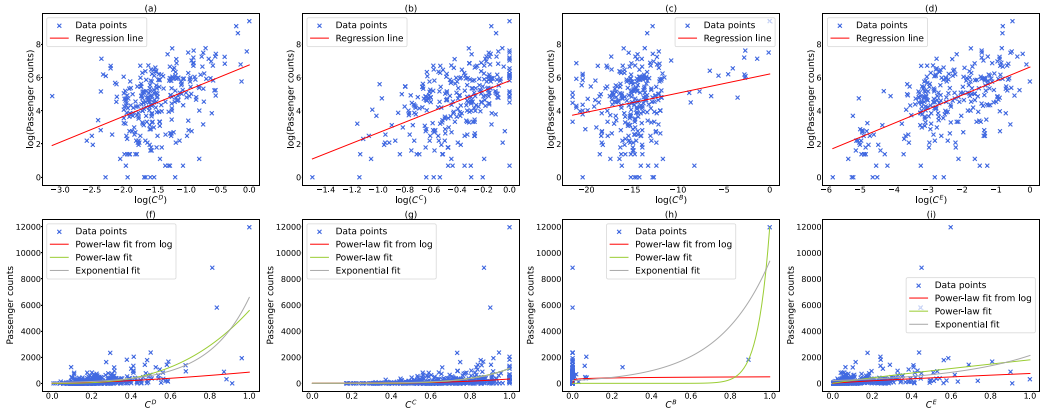
For each of the 24 cases (3 weights × 4 centralities × 2 functions), three correlation coefficients were calculated: Pearson, Spearman and Pearson for log-transformed variables. Values of correlation coefficients between centralities and passenger counts for all graph representations are summarised in Table 3. Correlation matrices for each case are included in Appendix A.1 in Fig. A.8. The correlations are the strongest for the frequency-weighted representation, followed by the unweighted representation. This corresponds to the expectations, due to the known property of passenger route choice behaviour where passengers value each transfer and their waiting times (resulting from high frequency bus arrivals) more highly than a minute of actual in-vehicle time (see for example route choice model results estimated by [29,30] based on passenger journey data available from smart card records).



(a) Scatter plots of passenger counts vs. centrality measures in the unweighted P-space representation. Each column shows a different centrality (from left to right: C^D, C^C, C^B, C^E). In each column, the the upper figure shows plots in log-log scale along with the best linear fit. In the bottom plot, the same relationship is plotted in linear scale, along with the best exponential and power-law fits, as well as power law fits from the linear fit in log scale.



(b) Frequency-weighted P-space representation. Explanation is the same as given in 3a.



(c) Time-weighted P-space representation. Explanation is the same as given in 3a.

Fig. 3. Relations between passenger counts and centrality measures in the P-space representations.

In most cases, the correlations are the highest for degree and eigenvector centrality. This is especially prominent in the frequency-weighted representation, where the values of the Pearson correlation coefficients for log-transformed variables reach values of about 0.8, which indicates a significant dependence. The correlations are high also in the unweighted representation, but observably lower than in the frequency-weighted case (0.6–0.7). The results are in line with the expectations for two reasons: first, from the transport

Table 3
Correlation coefficients between each centrality measure and passenger boarding counts for all three graph representations in P-space. r is the Pearson correlation coefficient, r_S is the Spearman correlation coefficient, and r_{log} is the Pearson correlation coefficient for logarithmic transforms of both variables.

	G_P			G_P^f			G_P^t		
	r	r_S	r_{log}	r	r_S	r_{log}	r	r_S	r_{log}
C^D	0.74	0.54	0.58	0.77	**0.81**	**0.81**	0.53	0.48	0.41
C^C	0.53	0.67	0.71	0.33	0.75	0.7	0.27	0.46	0.47
C^B	0.82	0.53	0.39	−0.02	0.47	0.25	0.58	0.13	0.22
C^E	0.6	0.72	0.67	0.77	**0.81**	**0.79**	0.36	0.58	0.58

Table 4
Correlation coefficients between each centrality measure and passenger boarding counts for all three graph representations in L-space. r is the Pearson correlation coefficient, r_S is the Spearman correlation coefficient, and r_{log} is the Pearson correlation coefficient for logarithmic transforms of both variables.

	G_L			G_L^f			G_L^t		
	r	r_S	r_{log}	r	r_S	r_{log}	r	r_S	r_{log}
C^D	0.43	0.34	0.35	0.76	**0.82**	**0.85**	0.12	0.18	0.06
C^C	0.32	0.51	0.55	0.32	0.75	0.7	0.31	0.51	0.46
C^B	0.4	0.34	0.4	0.53	0.37	0.47	0.43	0.35	0.37
C^E	0.62	0.52	0.55	0.74	0.63	0.67	−0.02	0.34	0.11

science perspective, due to passenger valuation of time. Second, from the network science perspective, in P-space, where most journeys are of length 1 or 2, degree centrality as a local measure is expected to exhibit strong correspondence, and augmented with service frequency, the transport perspective from the previous argument is naturally captured. While eigenvector centrality is a global measure, a similar discussion applies, since neighbours with higher importance contribute more than less important neighbours. Also from the network science perspective, weighted C^D and C^E are calculated for strength type weights, as is the case for frequency.

Correlations with closeness and especially betweenness centrality are low. This is in line with expectations discussed in Section 2.2. In the unweighted case, C^C shows medium correlation. In the frequency-weighted case, the Pearson correlation is low. The relation between C^C and P.C. in this case, as shown in the scatter plot in the second subfigure in bottom row of Fig. 3(b), is highly non-linear. This is subdued in the log-transforms, which is reflected in the high values of the log-transform Pearson and Spearman correlation coefficients. One noticeable outlier is the high Pearson correlation coefficient ($r_P = 0.82$) for betweenness centrality in the unweighted representation. The reason is the dense clustering of data points in the lower left part of the scatter plot and the few outliers at C^B and P.C. values an order of magnitude higher, as shown in the third subfigure in the bottom row of Fig. 3(b). This artefact disappears in the log-transform where the Pearson correlation coefficient is only 0.39 and the argument is strengthened by the relatively low value of the Spearman correlation.

An interesting observation are the relatively high correlations for C^C in the frequency-weighted representation. This might indicate closeness centrality as a potentially valuable indicator even in the frequency-weighted representation. On the other hand, in the in-vehicle-time-weighted graph, correlations to closeness centrality are low, despite – from the network science perspective – the in-vehicle time being the natural weight in shortest path calculations. This again strengthens the argument for passengers' higher valuation of waiting time.

Betweenness centrality shows the weakest correlation. While it is often the most relevant centrality measure in road networks and predicts nodes (representing junctions) with highest traffic flows and largest probability of congestion [31], in the P-space representation, where shortest paths are most often of length 1, it is not expected to be very informative. For many nodes, values of $C^B = 0$ were observed, meaning that no shortest path traverses them.

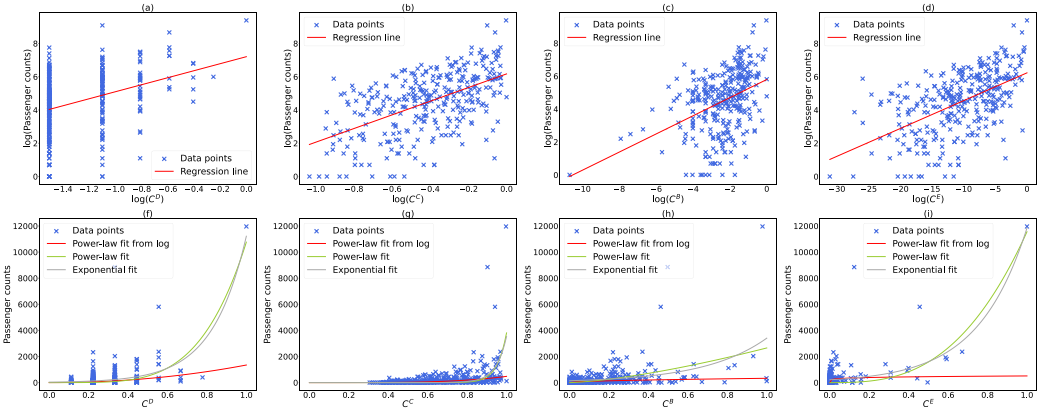
One interesting observation is that almost all power coefficients from the fits are positive, and most are above unity. The exceptions are C^B fits with values close to 0 in most cases, and the linear fit in G_P^f ($a = 0.95$) and linear ($a = 0.85$) and power-law ($b = 0.90$) fits in G_P^t for eigenvector centrality.

3.2.2. L-space

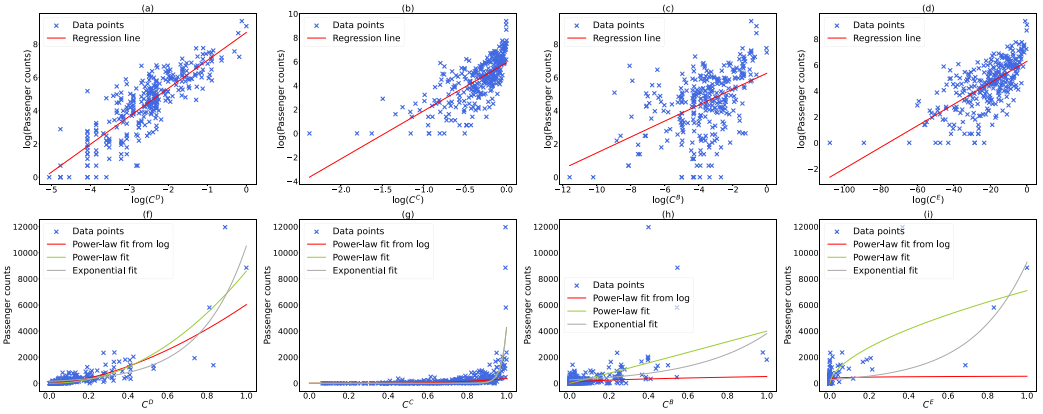
The results of the same analysis of the three L-space representations, G_L, G_L^f and G_L^t, are the subject of this section.

Relations between centralities and passenger counts, along with the best fits, are shown in Fig. 4. It can be seen that data are more scattered than in the P-space plots. In the time-weighted representation, linear fits in the log–log scale could not be performed within $p < 0.05$. Values of correlation coefficients for all graph representations are summarised in Table 4. Fit parameters are provided in Appendix A.2 Table A.7.

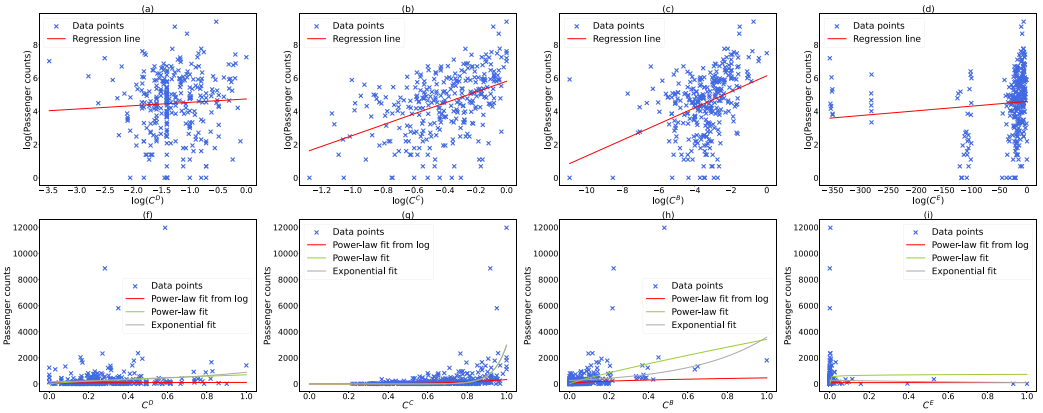
Correlation matrices are included in Appendix A.2 in Fig. A.9. Correlations are smaller than those for the P-space representation, which is in line with expectations, due to the paths in L-space not directly representing journey legs. Similarly to P-space, the highest correlations are observed for the frequency-weighted representation, especially for degree, closeness and eigenvector centrality.



(a) Scatter plots of passenger counts vs. centrality measures in the unweighted L-space representation. Each column shows a different centrality (from left to right: C^D, C^C, C^B, C^E). In each column, the the upper figure shows plots in log-log scale along with the best linear fit. In the bottom plot, the same relationship is plotted in linear scale, along with the best exponential and power-law fits, as well as power law fits from the linear fit in log scale.



(b) Frequency-weighted L-space representation. Explanation is the same as given in 4a.



(c) Time-weighted L-space representation. Explanation is the same as given in 4a.

Fig. 4. Relations between passenger counts and centrality measures in the L-space representations.
Note: for the colour version of the figure, please refer to the online version of the paper.

The high correlation of degree centrality to passenger counts in L-space is somewhat surprising, as is the high correlation to closeness centrality in the frequency-weighted representation, since it means summing the waiting times of all edges on the paths. However, this might be explained by the underlying structure and the shortest paths leading along separate routes. Thus, the

Table 5
Correlation coefficients between each centrality measure and passenger boarding counts for all three graph representations in C-space: unweighted basic C-space (C^u), weighted basic C-space (C^w), weighted C-space with self-loops (C^w_{SL}). r is the Pearson correlation coefficient, r_S is the Spearman correlation coefficient, and r_{log} is the Pearson correlation coefficient for the logarithmic transform of eigenvector centrality in the representation with self-loops. Value '-' indicates that the correlation coefficient was not calculated for the corresponding combination.

	C^u			C^w			C^w_{SL}		
	r	r_S	r_{log}	r	r_S	r_{log}	r	r_S	r_{log}
C^D	0.54	0.40	–	0.62	0.62	–	–	–	–
C^E	0.54	0.35	–	0.60	0.63	–	0.64	**0.78**	**0.82**

weight of the shortest path between nodes i and j along a route in the frequency-weighted representation would be of the form $w_p = w_{ik_1} + w_{k_1k_2} + \cdots + w_{k_nj}$. Since the path is along the same route, all weights are the same: $w_{ik_1} = w_{k_1k_2} = \cdots = w_{k_nk_j}$, so the result is the length in the unweighted representation, multiplied by the service frequency. Of course, this is an approximation, but in a star-like PTN as is the case in Ljubljana, most of the paths follow this rule.

The low correlations for closeness and betweenness centrality for the in-vehicle-time-weighted representation are a further indicator of the users' valuation of time. An interesting difference to the P-space shows are the very low correlations the time-weighted for degree and eigenvector representations (around 0 to 0.2). This is may be a consequence of the $1/x$ transformation for edge weights, lacking straightforward interpretation of connection strength in L-space.

3.3. Route-level analysis

The correlations between route importance and route-level passenger demand are explored in this section. In terms of passenger counts, two nodes are outliers, corresponding to the two routes with most frequent services in the Ljubljana PTN.

3.3.1. C-space

In the basic C-space representation, the correlations between route degree and eigenvector centralities for the unweighted and weighted representations and passenger counts are relatively low, falling in the range $\approx 0.35 - 0.55$ in the unweighted and ≈ 0.6 in the weighted representation (Table 5). Scatter plots (Fig. 5) do not show any observable relationships. Most significantly, the routes with the highest demand do not correspond to the most central routes. This is an indication of inadequacy of centralities as supply indicators at route level, which is in line with our expectations (Section 2.2), as the majority of the journeys consist of one leg, which is a node property in the C-space, and route frequencies, which are not captured in these representations, are known to be good predictors of demand.

3.3.2. C-space with self-loops

The results from the previous section are compared to the results for the doubly-weighed C-space representation with self-loops, where service frequencies are incorporated into the representation.

Scatter plots (Fig. 6) show a logarithmic dependence of passenger demand to eigenvector centrality. Fig. 6(a) shows the scatter plot in linear scale together with the best logarithmic fit. The four nodes with the lowest centrality measures are excluded from the fit. In Fig. 6(b), the results are shown in a logarithmic scale on the x-axis ($\log_{10} C^E$). The correlation coefficients are shown in Table 5. A significant Spearman and Pearson correlation values between the logarithm transform of C^E and passenger counts are observed (≈ 0.8), when the four lowest-centrality routes are excluded from the fit. The data is fitted with a piece-wise linear function; one fit for the lowest four nodes and another for the majority of the points. The results suggest that a minimum value of eigenvector centrality is required for a route to become ''interesting'' for passengers. From the piecewise linear fit, the critical value for the eigenvector centrality for the Ljubljana PTN is estimated to be at the order of magnitude $C^E \approx 10^{-3}$.

This result carries significant implications for PT systems planning, as it shows that some routes may be deemed redundant. The representation with self-loops offers the opportunity to explore the relative effects of route frequency and connectedness on passenger flows, if modifications can be made in the system operation. Alternatively, the four nodes with the lowest centrality measures can be discarded from the analysis and we can observe the limiting behaviour at higher centrality values. The advantage of this representation is that the relation between passenger counts vs. eigenvector centrality can be described by an elementary function where the outliers are taken into account and the routes with the highest number of passengers correspond to highest eigenvector centrality measures. Moreover, a logarithmic function, while divergent, increases slowly and can be interpreted to hint at a kind of saturation of demand with respect to supply. This reflects the fact that the daily number of PT users is finite and, within some variation, approximately constant. From the above analysis, we conclude that the C-space representation with self-loops is significantly more informative than the basic standard graph representation. Some of the implications for route-level accessibility are discussed in Section 4.

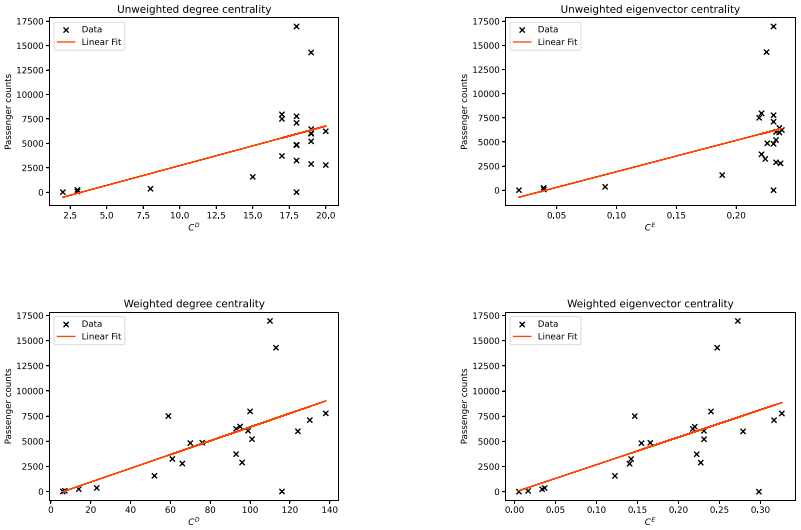
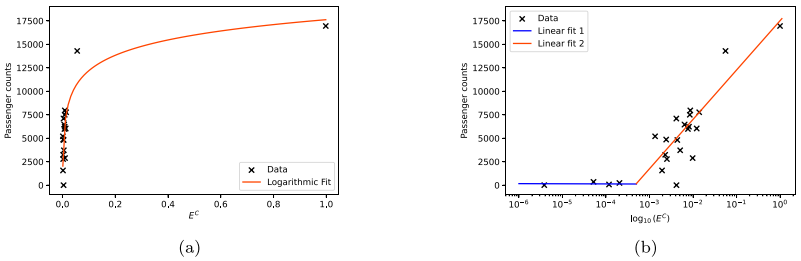
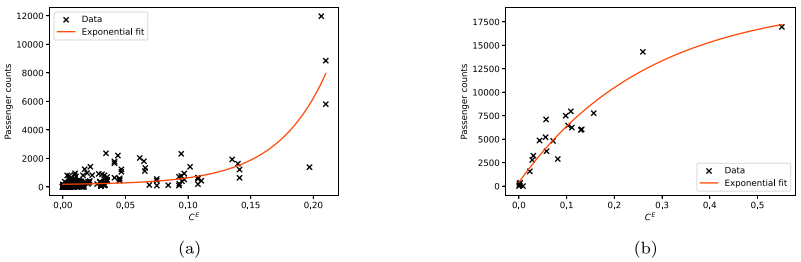


Fig. 5. Scatter plots of passenger counts in relation to degree (left) and eigenvector (right) centrality in unweighted (top row) and weighted (bottom row) C-space representations without self-loops. Added are the best linear fits in each case.



(a) (b)

Fig. 6. Scatter plots of passenger counts in relation to eigenvector centrality in the weighted C-space representations with self-loops. (a) Logarithmic fit. (b) Piece-wise-linear fit in the logarithmic x-scale.



(a) (b)

Fig. 7. Scatter plots of passenger counts in relation to eigenvector centrality in the B-space representation, along with the best exponential fits. plotted for the (a) stop node set (b) route node set.

3.4. Integrated stop- and route-level analysis

The results are shown in Fig. 7. Fig. 7(a) shows the scatter plot for eigenvector centrality vs. passenger counts for the stop set together with the best exponential fit. It should be noted that the relationship is not exactly exponential (or of any other elementary shape).

Fig. 7(b) shows the scatter plot for the route set together with the best exponential fit. The frequencies are included as edge weights and in comparison to the C-space representation, stop node centrality is included in the importance score. This turns out to be the best fit for centrality vs. passenger counts at the route level.

4. Discussion and conclusion

4.1. Key findings and study implications

The results of the analysis of the case study on the bus public transport network in Ljubljana, Slovenia, show significant correspondence between centrality-based PT supply indicators and passenger demand at all levels: stop, route and combined.

To answer our first research question, the analysis of the relation of each centrality measure to passenger counts at stop level showed significant correlations, especially for degree and eigenvector centrality in the frequency-weighted P-space representation, which is in line with our expectations (2.1.1, 2.2.1). The lowest correlation strengths were observed in in-vehicle-time-weighted representations. Our results strengthen the argument for the relative importance of service frequency, compared to in-vehicle time, as PT use predictors, based on passenger valuation of time (e.g. [30]).

The correlations were low for betweenness centrality in almost all cases, and generally also for closeness centrality. This is in contrast with [23], where it was observed that C^D, C^C and C^B are the most informative measures of centrality, while eigenvector, PageRank and clustering do not sufficiently discriminate between nodes. Low observed correlations for closeness and betweenness centrality in our analysis could also be due to the small size of the network and it would be instructive to study the relative importance of different centralities with respect to network size, and potentially general topology (e.g. star-like, grid, concentric etc.). While closeness centrality is a suitable accessibility indicator, the low correspondence to passenger demand suggests the use of other indicators, such as eigenvector centrality, in mobility-based planning.

Most of the relations are sufficiently described with power-law fits, suggesting the Pearson correlation coefficient for log-transformed variables as the most appropriate measure of correspondence (2.3, Eq. (9)). In almost all cases, the relations of passenger counts to centrality measures are super-linear. This may be an indicator of network inefficiency since supply at nodes does not sufficiently correspond to passenger demand. Alternatively, fast growth of passenger counts at higher values of centralities might indicate a disproportional centralisation of the network with high demands at only a few of the nodes. Moreover, the capacity of PT services is dimensioned to guarantee a certain level of service and reasonably accommodate the highest demand levels while minimising the amount of required resources. This can explain why service capacity is over-dimensioned for most of the PTN while being somewhat under-dimensioned for the stops characterised by the highest demand level (see for example [32]). The studied correlation coefficients and relationship fits could also serve as measures of network quality, as has already been proposed in [22].

Conversely, non-linear relationships may indicate a level of inadequacy of the introduced supply indicators. While network connectedness, as reflected in centrality measures, is expected to correlate to passenger demand, it does not offer a full picture of PT use. External factors, such as land use or closeness of hubs of other transport modalities are exogenous to the PT system, and while PTNs are designed to cater to these demands, there is not a perfect correspondence between actual supply and external factors of demand. Besides, the definitions of node centrality are designed to describe the complex nature of networks, and a linear correlation to passenger demand, as a quantity extrinsic to the network, might not be always expected.

To answer our second research question, in the basic C-space representation low correspondence between centrality-based supply indicators and route-level demand were observed, in line with expectations 2.2.2. The modified C-space representation with self-loops introduced in this paper and the corresponding eigenvector centrality measure as a proposed indicator of route-level supply, offers insights into interplay between connectedness at route level and route frequency. A regular relationship between the supply indicator and passenger demand of logarithmic shape strengthens the evidence for usefulness of this representation, as an elementary-function-shaped relation might indicate a deeper connection between network topology and passenger behaviour. The observation of a critical eigenvalue score as a supply indicator to start facilitating demand is a crucial result and may be used especially in evaluating performance of non-central routes.

The B-space representation explicitly connects the two levels. To explore the mutual impact and the interaction of the importance of the stops on a route and the importance of routes serving a stop, eigenvector centrality scores were calculated for the bipartite network and relations to passenger boarding counts were fitted separately for stops and routes. The results from our case study show exponential convergence of passenger counts with eigenvector centrality at route level, implying the potential value of this representation to enable multi-level approaches to supply–demand interaction analysis. This answers our third research question.

4.2. Study limitations and directions for further research

One obvious limitation for the extent of correspondence between passenger demand and supply indicators reported in our analysis is the lack of any land use or other extrinsic data. An important extension of this analysis will be to include external data to provide better understanding of passenger demand patterns, potentially also providing a partial answer to the question of the direction of supply–demand causality.

The results from route- and combined-level analysis can be used to guide PT systems planning and optimisation. Using the C-space representation with self-loops, simulation studies can give insights into the relative importance of frequency and shared stops, as well as their interaction. Using the proposed indicators as proxies of passenger demand can guide operators to find optimal route frequencies or possible route modifications by including more important stops. This can be explored more deeply and thoroughly by combining the route-level analysis with combined levels and considering also the stop importance. Where historical data is available, the changes in C- and B-space properties and their correlations to passenger demand over time can be explored.

We expect that the use of the route- and combined-level representations might potentially lead to stop-level demand predictions, which is of primary interest for operators. Inferences to stop level might be possible by using advanced computational methods, such as link prediction in B-space to infer the exit stops. We see this as a promising venue for future research.

Table A.6
Values of fit parameters in P-space for all four centralities and three weighted representations. Each row includes fit parameters for passenger counts vs. respective centrality. For each representation (unweighted: G_P, frequency-weighted: G_P^f, time-weighted: G_P^t), the three fit parameters are shown: a: power-law fit from log–log regression ($y = k_1 x^a$); b: power-law fit on original data ($y = k_2 x^b$); c: exponential fit on original data ($y = k_3 e^{cx}$).

	G_P			G_P^f			G_P^t		
	a	b	c	a	b	c	a	b	c
C^D	1.82	2.65	4.65	1.56	2.45	4.56	1.55	2.94	4.81
C^C	10.62	7.73	9.54	3.74	4.71	5.49	3.13	4.89	5.75
C^B	0.17	1.11	3.79	0.09	0.076	−8.29	0.12	16.60	3.78
C^E	1.54	4.23	5.52	0.95	2.09	3.91	0.85	0.90	2.33

Future research may conduct the same analysis for several PTNs to examine the transferability of these results to other contexts, and potentially identify some general patterns. Moreover, other centrality measures may be examined, based on their interpretation in describing passenger demand properties. A promising venue for further research is including a temporal aspect and study the correspondence between centrality-based supply indicators and passenger demand in different time periods, performing separate analyses for different parts of the day (rush hour etc.), days of week (workday or weekend/holiday) and at other time scales. As discussed above, including extrinsic data, such as land use and presence of hubs of other modalities will offer a more complete picture of the demand side and enable the construction of suitable supply measures.

In conclusion, this research deepens the understanding of the relationship between network structure and passenger demand by examining in detail several centrality measures and interpreting the correlations in terms of their definitions and scope. Particularly, the route- and combined-level analysis is novel and opens several potential novel research venues.

CRediT authorship contribution statement

Tina Šfiligoj: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Aljoša Peperko:** Writing – review & editing, Supervision, Conceptualization. **Patricija Bajec:** Writing – review & editing, Supervision, Conceptualization. **Oded Cats:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first and second author were partially supported by ARIS-Slovenian Agency for Research and Innovation (grants P2-0394 (A) and P1-0222, respectively). The authors thank Francesco Bruno for a detailed reading of the first version and valuable suggestions for improving the writing in the second version.

Appendix. Stop-level correlation matrices and fits

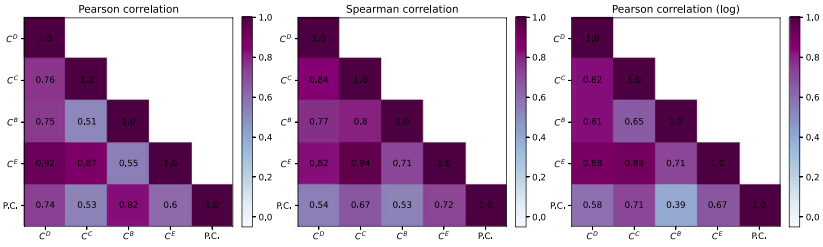
In this Appendix, additional results for stop-level analysis are included. Correlation matrices of all independent and dependent variables for each type of weighted representation in P-space are shown in Fig. A.8, and similarly for L-space in Fig. A.9. Fit parameters of power-law and exponential fits of passenger counts vs. each centrality relations are given in Table A.6 and Table A.7 for P-space and L-space, respectively.

A.1. P-space

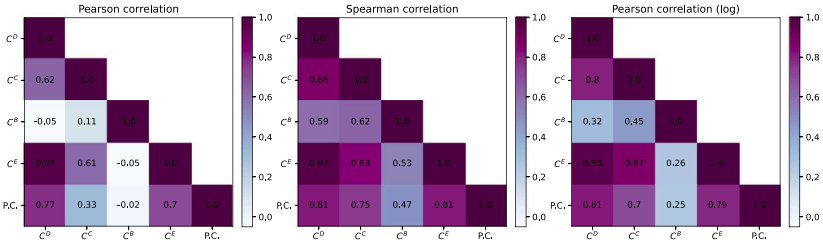
See Fig. A.8.
See Table A.6.

A.2. L-space

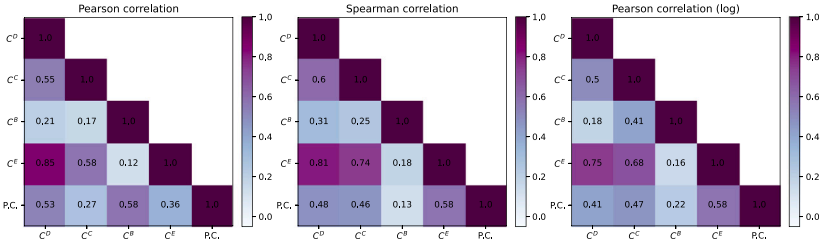
See Fig. A.9.
See Table A.7.



(a) Correlation matrices for the unweighted P-space representation G_P^u. From left to right: Pearson, Spearman and Pearson for log-transformed variables.



(b) Correlation matrices for the frequency-weighted P-space representation G_P^f. From left to right: Pearson, Spearman and Pearson for log-transformed variables.

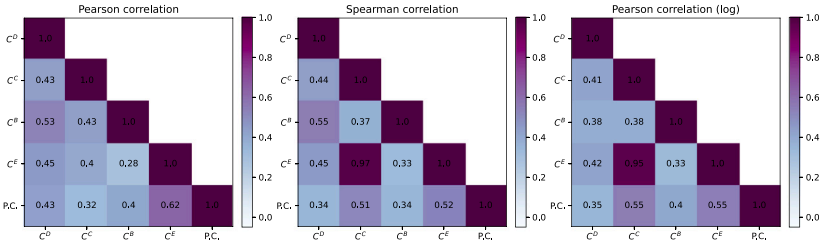


(c) Correlation matrices for the time-weighted P-space representation G_P^t. From left to right: Pearson, Spearman and Pearson for log-transformed variables.

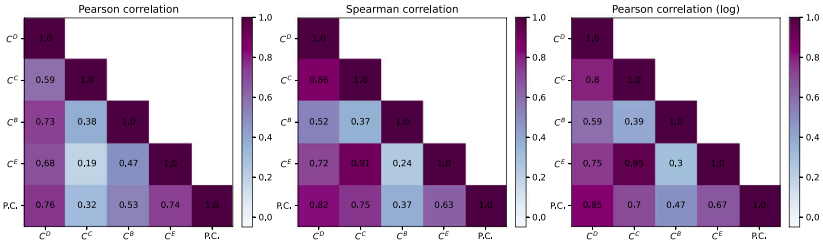
Fig. A.8. Correlation matrices for the P-space representations.

Table A.7
Values of fit parameters in L-space for all four centralities and three weighted representations. Each row includes fit parameters for passenger counts vs. respective centrality. For each representation (unweighted: G_L, frequency-weighted: G_L^f, time-weighted: G_L^t), the three fit parameters are shown: a: power-law fit from log–log regression ($y = k_1 x^a$); b: power-law fit on original data ($y = k_2 x^b$); c: exponential fit on original data ($y = k_3 e^{cx}$).

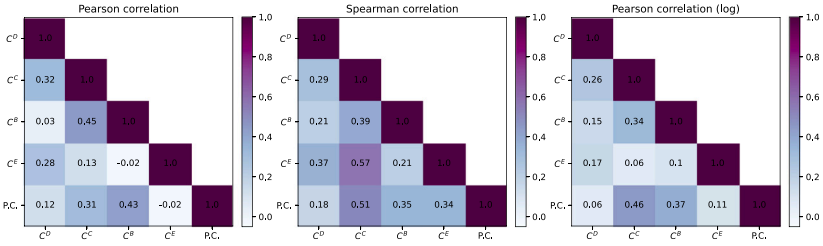
	G_L			G_L^f			G_L^t		
	a	b	c	a	b	c	a	b	c
C^D	2.11	4.43	5.79	1.68	2.10	4.24	0.20	0.74	1.69
C^C	4.11	20.51	18.77	4.01	38.28	38.51	3.26	15.10	15.10
C^B	0.55	1.44	3.17	0.48	1.03	2.68	0.48	0.87	2.63
C^E	0.17	2.87	4.01	0.08	0.59	3.53	0.003	0.063	−1.00



(a) Correlation matrices for the unweighted L-space representation G_L^u. From left to right: Pearson, Spearman and Pearson for log-transformed variables.



(b) Correlation matrices for the frequency-weighted L-space representation G_L^f. From left to right: Pearson, Spearman and Pearson for log-transformed variables.



(c) Correlation matrices for the time-weighted L-space representation G_L^t. From left to right: Pearson, Spearman and Pearson for log-transformed variables.

Fig. A.9. Correlation matrices for the L-space representations.

Data availability

Data will be made available on request.

References

[1] C. von Ferber, T. Holovatch, Y. Holovatch, V. Palchykov, Network harness: Metropolis public transport, Physica A: Stat. Mech. Appl. 380 (2007) 585–591.
[2] C. von Ferber, T. Holovatch, Y. Holovatch, V. Palchykov, Public transport networks: empirical analysis and modeling, Eur. Phys. J. B 68 (2009) 261–275.
[3] J. Sienkiewicz, J.A. Hołyst, Statistical analysis of 22 public transport networks in Poland, Phys. Rev. E—Stat. Nonlinear, Soft Matter Phys. 72 (4) (2005) 046127.
[4] S. Feng, B. Hu, C. Nie, X. Shen, Empirical study on a directed and weighted bus transport network in China, Physica A: Stat. Mech. Appl. 441 (2016) 85–92.
[5] X. Qing, Z. Zhenghu, X. Zhijing, W. Zhang, T. Zheng, Space P-based empirical research on public transport complex networks in 330 cities of China, J. Transp. Syst. Eng. Inf. Technol. 13 (1) (2013) 193–198.
[6] A. Háznagy, I. Fi, A. London, T. Nemeth, Complex network analysis of public transportation networks: A comprehensive study, in: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS, IEEE, 2015, pp. 371–378.
[7] S.D. Dimitrov, A.A. Ceder, A method of examining the structure and topological properties of public-transport networks, Physica A: Stat. Mech. Appl. 451 (2016) 373–387.
[8] A. Chatterjee, M. Manohar, G. Ramadurai, Statistical analysis of bus networks in India, PLoS One 11 (12) (2016) e0168478.
[9] K. Wang, X. Fu, Research on centrality of urban transport network nodes, in: AIP Conference Proceedings, vol. 1839, AIP Publishing LLC, 2017, 020181.
[10] R. de Regt, C. von Ferber, Y. Holovatch, M. Lebovka, Public transportation in great britain viewed as a complex network, Transp. A: Transp. Sci. 15 (2) (2019) 722–748.
[11] H. Pu, Y. Li, C. Ma, Topology analysis of lanzhou public transport network based on double-layer complex network theory, Physica A: Stat. Mech. Appl. 592 (2022) 126694.
[12] Y.-Z. Chen, N. Li, D.-R. He, A study on some urban bus transport networks, Physica A: Stat. Mech. Appl. 376 (2007) 747–754.
[13] T. Shanmukhappa, I.W.-H. Ho, K.T. Chi, K.K. Leung, Recent development in public transport network analysis from the complex network perspective, IEEE Circuits Syst. Mag. 19 (4) (2019) 39–65.

[14] L.-N. Wang, K. Wang, J.-L. Shen, Weighted complex networks in urban public transportation: Modeling and testing, Physica A: Stat. Mech. Appl. 545 (2020) 123498.
[15] S. Derrible, C. Kennedy, Applications of graph theory and network science to transit network design, Transp. Rev. 31 (4) (2011) 495–519.
[16] O. Cats, The robustness value of public transport development plans, J. Transp. Geogr. 51 (2016) 236–246.
[17] A. Kopsidas, K. Kepaptsoglou, Identification of critical stations in a metro system: A substitute complex network analysis, Physica A: Stat. Mech. Appl. 596 (2022) 127123.
[18] O. Cats, P. Krishnakumari, Metropolitan rail network robustness, Physica A: Stat. Mech. Appl. 549 (2020) 124317.
[19] K. Sugishita, Y. Asakura, Vulnerability studies in the fields of transportation and complex networks: a citation network analysis, Public Transp. 13 (1) (2021) 1–34.
[20] D. Luo, O. Cats, H. van Lint, G. Currie, Integrating network science and public transport accessibility analysis for comparative assessment, J. Transp. Geogr. 80 (2019) 102505.
[21] G. Sarlas, A. Páez, K.W. Axhausen, Betweenness-accessibility: Estimating impacts of accessibility on networks, J. Transp. Geogr. 84 (2020) 102680.
[22] D. Luo, O. Cats, H. van Lint, Can passenger flow distribution be estimated solely based on network properties in public transport systems? Transportation 47 (2020) 2757–2776.
[23] A.M. Senousi, X. Liu, J. Zhang, J. Huang, W. Shi, An empirical analysis of public transit networks using smart card data in Beijing, China, Geocarto Int. 37 (4) (2022) 1203–1223.
[24] Y. Sui, F. Shao, X. Yu, R. Sun, S. Li, Public transport network model based on layer operations, Physica A: Stat. Mech. Appl. 523 (2019) 984–995.
[25] A. Kopsidas, A. Douvaras, K. Kepaptsoglou, Exploring the association between network centralities and passenger flows in metro systems, Appl. Netw. Sci. 8 (1) (2023) 69.
[26] Z. Wang, D. Luo, O. Cats, T. Verma, Unraveling the hierarchy of public transport networks, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC, IEEE, 2020, pp. 1–6.
[27] J.B. Ingvardson, O.A. Nielsen, How urban density, network topology and socio-economy influence public transport ridership: Empirical evidence from 48 European metropolitan areas, J. Transp. Geogr. 72 (2018) 50–63.
[28] A. Hagberg, P.J. Swart, D.A. Schult, Exploring Network Structure, Dynamics, and Function Using Networkx, Tech. rep., Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2008.
[29] M. Dixit, O. Cats, N. van Oort, T. Brands, S. Hoogendoorn, Validation of a multi-modal transit route choice model using smartcard data, Transportation (2023) 1–21.
[30] M. Yap, H. Wong, O. Cats, Passenger valuation of interchanges in urban public transport, J. Public Transp. 26 (2024) 100089.
[31] P. Crucitti, V. Latora, S. Porta, Centrality measures in spatial networks of urban streets, Phys. Rev. E 73 (3) (2006) 036125.
[32] O. Cats, S. Glück, Frequency and vehicle capacity determination using a dynamic transit assignment model, Transp. Res. Rec. 2673 (3) (2019) 574–585.