Hindawi
Complexity
Volume 2021, Article ID 3280777, 19 pages
https://doi.org/10.1155/2021/3280777

Research Article

Modeling the Public Transport Networks: A Study of Their Efficiency

Mary Luz Mouronte-López [1,2]

[1]*Higher Polytechnic School, Universidad Francisco de Vitoria, Madrid, Spain*
[2]*Telefónica Chair, Universidad Francisco de Vitoria, Madrid, Spain*

Correspondence should be addressed to Mary Luz Mouronte-López; maryluz.mouronte@ufv.es

Received 19 May 2021; Revised 28 July 2021; Accepted 3 August 2021; Published 14 August 2021

Academic Editor: Jing-Hu Pan

The public transportation network (PTN) provides mobility and access to community resources, employment, medical care, infrastructures, and other resources in the city. This research studies the process of the formation of links among nodes in different real-world PTNs. We have found that this process may be appropriately explained by a generalized linear model (GLM) using local, global, and quasilocal similarity indexes as explanatory variables. In modeling, the response variable was described by a binomial probability density function, and the logit function was used as a link function. In the crossvalidation process, utilising a downsampling approach, both average accuracy and area under the receiver operating characteristic curve (AUC) metrics presented higher values than 0.99. The kappa parameter had magnitudes larger than 0.93 for most of the PTNs. In the final validation stage, recall and specificity metrics took the value 1. Accuracy and precision parameters were larger than 0.99 and 0.87, respectively, for the majority of PTNs. Only one of the PTNs required utilising a smoothed bootstrap approach in order to achieve better results. The similarity measures with the greatest influence on the model were determined. We also assessed the impact of link removal on the global efficiency of PTNs, considering several similarity indexes. Additionally, we find that most of the networks show low local and global efficiencies ($\leq$0.20), as well as travel times with a relevant variability, exhibiting standard deviations larger than 790 seconds. Significant similarities exist between the cumulative probability distributions of the local efficiency in all PTNs. With respect to the centrality measures, the eigenvector centrality presented a strong correlation with the hub/authority centralities (>0.80), while the pagerank showed a moderate, high, or very high correlation with the degree in all PTNs, >0.50.

1. Introduction

Link prediction methods have been the subject of research [1–4], which suggests several mechanisms to detect hidden connections. These mechanisms take into account the path information between pairs of nodes in order to estimate their common neighbors. They also consider a mutual information perspective in order to evaluate the similarity index between pairs of nodes. The conditional probability for the existence of a link is calculated, given the common neighbor of two nodes, as described in [5]. Finally, the weight of the links are considered, developing the mechanisms described in [6], which are based on the common neighbor, resource allocation (RA) [7], and adamic adar (AA) [8] indexes. The above is combined with the weighted mutual information (WMI) [9] score estimated between node pairs. Reference [10] suggests a new local information-based link prediction method, tie connection strength index (TCS), concerning the efficient paths between the target node-pair and their common neighbor. An adaptable parameter is presented in order to estimate the impact of the TCS and the topology of the network on the similarity of pairs of nodes. Reference [11] establishes a new type of triangle structure, which consists of one seed node, one common neighbor, and another node. Based on this, a new similarity index, named TRA index by the authors, is proposed for link prediction. The authors integrate the new triangle structure and the idea of RA [7] index [7]. Reference [12] proposed a new similarity

measure based on the AA score, information related to communities generated from the topological structure of the network and the degree centrality. The link prediction algorithms use two open implementations of a bulk synchronous parallel programming model [13]. They are Apache Giraph and Apache Graphx. Reference [14] demonstrates that similarities with respect to structural features (eigenvectors) optimize the link prediction task in multiplex networks. This is done using a layer reconstruction method (LRM), which considers the unconnected node pairs in the target layer as similar, provided that they are not only analogous from the point of view of the target layer but also from the perspective of other layers. Tests on real multiplex networks show that LRM takes advantage of existing information redundancy in different layers.

The application of link prediction methods in real contexts has also been analyzed. A great deal of research is done on the analysis of social networks. Reference [15] carries out a comprehensive review and discusses some link prediction applications in social networks such as recommender systems, community detection, anomaly detection, and influence analysis. Because social networks are highly dynamic with the come-and-go of nodes and links, some research considers temporal aspects. Reference [16] characterizes the likelihood of a link between two nodes from both existing connectivity topology and the popularity of both nodes. Several datasets are considered in order to test and calculate the performance of algorithms. Reference [17] builds a linear model for integrating neighborhood similarity measures and node specific information and uses an evolutionary algorithm to locate the coefficients, which optimizes the prediction of links. The authors assign different weights to each index using the Covariance Matrix Adaptation Evolution Strategy (CMAES) [18, 19]). In addition, the protein-protein interaction (PPI) networks (PPI) have been examined using link prediction methods. Reference [20] utilises the support vector machine learning method for protein-protein interaction (PPI) prediction. Features, often used in social networks, like some similarity index, have been progressively put into practice to make predictions in PPI [21, 22].

This paper studies the link formation process in several PTNs using various similarity measures, which have been applied in a link prediction theoretical framework. The most influential indexes in the pattern followed by link formation between pairs of nodes are determined.

PTNs have been examined from different points of view. Thus, models have been implemented to analyze travel behaviours. Reference [23] forecasts, based on surveys, some characteristics related to the passenger flow. Reference [24] implements a Bayesian network to detect the relationships between travel happiness and several parameters that affect travel behavior. Reference [24] checks pretravel information-seeking behaviours of the passengers using data collected during an extensive public transport on-board survey. For this purpose, the authors implement a multivariate binomial logistic regression model. The model takes into account factors related to sociodemographics, aspects of the travelers, characteristics of the trip, and devices used for information consultation.

The main novelty of our research is that it shows that the link formation pattern in PTNs can be appropriately explained by means of a generalized linear model (GLM), which has local, quasilocal, and global similarity measures between nodes as explanatory variables. The response variable, which establishes whether or not a link exists between pairs of nodes, is described by a binomial probability density function. The link function used is the logit function.

Studies exist that analyze topological parameters in PTNs (degree distributions, path length distribution, and betweenness), as well as growth models. However there are no analysis that we know of, which does this demonstration on PTNs. Research exists, which has developed growth models for PTNs, based on other considerations. Reference [25] replicates some statistical features of PTNs, describing their evolution in terms of adding routes in *P-space*. The authors use a self-avoiding walk (SAW) as a route model. In the aforementioned *P-Space* [26], one node symbolizes one stop, and one link joins a pair of stops, if at least one route exists that supports a direct service between them. Reference [27] developed an area-based model of highway growth. Specifically, a binary logit model in order to estimate the new route growth probability of divided highways and secondary highways using high-quality geographic information system (GIS) data of land-use, population distribution, and highway network for the Twin Cities Metropolitan Area from 1958 to 1990 was obtained in [28]. A growth model that iteratively invested in constructing new links or incrementing the capacity of those existing was implemented. The objective of the research was to establish the impact the demand distributions and operational costs have on the evolution of a PTN. The model considered parameters related to grid geometry, demand characteristics, operating mode parameters (operational speed per mode, cost per km, and capacity). On the contrary, the model described in this paper explains the appearance of links in PTNs based on exclusively topological parameters.

The PTNs also been studied as complex systems [29, 30] describes a geospatial layout for distributing stops and uses a maximum allowable walking distance in order to link the routes. The PTNs are optimized, considering aspects as efficiency and robustness. Reference [31] studies common problems that have been found when a complex system scheme is used for the analysis of the topology of a transportation system (such as mechanisms for the evaluation of the scale-freeness, metrics for the analysis of the network structure, and examination of the vulnerability of the networks using methods with an unacceptable computational time). The vulnerability of the PTNs has also been analyzed in depth [26, 32].

This paper studies the impact that the removal of links, with certain similarity characteristics, has on the global efficiency of PTNs. The relationships between similarity characteristics and the local efficiency of nodes are also checked. Other research has analysed the effect that the node elimination has on the global efficiency of PTNs [33], and the robustness of PTNs has been examined from other points of view, such as the evolution of the giant component when several nodes are deleted [26, 33]. The fault propagation

[20, 26] from nodes with certain topological characteristics (highest betweenness, degree, eigenvector centralities, and pagerank) has also been analyzed. However, a detailed study of the effect on the global efficiency in PTNs when certain links are removed according to similarity indexes analysed in this research has not been found.

This paper also examines the correlation between some centrality measures and relates them to other traffic flow characteristics. Some research exists [34–37] that analyze the correlations between centrality measures in networks of different types. However, we focus on the study of centralities in PTNs and relate them to the flow of vehicles. These characteristics, that we know, have not been previously studied specifically in the PTNs presented here. Moreover, the networks analyzed here are of very different sizes and nationalities, which suggests that they can also operate differently, bringing generality to the analysis. The correlation between centrality measures can explain some of the patterns found in PTN, when a target attack or a fault propagation is suffered by them [26].

The same applies to the study of travel times. It has been shown that, in general, the size, complexity, and variability of available routes in PTNs produce trip times that are highly different between routes. We also study the local efficiency, demonstrating that there are commonalities between PTNs with respect to this feature.

The PTNs studied are AVL, CFL, RGTR, and TICE in Luxembourg, which has 1,372 nodes and 340,684 links; Island Transit in USA, which has 358 nodes and 5,946 links; Lanta in USA, which consists of 2,150 nodes and 91, 583 links; Linja-Karjala Oy in Kuopio, Finland, which has 551 nodes and 63,339 links; Metlink in New Zealand, which has 3007 nodes and 355621 links; Prague Public Transit Company (PPTC), Regional Organiser of Prague Integrated Transport (ROPIT) in Prague, which consists of 5,152 stops and 1,602,778 links; STAR in France, which consists of 1,415 stops and 9,477,213 links; Thunder Bay Transit in Ontario, Canada, which consists of 825 nodes and 78,247 links; TransAntofagasta in Chile, which has 650 nodes and 58 724,362 links; and finally, Sage in California, which has 31 stops and 66 links. It can be observed that the networks are of small, medium, and large sizes.

The vulnerability of AVL, CFL, RGTR, TICE; Linja-Karjala Oy, STAR; Thunder Bay Transit; and Trans-Antofagasta networks was analyzed in [26].

The objectives of this research were as follows:

(1) To analyze whether a GLM, which has as input variables certain measures of similarity between nodes, can correctly explain the formation of links. To establish which of the measures have greater significance in this process.

(2) To detect the influence that the links can have on the global efficiency of the network, according to their similarity characteristics.

(3) To find common features in the networks that allow to characterize their efficiency and trip times).

(4) To determine the relationships that may exist between some centrality measures (eigen vector, pagerank, betweenness, hub, and authority), as well as with other traffic flow characteristics.

2. Materials and Methods

2.1. Overview of Used Resources. Information related to the stops and routes based on the studied networks, which is available on the websites, was utilised. Several programs in R [38] and Python [39] were specifically implemented to carry out this research, using the R.3.6.0 and 3.8.3 version, respectively. The networks and igraph packages were used. In addition, the proxfun, caret, nortest, stats, vip, and rose packages in R were utilised.

The programmes specifically developed to perform this research allowed:

Processing of information related to the PTNs to be able to work with it (routes, stops, stop times, trips, and calendars) (in Python, ProcessPTNInf.py).

Construction and simplification of the graphs that describe a PTN. Obtaining the similarity measures between nodes (in R, ConstGraphCalcSim.R).

Estimation of centralities (in R and python, CalcCentralities.py and CalcCentralities.R).

Building of a binary classification model, evaluating their results (in R, ModelingPTN.R).

Obtaining frequency and cumulative probability distributions related to efficiency and trip times (in R, CalcDistr.R).

Get graphs showing the results (in R, DrawGraphs.R).

These programs followed the typical development life cycle with phases of specification, detailed design, coding, and testing.

2.2. Overview of Used Methods

2.2.1. Generalized Linear Models. This is the generalized linear model (GLM) we have used for the simulation of link formation in PTNs.

Consider the response Y_i and the set of independent variables $X_i = (x_{i1}, x_{ip})$ for $i = 1, \ldots, n$. A GLM consists of both a random and a systematic component, as well as a link function.

Regarding the random component, it is assumed that Y_i, $1 \leq i \leq n$, are independent random variables described by a probability density function from the exponential family:

$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \quad (1)$$

where a, b, c are known functions, and θ, ϕ are parameters, called natural and dispersion parameters, respectively.

The systematic component relates some vector $(\eta_1, \ldots, \eta_n)$ to the p features.

$$\eta_i(\beta) = x_i^t, \quad \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad (2)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are called regression parameters.

The link function $g(\mu_i) = \eta_i = x_i^t\beta$ relates the linear predictor to the mean μ_i of y_i. If $\eta = \theta$, that is, if $\theta_i = \eta_i$, $\forall i$ holds. The link function is called the canonical link function.

The exponential family contains commonly used distributions such as gamma, normal, inverse Gaussian, Bernoulli, binomial, Poisson, geometric, negative binomial, and exponential.

In particular, a probability density function $f(y; \theta, \phi)$, characterized as a binomial distribution, where n is the number of trials, can be defined as

$$\begin{aligned} f(y;\theta,\phi) &= \binom{n}{y}\mu^y(1-\mu)^{n-y} \\ &= \exp\left[y\ln(\mu) + (n-y)\ln(1-\mu) + \ln\binom{n}{y}\right] \\ &= \exp\left[y\ln\left(\frac{\mu}{1-\mu}\right) + n\ln(1-\mu) + \ln\binom{n}{y}\right]. \end{aligned} \quad (3)$$

Therefore,

$$\begin{aligned} \theta &= \ln\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right), \\ \mathbf{b}(\theta) &= -\mathbf{n}\ln(1-\boldsymbol{\mu}) = \mathbf{n}\ln(1+\exp\boldsymbol{\theta}), \\ \mathbf{c}(\mathbf{y},\phi) &= \ln\binom{\mathbf{n}}{\mathbf{y}}. \end{aligned} \quad (4)$$

To evaluate the parameters of an exponential family, GLM maximum likelihood can be applied,

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta,\phi). \quad (5)$$

Therefore, log-likelihood for the sample $y_1, \ldots, y_n$ is

$$l(\theta) = \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi). \quad (6)$$

We use as link function g, a logit function. It returns values between 0 and 1 for any input,

$$g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right). \quad (7)$$

In order to maximize $l(\theta)$ over all choices of coefficients $\beta \in R^p$, it is necessary to consider that each natural parameter θ_i may be expressed using the mean μ_i of the exponential family distribution. Taking it into account, and recalling that a link function exists, such as

$$g(\mu_i) = \eta_i, \quad (8)$$

which joins the mean μ_i to the parameter $\eta_i = x_i^t\beta$. It is possible to compute β as in $\widehat{\beta}$ and then use these estimates to state that $g(\widehat{\mu}_i) = x_i^t\widehat{\beta} i = 1, \ldots, n$; $\widehat{\mu}_i = g^{-1}(x_i^t\widehat{\beta}), i = 1, \ldots, n$.

Therefore, it is possible to establish

$$l(\beta) = \sum_{i=1}^{n} y_i\theta_i - b(\theta_i), \quad (9)$$

where the terms that do not depend on $\theta_i, i = 1, 2, \ldots, n$, have been removed.

If the canonical link function g (8) considers

$$\theta_i = \eta_i = x_i^t\beta, \quad i = 1, \ldots, n, \quad (10)$$

$l(\beta)$ to maximize over β is

$$l(\beta) = \sum_{i=1}^{n} y_i x_i^t\beta - b x_i^t\beta. \quad (11)$$

In order to maximize $l(\beta)$ to form $\widehat{\beta}$, it is possible to carry out iteratively reweighted least squares regressions (IRLS) [40, 41]. Finally, the coefficients $\widehat{\beta}$ can be managed as a result of a single weighted least squares regression, the last one in the IRLS succession.

Specifically in this research, it is shown that the pattern of link formation in various PTNs can be well explained through a GLM. In this case, the response Y_i takes a categorical value, whether or not a link exists between two stops. The independent variables, $X_i = (x_{i1}, x_{ip})$, correspond to several indexes describing the similarity between stops. The probability density function $f(y; \theta, \phi)$ is characterized as a binomial distribution. The similarity indexes utilised as predictors are described in the labeled link building process in PTNs and the Supplementary materials section.

In order to check the importance of predictors using the t-test, it is required to examine if $\widehat{\beta}_j$ $\forall j$ is normally distributed. This is checked by applying the Anderson–Darling test [30] with a significance level $\alpha = 0.05$. The considered hypotheses are as follows:

(i) Null hypothesis H_0: "$\widehat{\beta}_j$ is normally distributed"

(ii) Alternative hypothesis H_a: "$\widehat{\beta}_j$ is not normally distributed"

If p – value $<\alpha$, H_0 is rejected, H_a is accepted. Else H_0 is taken.

The R package nortest was utilised for the calculation of the Anderson–Darling test.

Once it has been verified that $\widehat{\beta}_j$ $\forall j$ is normally distributed, t -tests [42] were carried out with a level of significance α. This allows us to know the contribution of each individual explanatory variable, X_{ij}, to the model. The possible hypotheses are as follows:

(i) Null hypothesis H_0: "explanatory variable X_{ij} has a slope that is equal to zero, that is, X_{ij} is not useful to predict Y_i, $\widehat{\beta}_j = 0$"

(ii) Alternative hypothesis H_a: "explanatory variable X_{ij} has a slope that is different from zero, that is, X_{ij} contribute to predict Y_i, $\widehat{\beta}_j \neq 0$

The results obtained in the test can be:

(i) If $p-$value $<\alpha$, H_0 is rejected, H_a is taken

(ii) Else H_0 is accepted, H_a is rejected

Next, the importance of the predictors is determined using a t statistic estimator, which is defined as the ratio of the estimated parameter $\widehat{\beta}_j$ to the standard error SE $(\widehat{\beta}_j)$ of the estimation,

$$t-\text{statistic } \widehat{\beta}_j = \frac{\widehat{\beta}_j}{\text{SE}\left(\widehat{\beta}_j\right)}. \tag{12}$$

For a given SE, the higher the value of the estimator, the higher value of the $t-$statistic.

If the null hypothesis is accepted, a high estimator produce evidence against it, similar to when the $t-$statistic is very far from the hypothesized value.

In order to implement the GLM model and to evaluate the importance of the predictors, the caret and vip packages in *R* are used.

2.2.2. Topological Representation of PTNs. A PTN can be represented in a topological space named L-Space in which a network is mapped as a graph $G = (N; L)$, where N is the set of nodes symbolizing the stops and L is the set of links established between them. In the L-Space, one node represents one stop, and one link means a union between two consecutive stops. This tells us that there is a link between two stops, if one stop is the successor of the other on a route.

2.2.3. Link Building Process in PTNs. In each network, it was analyzed whether a GLM could adequately describe the link formation process. As was explained in Section 2.2.1, the caret package in *R* was used in order to carry out the stages of training and validation of the model. The process was as follows.

The L-Space was constructed. All the loops and multiple links from the graph were deleted, obtaining a graph $G\prime$, where the maximal connected components were obtained. Then, with the largest cluster, the giant component (CG), the following operations were performed:

The number of pairs of connected and unconnected nodes were estimated, and several similarity measures were calculated for each one of them. Local, quasilocal, and global methods were applied.

The local similarity indexes used were: Adamic-Adar (dsimaa) [43], common neighbours (dsimcn), cosine (dsimcos) [44], cosine similarity on L+ (dsimcos_l) [45], hub promoted (dsimhpi) [46], jaccard (dsimjaccard) [47], hub depressed (dsimhdi) [3, 7], Leicht–Holme–Newman (dsimlhn_local) [48], preferential attachment (dsimpa) [49], and Sørensen (dsimsor) [50]. The global similarity measures used were: average commute time (dsimact) [37], normalized average commute time (dsimact_n) [51], Katz (dsimkatz) [52], L+ directly (dsiml) [45], matrix forest (dsimmf) [53], and random walk with restart (dsimrwr) [54]. Finally, the quasilocal measures of the similarity utilised were graph distance (dsimdis) and local path (dsimlp) [6, 55]. These indexes are described in detail in the Supplementary materials section.

The model has the values that describe the different similarities between pairs of nodes as input variables (features) and the indication of whether or not there is a link between them as output variable. In order to build the model, supervised learning is used. In this technique, the relations among the input variables (features) and outgoing ones (target) are learnt. That is, from some labeled examples (in each the correct input and output are known), the algorithm that is able to predict the value of the output for new cases not utilised in the learning (training process). For each PTN, a set of data is provided with different features, and the outcome or target (label) is known for each case (pair of nodes). The goal is to predict the label of new cases (pairs of nodes) with the minimum possible error. Since the outcome variable is a categorical value, whether or not a link exists, the prediction corresponds to a binary classification problem.

Crossvalidation is used as a procedure to estimate the model. Instead of splitting the dataset into a training and a test subset, in the crossvalidation mechanism, k equal partitions of the dataset are made. The model is trained k times: each time one of the partitions is taken as a test set, and the model is trained with the rest of the data (with the remaining $k-1$ folds). Each fold is used once as a test set. Finally, several predictions exist about the whole dataset. This process results in k estimates of a parameter related to the effectiveness of the model. An average of an estimated parameter (EP) can be made,

$$\langle \text{EP} \rangle = \frac{1}{k}\sum_{i=1}^{k} \text{EP}. \tag{13}$$

EP can be accuracy (14), area under the curve (AUC) [56], and kappa [57].

These parameters are described as follows:

TP: truth positives, TN: truth negatives, FP: false positives, and FN: false negatives.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{14}$$

AUC: AUC represents the probability that a classifier ranks a randomly selected positive instance higher than a randomly chosen negative instance. This EP can be defined, in general terms, as follows, given a binary classification task that has m positive and n negative instances, respectively. The outputs of a binary classifier can be considered as a rigorously ordered list for these instances, which can be appropriately represented by l_x, which is an indicator function of a set X. Therefore, c is a fixed classifier, where $y_{p1}, \ldots, y_{pm}$ are its outputs on the positive instances and $y_{n1}, \ldots, y_{mn}$ are its outputs on the negative instances. The AUC related to c is described [58] as

$$\mathrm{AUC} = \frac{\sum_{i=1}^{i=m} \sum_{j=1}^{n} l_{y_{pi} > y_{nj}}}{mn}, \tag{15}$$

which is the value of the Wilcoxon–Mann–Whitney statistic [59].

Kappa: this EP is defined as

$$\mathrm{Kappa} = \frac{p_0 - p_c}{1 - p_c}, \tag{16}$$

where p_0 = Accuracy and

$$p_c = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} + \frac{\mathrm{FP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} + \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} + \frac{\mathrm{TN} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{17}$$

Finally, an independent end estimation of the accuracy, recall, precision, and specificity of the model can be obtained using the validation set. The last three parameters are

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}},$$
$$\mathrm{specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{18}$$
$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

In addition, the confusion matrix as an estimation of the provided solution was obtained in the end validation for each PTN. Table 1 describes the confusion matrix general concept for a binary classification problem.

The final validation was performed on 20% of the total samples.

The selection of the similarity measures to be used as input variables to the model required checking the existing correlation between them. To determine whether this correlation should be estimated using Spearman's or Pearson's method, we checked whether the variables were normally distributed. The Anderson–Darling test [60] was applied with a significance level equal to 0.05. The following hypotheses were used:

(i) H_0: "the sample comes from a normal distribution"

(ii) H_a: "the sample does not come from a normal distribution"

If $p-\mathrm{value} < 0.05$, H_0 is rejected; otherwise, H_0 is accepted.

The *R* package nortest was utilised for the calculation of the Anderson–Darling test.

2.2.4. Study of the Efficiency. In a graph, G, the distance between the two nodes (i and j), $d(i, j)$, is the number of links that form the shortest path between them. If there is no link between i and j, then $d(i, j) = \infty$. The efficiency between i and j [60] can be defined as

$$\mathrm{Eff}_{ij} = \frac{1}{d(i, j)}, \quad \forall i \neq j. \tag{19}$$

Since Eff_{ij} is estimated based on the shortest path length between node pairs, an increase in $d(i, j)$ would result in a decrease in the local efficiency between i and j.

In addition, the global efficiency of G can be described as

$$\mathrm{GlobEff}(G) = \frac{1}{N(N-1)} \sum_{i \neq j} \mathrm{Eff}_{ij}. \tag{20}$$

This parameter is the average of the efficiencies calculated over all pairs of nodes in G. For a given number of nodes N, GlobEff (G) increases with the addition of links. According to the previous definition $0 \leq \mathrm{GlobEff}\ (G) \leq 1$, being the value 1 reached for a complete graph [61].

GlobEff (G) has been estimated in several PTNs as one of its features [62, 63]. This research analyses the impact that the elimination of links between pairs of nodes, with certain similarity characteristics, has on the GlobEff of the GC in $G\prime$. The result could help to achieve better network planning, since, depending on which links are removed or built, higher or lower GlobEff can be obtained. Common characteristics regarding efficiency in PTNs are also identified.

The relationship between GlobEff and network density is also analyzed. This last characteristic for undirected graphs such as PTNs can be defined as

$$\mathrm{density} = \frac{2 * \text{number of links in } G}{\text{number of nodes} * (\text{number of nodes} - 1)}. \tag{21}$$

2.2.5. Correlations between Topological Measurements. Certain investigations have been performed focusing on the study of centrality measures [35] in a PTN. In [36], the authors study some centralities in 58 existing social networks. Further studies examine the correlation between centrality metrics: using Pearson, Spearman, and Kendall methods [37]. The authors use the degree as the base to approximate three other metrics: closeness, betweenness, and eigenvector. They check the correlation between centrality metrics in several real networks, categorized as social, technological, and biological networks. Authors find that the betweenness occupies the highest coefficient, closeness is at the middle level, while eigenvector fluctuates dramatically between networks. They also put forward the idea that rank correlation performs better than the Pearson one in scale-free networks. In [40], several different real-world network graphs, representing several contexts (social club network, birds' social network, word adjacency network, airports network, games network, and

Table 1: Confusion matrix for a binary classification problem.

	Actual value (AV)	
Predicted value (PV)	AV 0 (no link exists)	AV 1 (A link exists)
PV 0 (no link exists)	Number of TN	Number of FP
PV 1 (A link exists)	Number of FN	Number of TP

related book network) with the number of nodes ranging from 34 to 332, were used. The authors classify the main centrality metrics into two categories: degree-based (degree and eigenvector centralities) and shortest path-based (betweenness, closeness, distance, and eccentricity centralities). They analyze the correlation between the aforementioned centrality metrics, showing that two degree-based centrality metrics (degree and eigenvector centrality) are highly correlated across all the studied networks. There is predominantly a moderate level of correlation between any two of the shortest path-based centrality metrics (betweenness, closeness, distance, and eccentricity). The authors explain that a poor correlation exists between a degree-based centrality metric and a shortest path-based centrality metric for regular random networks. As the variation in the degree distribution of the nodes increases, the correlation coefficient between the two classes of centrality metrics increases. Reference [34] uses a regression model to show a correlative relationship between passenger flow distribution and the conventional network properties (in/out degree, betweenness, and closeness) for the train system in Hague and Amsterdam cities.

Due to the classification, social, technological, and biological networks can encompass networks of very different types, and our investigation focuses on the study of centralities in PTNs. These correlations are studied in G'. Specifically, the following centralities are calculated:

(i) The degree of a node i, $k(i)$, for an undirected graph, G, such as a PTN, is [26, 64]

$$k(i) = \sum_{j=1}^{N} A_{ij}, \tag{22}$$

where

A_{ij} is the element ij of the adjacency matrix, A, such as $A_{ij} = 1$, if the node i is linked to node j and 0, otherwise.

(ii) The minimum distance between two nodes i, j in G, l, is the length of the shortest path between them.

(iii) The betweenness centrality of a node i in G, $\mathrm{BC}(i)$,is [26, 65]

$$\mathrm{BC}(i) = \sum_{u \neq i \neq w} \frac{\gamma_{u,w}(i)}{\gamma_{u,w}}, \tag{23}$$

where $\gamma_{u,w}$ is the total number of shortest paths from node u to node w, and $\gamma_{u,w}(i)$ is the number of those paths that pass through i.

(iv) Regarding the eigenvector centrality of a node i in G, $\mathrm{EC}(i)$ [26, 65, 66]: $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_N$ are the eigenvalues of the adjacency matrix $A = \{A_{ij}\}$ of G. Then, the largest eigenvalue of matrix A is $\lambda_{\max}$ with an eigenvector $e = [e_1, e_2, \ldots, e_N]^T$ such that $\lambda_{\max} * e_i = \sum_{j=1}^{N} A_{ij} * e_j$. The eigenvector centrality for node i represented as $\mathrm{EC}(i)$ can be defined as

$$\mathrm{EC}(i) = \frac{1}{\lambda_{\max}} \sum_{j=1}^{N} A_{ij} * e_j. \tag{24}$$

(v) Pagerank, PR, of a node i in G, is [26, 66–68]

$$\mathrm{PR}(i) = \frac{q}{N} + (1-q) \sum_{j: j \longrightarrow i} \frac{\mathrm{PR}(j)}{k_{\mathrm{out}}(j)}, \quad i = 1, 2, 3, \ldots, N, \tag{25}$$

where [26]

N is the number of nodes in G, $\mathrm{PR}(j)$ is the pagerank of a node j, and $k_{\mathrm{out}}(j)$ is the outdegree of node j, being the sum of $(\mathrm{PR}(j)/k_{\mathrm{out}}(j))$ executed over the nodes pointing towards i. In the case of the PTNs, it is considered that G is an undirected graph; therefore, $k_{\mathrm{out}}(j) = k(j)$.

q is the damping parameter, $\in [0, 1]$.

(vi) A hub is a node that points to many relevant nodes, and an authority node is the one that is focused on by many important nodes. Both are based on the eigenvectors related to the highest eigenvalues of the matrices AA^T and A^TA.

The hub centrality of the node i, denoted by HC (i), is the i-th entry of the following vector y satisfying equation:

$$AA^T y = \lambda y, \quad \text{where } \lambda \in R \text{ is the highest eigenvalue of } AA^T. \tag{26}$$

Similarly, the authority of a node i, symbolized by AC (i), is the i-th entry of the following vector x satisfying equation:

$$A^T Ax = \lambda x, \quad \text{where } \lambda \in R \text{ is the highest eigenvalue of } A^TA. \tag{27}$$

For an undirected graph, such as a PTN, the adjacency matrix A is symmetric. The two scores, AC(i) and HC(i), are identical.

3. Results and Discussion

3.1. Link Building Process in PTNs. As was previously displayed in 2.2.2, the network was represented in the *L-Space*. All loops and multiple links were eliminated, obtaining graph $G\prime$. This is where we calculate the existing maximum number of connected components. Table 2 contains information collected after the explained process, for all analysed networks, the number of links and existing nodes and clusters in $G\prime$. In addition, there are the number of nodes and links present in the largest cluster GC. As well as the fact some of them have several clusters, detection of clusters in cities over PTNs can also allow us to find urban groups, which are strongly connected through transportation. The comparison between PTN clusters and urban agglomerations can be used to estimate whether the PTNs are capable of supporting these human distributions [69]. Identifying under- and overserviced areas can also help in policy decisions, including infrastructure planning and local development [70].

Table 2: Number of nodes, links, clusters, and characteristics of the GC in G' for all analyzed networks.

Network	Number of nodes	Number of links	Number of clusters	GC	
				Number of nodes	Number of links
AVL, CFL, RGTR, TICE	1328	1924	3	1370	1921
Island Transit	358	420	2	271	313
Lanta	2150	2330	1	2150	2330
Linja-Karjala Oy	534	700	1	534	700
Metlink	3007	3583	3	2998	3574
PPTC, ROPIT	5152	6757	18	4985	6599
Sage		36	1	31	36
STAR	1415	1993	2	1386	1965
Thunder Bay Transit	818	885	6	813	885
TransAntofagasta	645	962	1	645	962

As was explained in 2.2.1, we used the caret package in R for the building of the model. As described in 2.2.3, the model was trained k times: each time one of the partitions was taken as a test set, and the model was trained with the rest of the data (with the remaining $k-1$ folds). Each fold was used once as a test set. Finally, several predictions exist about the whole dataset. This process results in k estimates of the *accuracy*, *AUC*, and kappa parameters. Additionally, if two similarity measures had a correlation greater than 0.9, one of them was not considered in the prediction. Table 3 shows the similarity indexes that present a Spearman correlation higher than 0.9 with another.

In order to know the method to be used for the calculation of correlations, Pearson or Spearman, the Anderson–Darling test was applied with a significance level $\alpha = 0.05$. All networks showed a $p-$value <0.05. Therefore, the null hypothesis, H_0 was rejected, inferring that the distributions did not follow a normal pattern. Spearman's method was used to calculate correlations.

The importance of each predictor in the model was estimated calculating the absolute value of the $t-$statistics [71], whose definition has been presented in 2.2.1 The importance of predictors is shown in Table 4.

Tables 5 and 6 show, in each PTN, the average of the estimators (accuracy, AUC ,and kappa) calculated over the k times that the model was trained. Since the number of links between pairs of nodes was much lower than the number of unconnected pairs of nodes, the down-sampling approach was utilised, randomly removing the observations. In order to improve the results, artificial balanced samples were generated according to a smoothed bootstrap procedure [60] in the Thunder Bay Transit network. The rose package in R was used.

Table 7 shows, in each network, the confusion matrix [72] obtained in the final validation. In Table 8, accuracy, recall, precision, and specificity parameters are presented.

All networks showed good results applying down-sampling, according to the parameters chosen for the evaluation of the model. In the crossvalidation process, average accuracy and AUC values were higher than 0.99 and kappa larger than 0.93. In the validation stage, accuracy and recall showed values higher than 0.99, and specificity had a value equal to 1. The only exception was the Thunder Bay Transit network, where it was necessary to apply the rose method in order to achieve better kappa and precision values.

As a result, the process of building links was appropriately modeled using a GLM, which had some measures of similarities between nodes as input variables. The response variable, which establishes the existence or not of a link between pairs of nodes, is appropriately described by a binomial probability density function. The link function used is the logit function, as we explained in 2.2.1. The model has the novelties described in Section 1, with respect to other models that have already been developed for PTNs.

In most networks, the figure with the highest influence was dsimdis, followed by simact. In addition, the simcos_l and simlp showed high or moderate importance in some networks.

3.2. Study of Trip Times. The trip times are analyzed in order to estimate things in common between networks. Several statistical parameters are calculated (average, standard deviation, median, moda, maximum, and minimum values). The results and the frequency distribution are displayed in Table 9 and Figure 1, respectively.

The cumulative probability distributions are also checked. They are shown in Figure 2. The stats package in R was used. The similarity between two distributions is examined, applying the Kolmogorov–Smirnov test [73]. A significance level equal to 0.05 is taken, while the following hypotheses are considered:

(i) Null hypothesis (H_0): "the samples come from the same distribution."

(ii) Alternative hypothesis (H_a):"the samples come from different distributions."

If a $p-$value <0.05 is obtained in the test, the null hypothesis is rejected. Table S.1 shows the results obtained in the test.

It can be noted that similarities do not exist between the PTNs in relation to the trip times. All networks presented a high standard deviation. The lowest is 14.02 minutes (790.23523 seconds) and the highest is 11.12 hours (42,027.19610 seconds). This shows that the size, the complexity, and variability of available routes in the PTNs cause trip times to be highly inconsistent between routes. Trip times allow the evaluation of how travelers choose a service based on whether or not it is convenient. Trip times have

Table 3: Similarity measures that present a Spearman correlation higher than 0.9 with another.

Network	Similarity measures
AVL, CFL, RGTR, TICE	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsiml, dsimkatz, dsimmf, dsimrwr.
Island Transit	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsimact_n, dsimdis, dsimkatz, dsimmf, dsimrwr, dsiml
Lanta	dsimcn,d simcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsimact_n, dsiml, dsimkatz, dsimmf, dsimrwr
Linja-Karjala Oy	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsiml, dsimkatz, dsimmf, dsimrwr
Metlink	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsimact_n, dsiml, dsimkatz, dsimmf, dsimrwr
PPTC, ROPIT	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsimact_n, dsiml, dsimkatz, dsimmf, dsimrwr
Sage	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsimact_n, dsimdis, dsimkatz, dsimmf, dsimrwr Dsiml
STAR	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsiml, dsimkatz, dsimmf, dsimrwr
Thunder Bay Transit	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard, dsimlhn_local, dsimsor, dsimact_n, dsiml, dsimkatz, dsimmf, dsimrwr
TransAntofagasta	dsimcn, dsimcos, dsimhdi, dsimhpi, dsimjaccard,d simlhn_local, dsimsor, dsiml, dsimkatz, dsimmf, dsimrwr

Table 4: Importance of predictors.

Network	Similarity measures	Importance	Network	Similarity measures	Importance
AVL, CFL, RGTR, TICE	Dsimdis	100	Island Transit	Dsimact	100
	Dsimlp	13.20776		dsimcos_l	53.78080
	Dsimpa	11.61152		Dsimlp	7.17716
	Dsimaa	10.37026		Dsimpa	6.75717
	dsimcos_l	9.554070		Dsimaa	0
	dsimact_n	4.423947			
	Dsimact	0			
Lanta	dsimcos_l	100	Linja-Karjala Oy	Dsimdis	100
	Dsimdis	94.94621		Dsimlp	25.23217
	Dsimpa	10.94917		Dsimact	23.95510
	Dsimlp	7.372249		Dsimaa	17.12581
	Dsimaa	6.737758		Dsimpa	12.22195
	Dsimact	0		dsimact_n	4.047694
				dsimcos_l	0
Metlink	Dsimdis	100	PPTC, ROPIT	Dsimdis	100
	dsimcos_l	48.71324		Dsimlp	19.90804
	Dsimaa	10.70750		dsimcos_l	14.18266
	Dsimpa	7.501585		Dsimact	11.82702
	Dsimact	2.410628		Dsimaa	3.308090
	Dsimlp	0		dsimact_n	1.716437
				Dsimpa	0
Sage	Dsimact	100	STAR	Dsimdis	100
	dsimcos_l	26.16491		Dsimlp	9.89944
	Dsimlp	16.60197		Dsimact	7.32344
	Dsimpa	11.17248		Dsimaa	5.10874
	Dsimaa	4.439547		Dsimpa	3.83377
	Dsimhpi	0		dsimcos_l	0.10442
				dsimact_n	0
Thunder Bay Transit	Dsimdis	100	TransAntofagasta	Dsimdis	100
	dsimcos_l	40.08267		Dsimact	55.94523
	Dsimaa	38.62297		Dsimlp	48.88970
	Dsimlp	34.96505		dsimcos_l	41.40694
	Dsimpa	24.13330		dsimact_n	28.06088
	Dsimact	0		Dsimaa	27.07529
				Dsimpa	0

Table 5: In AVL, CFL, RGTR, TICE, Island Transit, Lanta, Linja-Karjala Oy, Metlink networks, the average estimation of accuracy, AUC, and kappa calculated over the k times in which the model was trained.

Network	Training approach	Accuracy	AUC	Kappa
AVL, CFL, RGTR, TICE	Downsampling	0.99996	1	0.98208
Island Transit	Downsampling	0.99986	1	0.98406
Lanta	Downsampling	1	1	1
Linja-Karjala Oy	Downsampling	1	1	1
Metlink	Downsampling	0.99994	1	0.93209

Table 6: In PPTC, ROPIT, Sage, STAR, Thunder Bay Transit, TransAntofagasta networks, the average estimation of accuracy, AUC, and kappa calculated over the k times in which the model was trained.

Network	Training approach	Accuracy	AUC	Kappa
PPTC, ROPIT	Downsampling	1	1	0.99886
Sage		1	1	1
STAR	Downsampling	1	1	1
Thunder bay Transit	Downsampling	0.99933	0.99960	0.79877
	Smoothed bootstrap	0.99999	1	0.99718
TransAntofagasta	Downsampling	1	1	1

Table 7: Final validation. Confusion matrix.

Network	Training approach	Confusion matrix		
			Actual value	
		Predicted value	AV 0	AV 1
AVL, CFL, RGTR, TICE	Downsampling	PV 0	349548	0
		PV 1	14	384
			Actual value	
		Predicted value	AV 0	AV 1
Island Transit	Downsampling	PV 0	14506	0
		PV 1	2	62
			Actual value	
		Predicted value	AV 0	AV 1
Lanta	Downsampling	PV 0	923138	0
		PV 1	0	466
			Actual value	
		Predicted value	AV 0	AV 1
Linja-Karjala Oy	Downsampling	PV 0	56644	0
		PV 1	0	140
			Actual value	
		Predicted value	AV 0	AV 1
Metlink	Downsampling	PV 0	1795467	0
		PV 1	104	714
			Actual value	
		Predicted value	AV 0	AV 1
PPTC, ROPIT	Downsampling	PV 0	4966405	0
		PV 1	3	1319
			Actual value	
		Predicted value	AV 0	AV 1
STAR	Downsampling	PV 0	383136	0
		PV 1	0	393
			Actual value	
		Predicted value	AV 0	AV 1
	Downsampling	PV 0	131588	0
Thunder Bay Transit		PV 1	89	177
		Predicted value	AV 0	AV 1
	Smoothed bootstrap	PV 0	131676	0
		PV 1	1	177

Table 7: Continued.

Network	Training approach	Confusion matrix		
			Actual value	
		Predicted value	AV 0	AV 1
TransAntofagasta	Downsampling	PV 0	82691	0
		PV 1	0	192
			Actual value	
		Predicted value	AV 0	AV 1
Sage	Downsampling	PV 0	171	0
		PV 1	0	7

Table 8: Final validation. Accuracy, recall, precision, and specificity parameters.

Network	Training approach	Accuracy	Recall	Precision	Specificity
AVL, CFL, RGTR, TICE	Downsampling	0.99999	1	0.96482	1
Island Transit	Downsampling	0.99986	1	0.96875	1
Lanta	Downsampling	1	1	1	1
Linja-Karjala Oy	Downsampling	1	1	1	1
Metlink	Downsampling	0.9999421	1	0.87286	1
PPTC, ROPIT	Downsampling	1	1	0.99773	1
STAR	Downsampling	1	1	1	1
Thunder Bay Transit	Downsampling	0.99933	1	0.66541	1
	Smoothed bootstrap	0.99999	1	0.99438	1
TransAntofagasta		1	1	1	1
Sage		1	1	1	1

Table 9: Trip time metrics (seconds).

Network	Average	Standard deviation	Median	Max	Moda	Min
AVL, CFL, RGTR, TICE	1,731.33538	790.23523	1800	4020	2280	60
Island Transit	5,385.63025	1,298.01961	2100	69000	2400	120
Lanta	3,532.24305	1,505.97579	3360	10140	1800	420
Linja-Karjala Oy	2,071.65844	841.08391	2040	4140	1260	60
Metlink	2,111.43848	1,293.83504	1980	43800	1500	240
PPTC, ROPIT	1,911.43488	1,087.87538	1820	30180	960	60
Sage	12,000	7,842.19357	11850	24000	1200	1200
STAR	43,533.94450	42,027.19610	83460	86340	84960	60
Thunder bay Transit	1,751.91089	892.014494	1380	4500	1200	300
TransAntofagasta	10,297.08240	1297.94464	10254	13244	9827	5999

been considered by some researchers to evaluate the performance of PTNs [74, 75].

3.3. Study of Efficiency

3.3.1. Local Efficiency. All networks showed a large majority of nodes with low local efficiency ≤0.20, as can be noted in Figures 3 and 4.

As was done with trip times, the similarity between local efficiency distributions is examined, applying the Kolmogorov–Smirnov test. A significance level equal to 0.05 is taken, resulting in the following hypotheses being considered:

(i) Null hypothesis (H_0): "the samples come from the same distribution."

(ii) Alternative hypothesis (H_a): "the samples come from different distributions."

If the p – value obtained in the test is <0.05, the null hypothesis is rejected.

The networks presented high analogies in the cumulative distributions of local efficiency. The test yielded a p – value >0.05 in all pairwise comparisons performed, as can be appreciated in Table S.2. Therefore, in general, if a stop is unavailable, the remaining connections between its neighbours are distinct from direct connections. This is revealed by the low value of the local efficiency [76].

3.3.2. Global Efficiency. The calculation of the GlobEff was carried out in the GC of G_1, and it can be observed, according to the results depicted in Table 10, that the higher the density of G_1, the higher the GlobEff.

Most of the analyzed networks presented a GlobEff of small value (<0.20). Some pieces of research use the GlobEff as a parameter to compare PTNs [77, 78], and others apply it to identify hubs [79, 80]. Consequently, the degree of a node

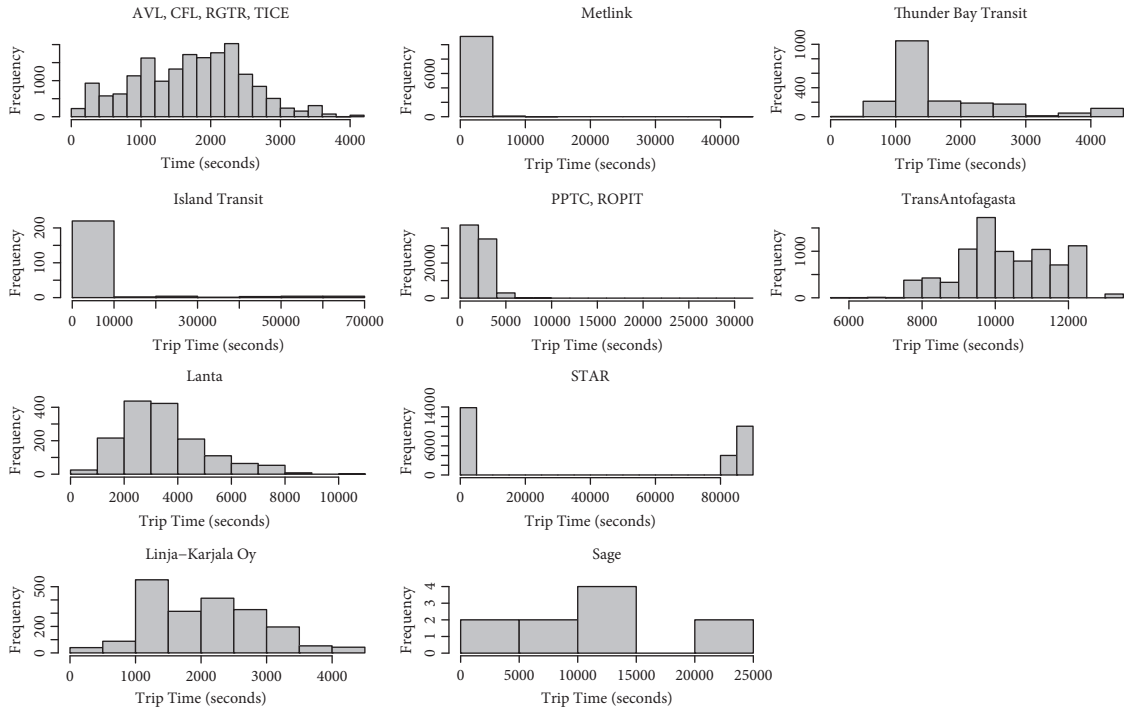

Figure 1: Histogram of trip time.

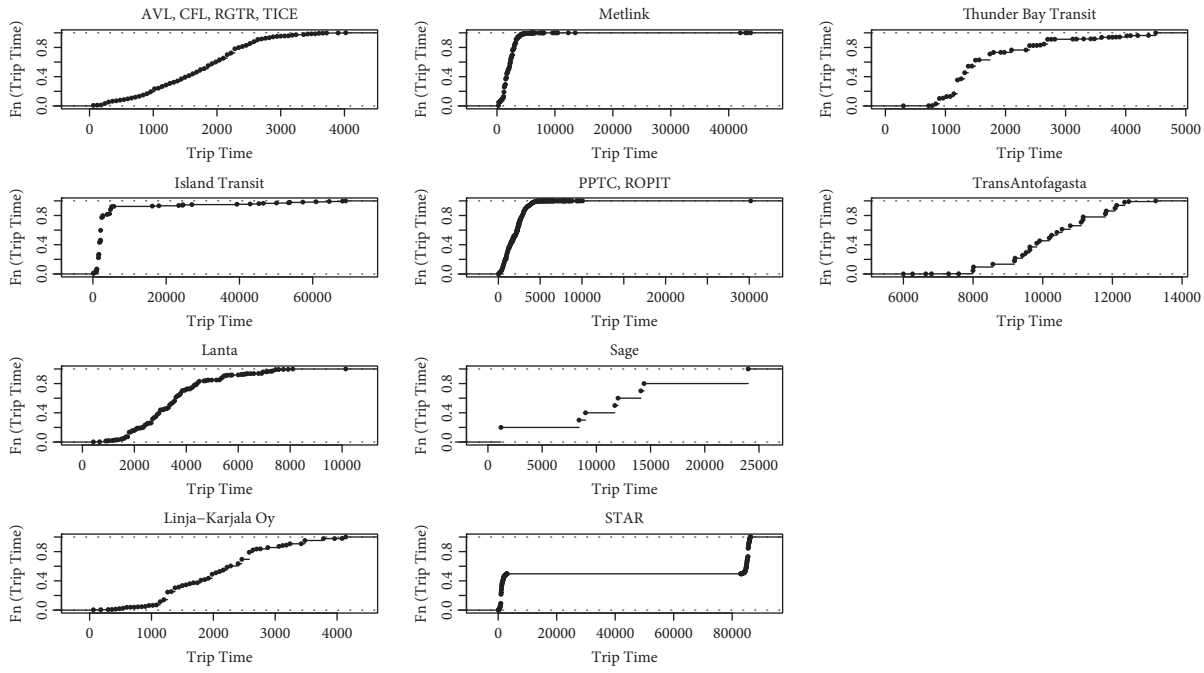

Figure 2: Cumulative probability distribution of trip times.

is ranked by comparing the changes in PTN efficiency after eliminating the node. In contrast, this research analyses the variation in GlobEff when links with certain similarity characteristics were removed. The results are shown in Table 11. Similarity measures with a correlation higher than 0.9 with another were not considered. It can be noted that in most of the networks, the link deletion in which a 75% reduction was reached most quickly was dsimpa and dsimlp, and the one that took the longest to reach was dsimcos_l. Figures 5–7 show the variation in GlobEff when certain links are removed.

Table 11 shows, for each similarity measure, the number of removed links that causes the reduction of GlobEff by 75%.

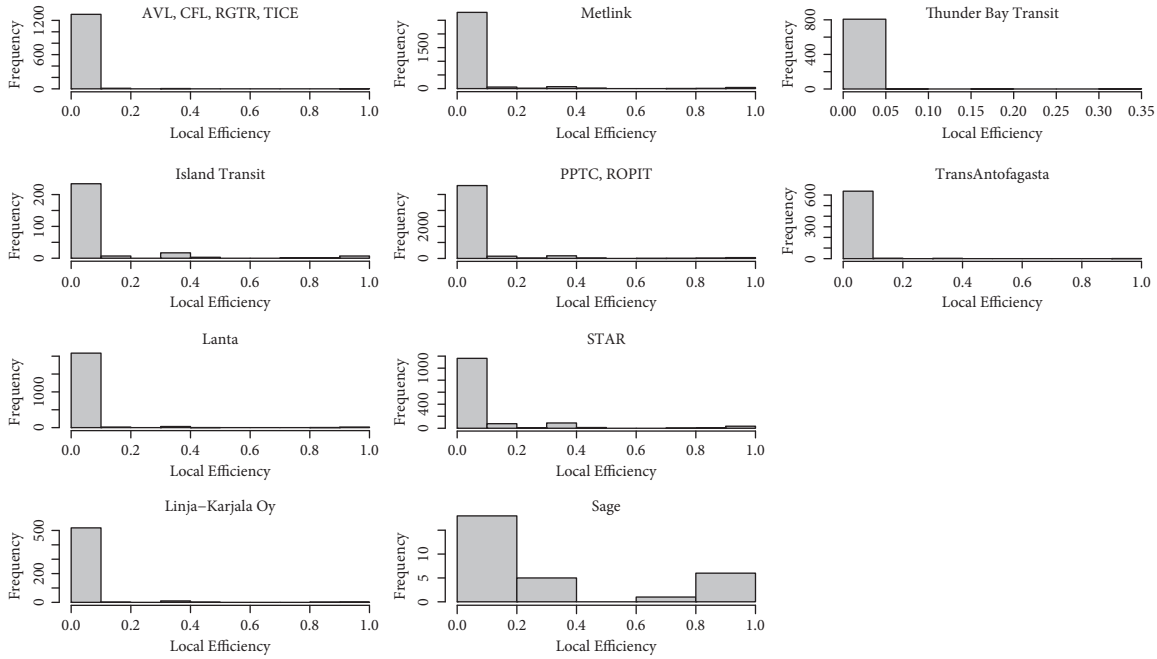

Figure 3: Histogram of local efficiency.

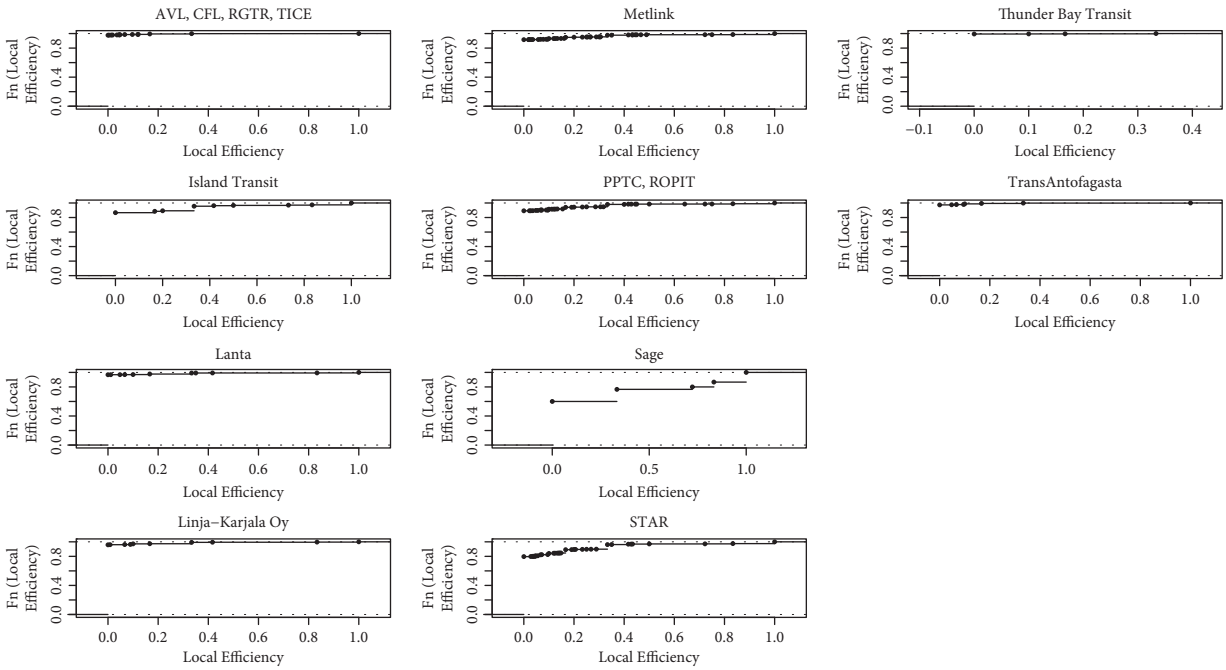

Figure 4: Cumulative probability distribution of local efficiency.

3.3.3. Correlations between Topological Measurements. The eigenvector, betweenness, pagerank, degree, hub, and authority centralities were calculated in $G\prime$, in order to study the correlation between them. The correlation of these variables with the amount of transport arriving and departing weekly from a stop were also estimated. Enabling us to know which method, Pearson or Spearman, should be used in the calculation, the Anderson–Darling test with a significance level $\alpha = 0.05$ was applied. In this way, it could be known whether or not the variables were normally distributed. The test yielded a p – value <0.05 for all variables, so the null hypothesis H_0 was rejected, and the alternative hypothesis H_a was accepted.

The correlations obtained by applying Spearman's method are shown in Tables S.3–S.12. In all networks, the eigenvector centrality presented a strong correlation with hub and authority centralities. Pagerank showed a moderate, high, or very high correlation with the degree. Therefore,

Table 10: GlobEff and density in GC in G' for all analyzed networks.

Network	GlobEff (GC in G')	Density
AVL, CFL, RGTR, TICE	0.14604	0.00219
Island Transit	0.07901	0.00856
Lanta	0.03248	0.00101
Linja-Karjala Oy	0.15180	0.00492
Metlink	0.05731	0.00080
PPTC, ROPIT	0.05291	0.00053
Sage	0.24383	0.07742
STAR	0.11220	0.00205
Thunder Bay Transit	0.07584	0.00268
TransAntofagasta	0.16649	0.00463

Table 11: Number of removed links that cause a 75% of reduction in the GlobEff.

Network	Similarity measures	Number of removed links	Network	Similarity measures	Number of removed links
AVL, CFL, RGTR, TICE	Dsimpa	655	Island transit	dsimpa	80
	Dsimlp	697		dsimlp	83
	Dsimact	702		dsimaa	110
	Dsimdis	890		dsimact	123
	Dsimaa	892		dsimcos_l	195
	dsimact_n	1052			
	dsimcos_l	1084			
Lanta	Dsimpa	177	Linja-Karjala Oy	dsimpa	181
	Dsimlp	203		dsimlp	206
	Dsimact	358		dsimact	221
	Dsimaa	549		dsimdis	302
	Dsimdis	616		dsimaa	303
	dsimcos_l	1450		dsimact_n	395
				dsimcos_l	399
Metlink	Dsimpa	1,193	PPTC, ROPIT	dsimpa	1500
	Dsimlp	1,217		dsimlp	1730
	Dsimact	1,718		dsimact	1850
	dsimcos_l	2,202		dsimaa	2550
	Dsimaa	2,493		dsimdis	2550
	Dsimdis	2,538		dsimcos_l	3750
Sage	Dsimlp	20	STAR	dsimact	545
	Dsimpa	20		dsimpa	640
	Dsimaa	21		dsimlp	714
	Dsimact	21		dsimdis	1,050
	dsimcos_l	24		dsimaa	1,053
				dsimact_n	1,172
				dsimcos_l	1,198
Thunder Bay Transit	Dsimpa	89	TransAntofagasta	dsimlp	386
	Dsimlp	93		dsimact	398
	Dsimact	118		dsimaa	512
	Dsimdis	327		dsimact_n	535
	Dsimaa	393		dsimdis	550
	dsimcos_l	422		dsimcos_l	553

also in this network, a high degree usually has a significant influence. The pagerank and degree only presented a moderate or high correlation with betweenness in some networks, demonstrating that specifically in these few networks a node with a high degree also usually presents an important level of connectivity. Eigenvector and degree, in most networks, exhibited a low or very low correlation. Furthermore, the number of weekly buses arriving and departing from a bus stop showed no strong correlation with any of the centrality measures. Strong correlations between

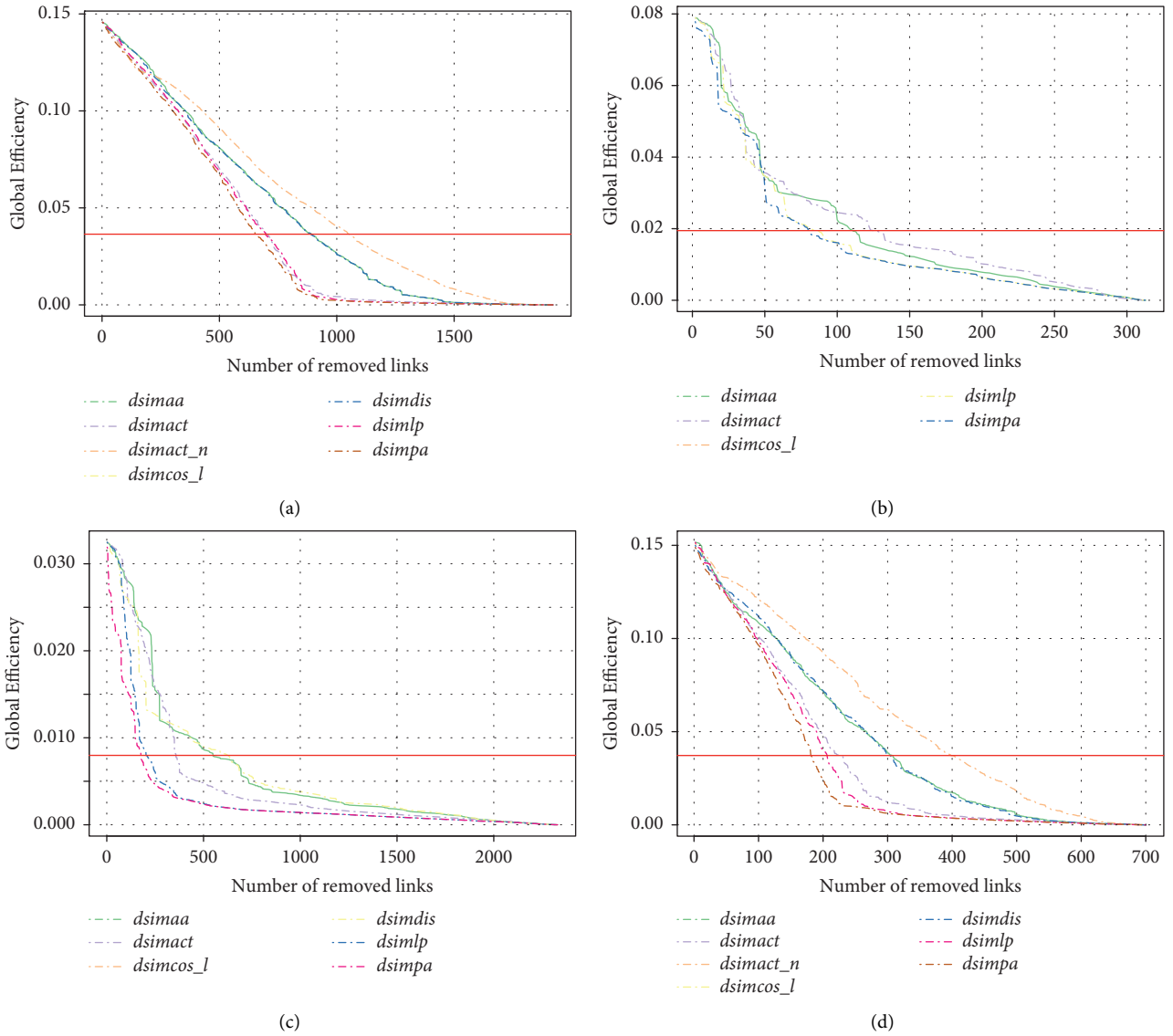

Figure 5: Variation in global efficiency when links with certain similarity characteristics are removed in AVL, CFL, RGTR, and TICE (a), Island Transit (b), Lanta (c), and Linja-Karjala Oy (d) networks.

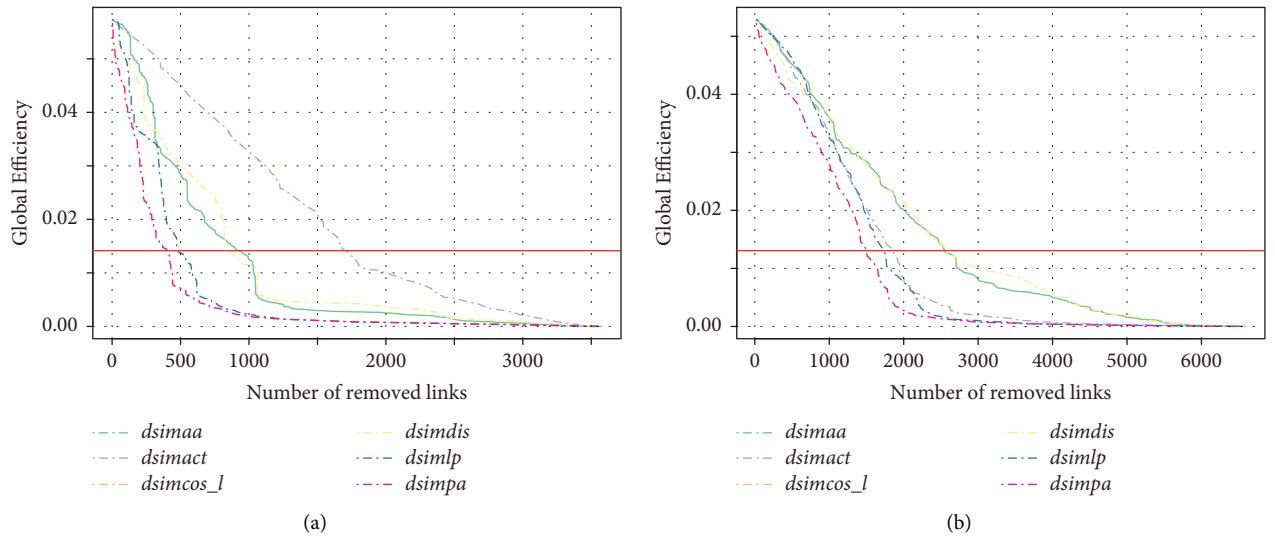

Figure 6: Continued.

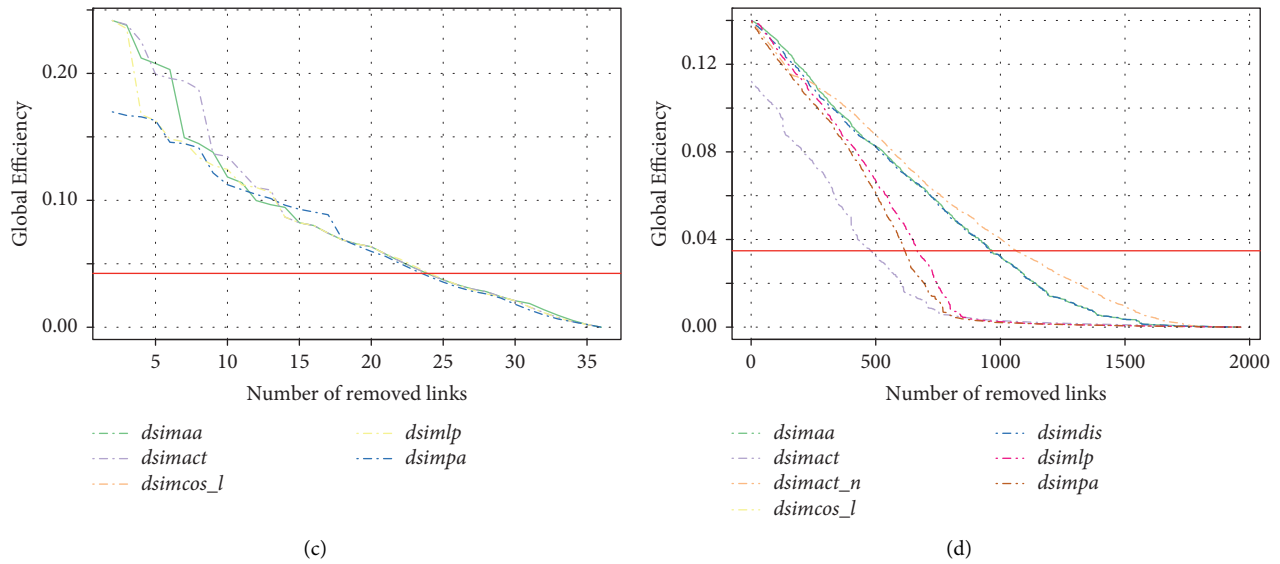

Figure 6: Variation in global efficiency when links with certain similarity characteristics are removed in Metlink (a), PPTC, ROPIT (b), Sage (c), and STAR (d) networks.

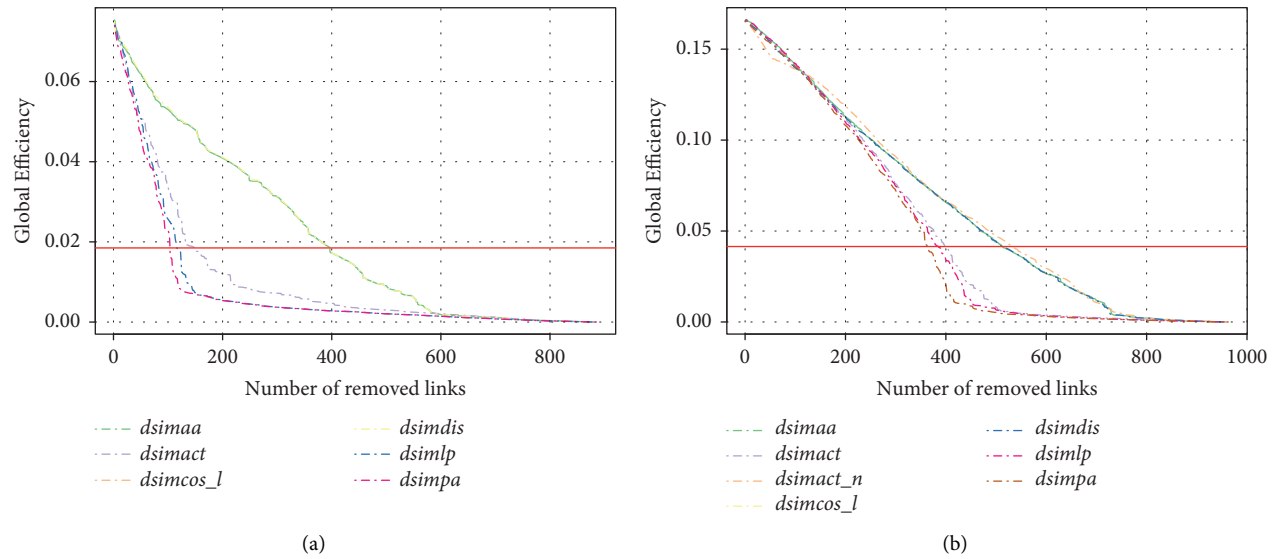

Figure 7: Variation in global efficiency when links with certain similarity characteristics are removed in Thunder Bay Transit (a) and TransAntofagasta (b) networks.

degree and pagerank and degree and betweenness have also been found in some Chinese PTNs [78].

4. Conclusions

Regarding the model followed by the formation of links between stops, this research shows that it can be correctly explained through a generalized linear model, which has certain similarity measures as input variables. Although the similarity measures that explain the model are different among networks, in most of them, dsimdis has a higher significance. It has a value equal to 100. In addition, dsimcos_l and dsimlp presented relevant importance in some PTNs with values higher than 30. Additionally, dsimact and dsimpa showed values equal to 100 and larger than 10, respectively, in certain PTNs.

Regarding travel times, these showed a high variability between networks (with standard deviations greater than 790.23 seconds), as well as very different cumulative probability distributions ($p-$value ≥ 0.05 in Kolmogorov–Smirnov test).

The study of local efficiency reveals that its cumulative distributions have strong analogies in all network distributions (Kolmogorov–Smirnov test showed $p-$values

<0.05). The local efficiency showed values ≤0.2 in the most of PTNs. Similarly, the overall efficiency exhibited reduced values (≤0.25). This seems to be a common feature of PTNs.

With respect to the centrality measures, they did not show correlation with the flow of vehicles, suggesting that traffic dynamics in the network may be strongly influenced by other different parameters as opposed to topological ones. In all networks, strong correlations of the eigenvector centrality with the hub and authority centralities were detected (with values higher than 0.80). The pagerank showed moderate, high, or very high correlation with the degree (it was larger than 0.5 in all networks). Therefore, these correlation characteristics seem to be a commonality in PTNs.

This research can be continued with a detailed study on the interactions between the different existing modes of transport modes in the cities. A multimodal transportation system, embodied as a multiplex network, can be considered in order to face the problem of urban mobility. In a multiplex network, a node symbolizes a specific origin/destination stop, which exists in each of the network layers. Nevertheless, the links are represented by a different layer of interaction determined by the type of transportation mode used for connecting two nodes.

Data Availability

Information of stops, routes and trip times of AVL, CFL, RGTR, and TICE; Island Transit; Lanta; Linja-Karjala Oy; Metlink; PPTC, ROPIT; Sage; STAR; Thunder Bay Transit; and TransAntofagasta were retrieved from the operating companies' public web sites, the Deconet Public Transport Network Data, and GTFS Data Exchange repositories.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially funded by Telefonica Chair at Francisco de Vitoria University.

Supplementary Materials

Supplementary Material includes (i) description of similarity measures (local, global, and quasilocal methods), (ii) tables related to the study of the trip times, (iii) tables regarding analysis of the local efficiency, and (iv) tables related to correlations between centrality measures. (*Supplementary Materials*)

References

[1] I. Ahmad, M. U. Akhtar, S. Noor, and A. Shahnaz, "Missing link prediction using common neighbor and centrality based parameterized algorithm," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.

[2] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, vol. 49, no. 4, pp. 1–33, 2017.

[3] L. Lu and T. Zhou, "Link prediction in complex networks: a survey," *Physica Statistical Mechanics and Its Applications A*, vol. 390, no. 6, pp. 1150–1170, 2011.

[4] D. Malhotra and R. Goyal, "Link prediction in complex networks using information-theoretic measures," *Journal of Complex Networks*, vol. 8, no. 4, 2020.

[5] F. Tan, Y. Xia, and B. Zhu, "Link prediction in complex networks: a mutual information perspective," *PLoS One*, vol. 9, Article ID e107056, 2014.

[6] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[7] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[8] F. Gao, K. Musial, C. Cooper, and S. Tsoka, "Link prediction methods and their accuracy for different social networks and network metrics," *Scientific Programming*, vol. 2015, Article ID 172879, 13 pages, 2009.

[9] B. Zhu and Y. Xia, "Link prediction in weighted networks: a weighted mutual information model," *PLoS One*, vol. 11, pp. 1–13, 2016.

[10] Y. Yang, J. Zhang, X. Zhu, J. Ma, and X. Su, "Link prediction based on the tie connection strength of common neighbor," *International Journal of Modern Physics C*, vol. 30, no. 11, Article ID 1950089, 2019.

[11] S. Bai, L. Li, J. Cheng, S. Xu, and X. Chen, "Predicting missing links based on a new triangle structure," *Complexity*, vol. 2018, Article ID 7312603, 11 pages, 2018.

[12] A. Mohan, R. Venkatesan, and K. V. Pramod, "A scalable method for link prediction in large real world networks," *Journal of Parallel and Distributed Computing*, vol. 109, pp. 89–101, 2017.

[13] S. Gorlatch, U. Banerjee, R. De Nicola et al., "BSP (Bulk synchronous parallelism)," in *Encyclopedia of Parallel Computing*, D. Padua, Ed., Springer, Boston, MA, USA, pp. 192–199, 2011.

[14] A. M. Abdolhosseini-Qomi, S. H. Jafari, A. Taghizadeh, N. Yazdani, M. Asadpour, and M. Rahgozar, "Link prediction in real-world multiplex networks via layer reconstruction method," *Royal Society Open Science*, vol. 7, no. 7, Article ID 191928, 2020.

[15] N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, "Applications of link prediction in social networks: a review," *Journal of Network and Computer Applications*, vol. 166, Article ID 102716, 2020.

[16] T. Wang, X.-S. He, M.-Y. Zhou, and Z.-Q. Fu, "Link prediction in evolving networks based on popularity of nodes," *Scientific Reports*, vol. 7, no. 1, 2017.

[17] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, "An evolutionary algorithm approach to link prediction in dynamic social networks," *Journal of Computational Science*, vol. 5, no. 5, pp. 750–764, 2014.

[18] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[19] N. Hansen, "The CMA evolution strategy: a comparing review," in *Towards a New Evolutionary Computation. Studies in Fuzziness and Soft Computing*, J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, Eds., vol. 192, pp. 75–102, Springer, Berlin, Germany, 2006.

[20] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[21] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, Article ID 046122, 2009.

[22] C. Lei and J. Ruan, "A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity," *Bioinformatics*, vol. 29, no. 3, pp. 355–364, 2013.

[23] X. Li, "Research on passenger flow forecast of urban rail transit," *E3S Web of Conferences*, vol. 213, Article ID 03026, 2020.

[24] G. Yeboah, C. D. Cottrill, J. D. Nelson, D. Corsar, M. Markovic, and P. Edwards, "Understanding factors influencing public transport passengers' pre-travel information-seeking behaviour," *Public Transport*, vol. 11, no. 1, pp. 135–158, 2019.

[25] C. von Ferber, T. Holovatch, Y. Holovatch, and V. Palchykov, "Public transport networks: empirical analysis and modeling," *The European Physical Journal B*, vol. 68, no. 2, pp. 261–275, 2009.

[26] M. L. Mouronte-López, "Analysing the vulnerability of public transport networks," *Journal of Advanced Transportation*, vol. 2021, Article ID 5513311, 22 pages, 2021.

[27] D. Levinson and W. Chen, "Area-based models of highway growth," *Journal of Urban Planning and Development*, vol. 133, no. 4, pp. 250–254, 2007.

[28] O. Cats, A. Vermeulen, M. Warnier, and H. van Lint, "Modelling growth principles of metropolitan public transport networks," *Journal of Transport Geography*, vol. 82, Article ID 102567, 2020.

[29] M. L. Mouronte and R. M. Benito, "Structural properties of urban bus and subway networks of Madrid," *Networks and Heterogeneous Media*, vol. 7, no. 3, pp. 415–428, 2012.

[30] B. P. V. Samson, G. A. T. Velez, J. R. Nobleza, D. Sanchez, and J. T. Milan, "Optimizing the efficiency, vulnerability and robustness of road-based para-transit networks using genetic algorithm," *Lecture Notes in Computer Science*, vol. 2018, pp. 3–14, 2018.

[31] M. Zanin, X. Sun, and S. Wandelt, "Studying the topology of transportation systems through complex networks: handle with care," *Journal of Advanced Transportation*, vol. 2018, Article ID 3156137, 17 pages, 2018.

[32] E. Rodríguez-Núñez and J. C. García-Palomares, "Measuring the vulnerability of public transport networks," *Journal of Transport Geography*, vol. 35, pp. 50–63, 2014.

[33] B. Berche, C. von Ferber, T. Holovatch, and Y. Holovatch, "Resilience of public transport networks against attacks," *The European Physical Journal B*, vol. 71, no. 1, pp. 125–137, 2009.

[34] D. Luo, O. Cats, and H. van Lint, "Can passenger flow distribution be estimated solely based on network properties in public transport systems?" *Transportation*, vol. 47, no. 6, pp. 2757–2776, 2020.

[35] M. L. Mouronte and R. M. Benito, "Structural analysis and traffic flow in the transport networks of Madrid," *Networks and Heterogeneous Media*, vol. 10, no. 1, pp. 127–148, 2015.

[36] T. W. Valente, K. Coronges, C. Lakon, and E. Costenbader, "How correlated are network centrality measures?" *Connections*, vol. 28, no. 1, pp. 16–26, 2008.

[37] E. Costenbader and T. W. Valente, "The stability of centrality measures when networks are sampled," *Social Networks*, vol. 25, no. 4, pp. 283–307, 2003.

[38] R Project for Statistical Computing, 2020, https://www.r-project.org/.

[39] Python, Python Lenguaje de Programación, 2020, https://www.python.org/.

[40] C. Shao, P. Cui, P. Xun, Y. Peng, and X. Jiang, "Rank correlation between centrality metrics in complex networks: an empirical study," *Open Physics*, vol. 16, no. 1, pp. 1009–1023, 2018.

[41] P. J. Green, "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives," *Journal of the Royal Statistical Society: Series B*, vol. 46, no. 2, pp. 149–170, 1984.

[42] P. Kragh and L. Theil, *Regression with Linear Predictors*, Springer, New York, NY, USA, 2010.

[43] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[44] A. Rodriguez, B. Kim, M. Turkoz et al., "New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network," *Scientometrics*, vol. 103, no. 2, pp. 565–581, 2015.

[45] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[46] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[47] J. Paul, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[48] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E*, vol. 73, no. 2, pp. 1–10, 2006.

[49] H. Chen, X. Li, and Z. Huang, "Link prediction approach to collaborative filtering," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference of Digital Library*, pp. 141-142, Denver, CO, USA, June 2005.

[50] T. J. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biologiske Skrifter, Kongelige Danske Videnskabernes Selskab*, Københavnn: I Kommission Hos E. Munksgaard, Denmark, 1948.

[51] D. J. Klein and M. Randić, "Resistance distance," *Journal of Mathematical Chemistry*, vol. 12, no. 1, pp. 81–95, 1993.

[52] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[53] P. Y. Chebotarev and E. V. Shamis, "The matrix-forest theorem and measuring relations in small social groups," *Automation and Remote Control*, vol. 58, no. 9, pp. 1505–1514, 1997.

[54] J.-Y Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery"" in *Proceedings of the 10th ACM SIGKDD International Conference Knowledge Discovery & Data Mining KDD'04*, pp. 653–658, Seattle, WA, USA, August 2004.

[55] L. Lü, C. H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, no. 4, Article ID 046122, 2009.

[56] H. Aidos, P. Robert, W. Duin, and A. Fred, "The area under the ROC curve as a criterion for clustering evaluation"" in

Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods (ICPRAM 2013), Barcelona, Spain, February 2013.

[57] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[58] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS '03)*, pp. 313–320, British Columbia, Canada, December 2003.

[59] M. Neuhauser, *Nonparametric Statistical Tests: A Computational Approach*, Chapman and Hall/CRC, New York, NY, USA, 2017.

[60] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 92–122, 2012.

[61] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical Review Letters*, vol. 87, no. 19, pp. 198701–198705, 2001.

[62] T. B. Athul and G. Suresh Singh, "Total graph of regular graphs," *Advances in Mathematics: Scientific Journal*, vol. 9, no. 6, pp. 4213–4220, 2020.

[63] E. Bryan, C. VerSchneider, A. Darren, and Narayan, "Efficiency of star-like networks and the Atlanta subway network," *Physica A*, vol. 392, pp. 5481–5489, 2013.

[64] M. Natarajan, "Correlation coefficient analysis of centrality metrics for complex network graphs," in *Intelligent Systems in Cybernetics and Automation Theory. CSOC 2015. Advances in Intelligent Systems and Computing*, R. Silhavy and S. Roman, Eds., vol. 348Czech, Springer, 2015.

[65] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[66] M. Dehmer, *Structural Analysis of Complex Networks*, Birkhäuser, Vienna, Austria, 2011.

[67] E. Estrada, D. J. Higham, and N. Hatano, "Communicability betweenness in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 388, no. 5, pp. 764–774, 2009.

[68] C. Durón, "Heatmap centrality: a new measure to identify super-spreader nodes in scale-free networks," *PLoS One*, vol. 15, no. 7, Article ID e0235690, 2020.

[69] N. Perra and S. Fortunato, "Spectral centrality measures in complex networks," *Physical Review E*, vol. 78, no. 3, Article ID 036107, 2008.

[70] H. Yue, Q. Guan, Y. Pan, L. Chen, and J. Lv, "Detecting clusters over intercity transportation networks using K-shortest paths and hierarchical clustering: a case study of mainland China," *International Journal of Geographical Information Science*, vol. 2019, pp. 1–29, 2019.

[71] A. Bramson, M. Hori, B. Zha, and H. Inamoto, "Scoring and classifying regions via multimodal transportation networks," *Applied Network Science*, vol. 4, no. 1, p. 97, 2019.

[72] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: a measure driven view," *Information Sciences*, vol. 507, pp. 772–794, 2020.

[73] B. Llc, "Normality tests: Kolmogorov-Smirnov test, Pearson's chi-square test, Anderson-darling test, D'agostino's K-squared test, Jarque-bera Test," 2010.

[74] Y. Boujelbene and A. Derbel, "The performance analysis of public transport operators in Tunisia using AHP method," *Procedia Computer Science*, vol. 73, pp. 498–508, 2015.

[75] P. P. Kumar, M. Parida, and M. Swami, "Performance evaluation of multimodal transportation systems," *Procedia-Social and Behavioral Sciences*, vol. 104, pp. 795–804, 2013.

[76] L.-G. Mattsson and E. Jenelius, "Vulnerability and resilience of transport systems - a discussion of recent research," *Transportation Research Part A: Policy and Practice*, vol. 81, pp. 16–34, 2015.

[77] J. Lin and Y. Ban, "Complex network topology of transportation systems," *Transport Reviews*, vol. 33, no. 6, pp. 658–685, 2013.

[78] L. Zhang, J. Lu, B.-B. Fu, and S.-B. Li, "A review and prospect for the complexity and resilience of urban public Transit network based on complex network theory," *Complexity*, vol. 2018, Article ID 2156309, 36 pages, 2018.

[79] X. Zhang, W. Li, J. Deng, and T. Wang, "Research on Hub node identification of the public transport network of Guilin based on complex network theory," in *Safe, Smart, and Sustainable Multimodal Transportation Systems (CICTP 2014)*, J. Ma, D. Pan, H. Huang, and Y. Yin, Eds., American Society of Civil Engineers, Reston, VA, USA, pp. 302–1309, 2014.

[80] X. Yang, S. Lu, W. Zhao, and Z. Zhao, "Exploring the characteristics of an intra-urban bus service network: a case study of Shenzhen, China," *ISPRS International Journal of Geo-Information*, vol. 8, no. 11, pp. 486–517, 2019.