國立臺灣大學理學院地理環境資源學系

碩士論文

Department of Geography
College of Science
National Taiwan University

Master Thesis

考量地理特性的 PageRank 演算法:
評估地理網絡節點之重要性

Geographically Modified PageRank Algorithms: Measuring the Importance of Nodes in a Geospatial Network

陳威全

Wei Chien Benny Chin

指導教授：溫在弘 博士

Advisor: Tzai-Hung Wen, Ph.D.

中華民國 102 年 7 月

July, 2013

國立臺灣大學碩士學位論文
口試委員會審定書

考量地理特性的 PageRank 演算法：
評估地理網絡節點之重要性

Geographically Modified PageRank Algorithms: Measuring the Importance of Nodes in a Geospatial Network

本論文係陳威全君（R00228026）在國立臺灣大學地理環境資源學系、所完成之碩士學位論文，於民國 102 年 6 月 21 日承下列考試委員審查通過及口試及格，特此證明。

口試委員：溫在弘 （簽名）

（指導教授）

黃書源 _____

_____ _____

_____ _____

_____ _____

系主任、所長 _____（簽名）

（是否須簽章依各院系所規定）

謝辭

感謝大家。大學至碩士班在台灣的這六年以來，感謝每一位師長的教導，及學長姐、學弟妹、朋友們的支持與幫助。

感謝指導教授溫在弘老師，在整個碩士生生涯中提供我很多的指導，除了研究上的各個細節方面，從溫老師這我學到的更重要的是做研究的態度與思維模式。除此之外，也非常感謝溫老師的鼓勵與支持，我才順利的到美國地理學會年會進行發表，這對我的人生來說，是一個非常重要的經驗與體驗。感謝兩位口委老師，黃崇源老師及林楨家老師，特意撥出時間來評閱我的論文紙本及參與我的口試與提供我論文撰寫上的建議與提出我研究中不足之處。在做研究的過程中，黃老師在數次討論中給的建議，是讓我研究可以做得更順利的關鍵之一。林老師在專討一就對我的研究提供了很重要的意見與概念，在專討二抽空擔任我的評閱老師時也提供了我很多重要的提醒，還有在 AAG 也抽空來聽我的發表。感謝各位老師的協助，如今我才能真的完成這份研究，寫完這份論文。另外，特別感謝李美慧老師，如果沒有李老師在大三、大四時對我的細心指導，從我完全不知道什麼是做研究，帶著我做完了大專生研究計劃及大四專題，也讓我對做研究有了興趣，我才會選擇念碩士，也才會有今天。

感謝研究室的各位，尤其是在這一年多以來，跟我討論並協助我解決在研究及學習上遇到的問題的逸翔、佳蓉，幫我處理很多學習以外的各種問題包括報賬、申請各種東西等等的玉珊。感謝各位一同打拼的碩士班同學，一起到 AAG 去發表的同學們，尤其是育棋及奕堯，在整趟行程上的規劃等。

最後，感謝我家人對我要到台灣來升學 6 年的這個決定的精神上的支持，幾資源上的資助！還有這六年來時刻的陪伴與無限的支持的女朋友素華。還有，一起從高中畢業以後跑來台灣唸書的老朋友們，文傑、康淵、青媚等等。懵懂的從高中畢業以後，跌跌撞撞地完成這六年的升學路上，感謝大家的陪伴。

感謝各位。希望大家事事順意。

威全

2013 年 8 月 1 日

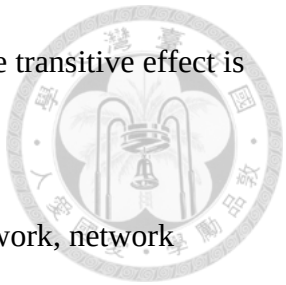
摘要

地理空間網絡是指以網絡方式再現空間單元之間關係的網絡。在社會網絡分析領域中，已有相當完善的研究在討論網絡核心性、小世界與無尺度網絡等各種網絡拓撲結構特性。PageRank (PR) 是基於網頁間連線結構的網絡分析方法，其是 Google 搜尋引擎排列搜尋結果的演算法。透過分析地理網絡拓撲結構，可以獲取地理網絡中節點的重要性資訊。然而，大部分網絡分析方法著重於討論網絡拓撲結構，這些方法所依據的原理並未考慮地理結構的特性，包括節點間距離遞減效果。所以，本研究將距離遞減效果及引力模型(即同時整合距離遞減效果及吸引效果)，並發展出兩個以地理特性及引力模型修改的 PR 演算法，包括 Inverse-Distance PageRank (IDPR) 及 Geographical PageRank (GPR)。為測試此 2 種演算法，本研究進行 2 個測試。其一為建立台灣城市間的 1 小時交通可及性網絡，經計算後與人口分佈及城際交通流量 (作為實際資料) 作斯皮爾曼等級相關分析 (Spearman Rank Correlation)，並與 PR 及 Weighted PageRank (WPR)演算法作比較討論；其二是以上述一小時交通網絡為基底狀態，討論在加入台灣高速鐵路系統之後，節點重要性的變化。本研究第一個測試的結果顯示 IDPR 及 GPR 與實際資料間的相關性最高，而第二部分測試結果顯示 WPR 及 GPR 的重要性變化具有網絡的遞移效果。因此，本研究認為 IDPR 及 GPR 較為適合用在地理空間網絡中，節點重要性的測定，而若遞移效果是關鍵的特性，則應考慮使用 GPR 作為分析方式。

Keywords: 距離遞減效果, 引力模型, PageRank, 地理空間網絡, 網絡遞移效果

Abstract

A geospatial network represents the spatial relationships with the network perspective. Within the scope of social network analysis, the network topology characteristics, including network centrality, small world and scale-free properties, have been well studied, and these concepts can also provide important implications on measuring the important of places in the geospatial network. PageRank (PR), which is an important link analysis algorithm, is what Google uses to determine how important a page is on the web. However, most measures of network analysis were designed to understand network topological structures rather than geographical structures. Therefore, these measures have not considered the geographical relationships as their main concern, including geographical distance decay effect between nodes. This study incorporates geographic properties, including distance-decay and spatial interactions among nodes, and proposes two modified PR algorithms, Inverse-Distance PageRank (IDPR) and Geographical PageRank (GPR). To test the performance of the index of importance (including IDPR and GPR), this study did two experiments with the inter-city network of Taiwan. In the first experiment, this study calculated the index of importance, and this study used the population data and inter-townships car flow data as observed data to check the Spearman Rank Correlation, and compared the correlation results with existing algorithms: PR and Weighted PageRank (WPR); in the second experiment, this study explore the changes of node's importance between before and after the construction of Taiwan High Speed Rail System. Our findings in the first experiment showed that IDPR and GPR are better correlated to the observed data, and our findings in the second experiment showed that the GPR and WPR could capture the transitive effect. Since IDPR and GPR take the distance decay effect into account, results using the algorithms can capture more geographical properties. In conclusion, IDPR and GPR



are better metrics to be used in geospatial network analysis; but, if the transitive effect is an important feature in the analysis, GPR is a better metric.

Keywords: Distance decay, gravity model, PageRank, geospatial network, network transitive effect

Table of Contents

Table of Figures

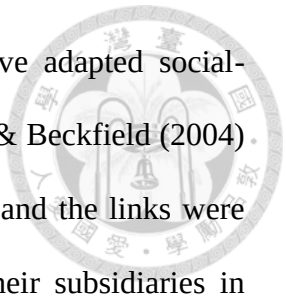
Table of Tables

1. Introduction

1.1. Background

The world contains extensive connections (Tobler, 1970; Commoner, 1971). Commoner (1971) once said that “there is one ecosphere for all living organisms, and what affects one, affects all”, concluding that “Everything is connected to everything else” — the first of The Four Laws of Ecology. This law can be extended from living organisms to, for example, geography, in which every spatial feature is connected because these features are located in a finite space. Tobler (1970, 2004) examined the population changes in Detroit and observed that population change in a given place depends on the previous population of not only that place but all other places, in other words “everything is related to everything else” — The First Law of Geography. In summary, elements in a system are connected, and their relationships should not be ignored. This implied that there is a structure of relationships between things, and this interaction would be resulting in the situation of each thing or place; for example, the population distribution of each city is a result of the interaction between cities (Tobler, 1970).

A geospatial network is a simplified version of the real world in the form of nodes and links, which emphasize the relationships between places. The nodes are formed by spatial features, which can be represented as points such as ports, airports, or buildings or as areas such as countries, cities, or regions. The links are formed by the relationships between the spatial features, such as the commuting path between regions or the airlines between airports. Analysis of the structure of geospatial networks can be used to elucidate the interactions among these features. To explore and understand the



spatial features and the structures within them, recent studies have adapted social-network-analysis frameworks to geography. For example, Alderson & Beckfield (2004) created a world-city network in which the nodes were world cities and the links were formed by the interactions among multinational enterprises and their subsidiaries in different cities; they explored the economic status and positions of each city with this network. El-Geneidy & Levinson (2011) and Reggiani et al. (2011) created a commuting network to calculate accessibility. Jiang (2009), Jiang et al. (2009), and Jiang & Jia (2011) created a street-street topology network to elucidate human movement on streets. Wang et al. (2011) created a network relationship between streets and used this network to explain the distribution of land-use relative importance. Ducruet et al. (2010) created an inter-port network to measure the vulnerability of each port. Ducruet et al. (2011), Guimera et al. (2005), Reggiani et al. (2009), and Scholz (2011) used the airline networks to explore the concentration of air transportation and find the hubs or hot spots. In summary, these studies created networks of spatial features and used them to explore geographical questions. In addition, these geospatial-network studies considered the relative impacts of geographical issues, the network position of the connected spatial features, and the vulnerability or strength of places (nodes) within the networks.

Recent studies have used social-network analysis metrics to retrieve information from the real world (Alderson & Beckfield, 2004; Guimera et al., 2005; Reggiani et al., 2009; Ducruet et al., 2010; Jiang, 2009; Jiang et al., 2009; Jiang & Jia, 2011; Ducruet et al., 2011; Reggiani et al., 2011; Wang et al., 2011; Scholz, 2011). In social-network analysis, network-centrality metrics are the most basic tools for understanding the node's or link's position and importance within the network. Traditional centrality metrics (Freeman, 1978) include degree centrality, closeness centrality, and

betweenness centrality. Of these measurements, degree centrality, a count of a node's linked neighbors, is the most intuitive and simple; closeness centrality and betweenness centrality describe the role or placement of a node within the network (Freeman, 1978). For example, in a friendships network, a person with a relatively high number of friends has high degree centrality; a person indirectly connected to all other people by relatively few steps (such that they rely on fewer people to spread information) has high closeness centrality; a person connecting two or more sub-groups has high betweenness. In geospatial-network analysis, network centralities were used to explore the places' characteristics in connection with interactions within the network. For example, Alderson & Beckfield (2004) examined the interactions among multinational enterprises and their subsidiaries across cities using out-degree centrality to show each city's influence on the world economy; in-degree centrality to show each city's ability to attract investment from other cities; closeness centrality to show each city's independence (as opposed to vulnerability); and betweenness centrality to show the potential of each city to act as a bridge, brokering the interactions (investment flow) among cities or subgroups of cities. In summary, this study uses various centrality measurements to show different types of importance in a geospatial network; this gives an implication and example for using network analysis perspective to explore geographical issues. Network centrality is used frequently in geospatial-network studies (Ducruet et al., 2011; Guimera et al., 2005; Reggiani et al., 2009; and Scholz, 2011); however, network centrality does not capture the transitive effect of network topology and, thus, underestimates or overestimates the importance of the nodes.

The transitive effect in network topology is a concept derived from the social network of trust. For example (see Figure 1.1), if Alice (A) trusts Bob (B), and Bob trusts Claire (C), then, Alice indirectly trusts Claire. In other words, Bob's trust in

Claire transfers to Alice (Jøsang et al., 2006; Jøsang et al., 2007). This concept is called "trust transitivity". In mathematics, a relation is said to be transitive if a→b and b→c together imply that a→c. Trust transitivity is relevant to reputation or recommendation systems (Jøsang et al., 2007; Symeonidis et al., 2010); for example, a person's reputation can be measured by his / her friends' reputations (Figure 1.2). If person (C)'s friends (who recommended him / her) have good reputations (B), then (C) should have a good reputation, too. In contrast, if person (E) has friends with poor reputations (D), then (E) should have a poor reputation relative to (C). PageRank (PR), the algorithm behind the Google search engine, employs this concept to rank pages, using the collection of the importance of the web pages that point to a given web page to evaluate that page's importance (Brin & Page, 1998). In a geospatial network, the importance of a high-level city should be transferred to its linked neighbors. Cities connected to a capitol city would thus be more important than cities that are not. PR is used to calculate the importance of nodes in a geospatial network, as in the street-street topology network used to measure and predict human movements (Jiang 2009, Jiang et al. 2009, Jiang & Jia 2011).

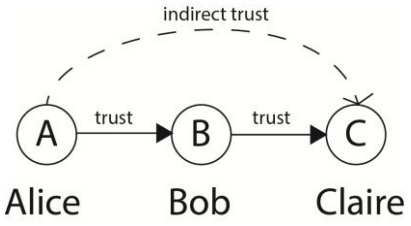


Figure 1.1 An illustration of trust transitivity.

The trust from Bob to Claire was transferred to Alice.

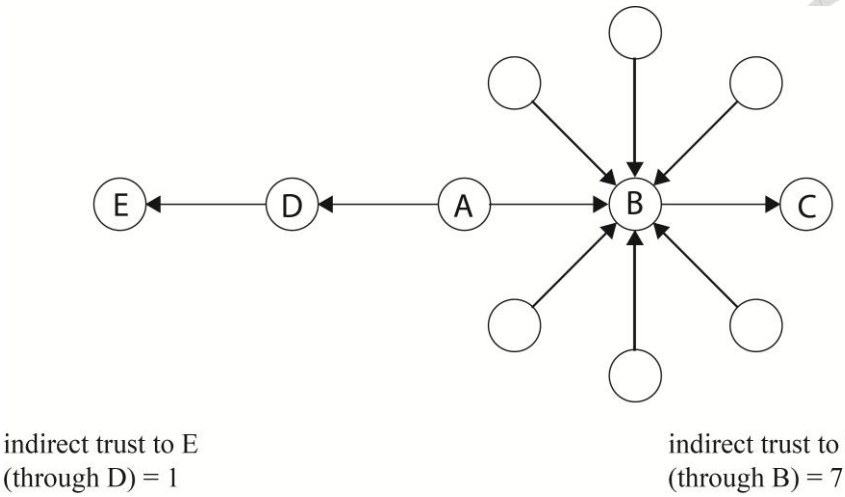


Figure 1.2 An illustration of reputation system and transitive effect.

An illustration of reputation system and transitive effect. B gains 7 direct trusts, therefore, C as the person B trusts, would gain 7 indirect trusts; on the other hand, D gain only 1 direct trust, therefore, E as the person D trusts, would only gain 1 indirect trust. Thus, C would get more trusts than B. And, the trusts is said to be transfer from B and D to C and E respectively.

In addition to the transitive effect, which can be understood as a forward effect (in which the influence moves from source to target), there is another effect within networks that can be understood as a backward effect (in which the influence moves from target to source), namely, the attractive effect. For example (Figure 1.3), if Bob needs to recommend someone to Alice, and there are two persons Bob (B) trusts (two candidates), Claire (C) and Freddy (F), Bob might say, "I recommend Claire and Freddy. Claire seems to be a more popular choice". Thus, Claire would gain more attention from Alice than would Freddy. In this example, Bob used in-degree centrality to compare the attractiveness of Claire and Freddy and gave this information to Alice, allowing Alice to differentiate the popularities of Claire and Freddy. This interaction is a backward effect; each node has a different level of attractiveness, leading to different levels of attention from the source. Alderson and Beckfield (2004) use in-degree centrality to examine the prestige of each city in terms of the ability to attract investment, which is also applied in

geospatial networks. In the original version of the PR (Brin & Page 1998), the attention given to each target is the same. For example, Bob would recommend both Claire and Freddy to Alice, but he would not tell Alice about their differences, so Claire and Freddy are equivalent to Alice and would gain same attention from her. In other words, the importance given by Bob is equally divided and transferred to Claire and Freddy. Thus, the original PR does not consider the attractive effect. To capture the attractive effect, Weighted PageRank (WPR) was introduced (Xing & Ghorbani, 2004). As in Alderson & Beckfield (2004), WPR considers the in-degree centrality as the attractiveness of nodes and uses this attractiveness level as a weight in the PR algorithm.

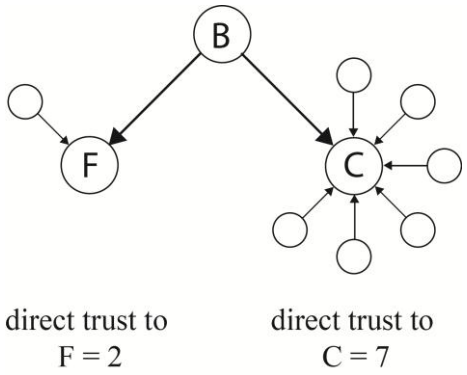


Figure 1.3 An illustration of attractive effect.

C gained more trusts than F, and this makes C more attractive and reliable. Thus, B should assign more trust to C than to F.

The nodes in a geospatial network represent spatial features such as airports, railway stations, or cities with set physical locations that should be considered in analyzing the geospatial network, whereas the nodes in a social network or hyperlink network represent people or web pages, which do not have a fixed geographical location. Tobler's first law of geography concluded that "near things are more related than distant things" (Tobler, 1970; Tobler, 2004), implying that the effect of geographical distance is an impedance -- the distance-decay effect. Distance decay is extremely important in

urban and economic geography (Fotheringham, 1981) and has been embedded in spatial analysis methods such as kernel-density estimation, which is a major tool for estimating the density surface of a spatial process that uses a distance-decay curve to assign weight to each event. In the distance-decay effect, the relative importance of interaction between two places declines as the geographical distance between them increases (Tobler, 1970; Fotheringham, 1981). For example, commuting flow from a near city is greater than the flow from a far city (Gargiulo *et al.*, 2012), and more airline passengers travel between close cities than between far cities (Fotheringham, 1981). As geospatial-network analysis emphasizes the relationships and interaction between places, the distance-decay effect should not be neglected; however, the measurements introduced above are all designed to elucidate social networks or the topological structure of networks, which do not consider the geographical distance between nodes in the calculation.

The gravity model, which also called the spatial-interaction model, is used in social and economic studies to describe certain behaviors and spatial processes. The gravity model mimics gravitational interaction as described by Newton's law of universal gravitation, which states that "every point mass attracts every single other point mass by a force pointing along the line intersecting both points." The first part of this law states that there are relationships between every pair of single point masses or bodies, which is similar in concept to Commoner's four laws of ecology and Tobler's first law of geography. The second part of this law describes a force between the two points, which is the gravitational force, where the gravitational force is proportional to the product of the two masses and inversely proportional to the square of the distance between them. The gravity models (Reily, 1931; Stewart, 1950; Tinbergen, 1963) use a similar idea to describe phenomena in geography or spatial processes, like the

commuting flow between cities (Gargiulo *et al.*, 2012), and the air-transportation flow (Fotheringham, 1981). The gravity models in geographical studies replace "bodies" and "mass" with "places" and "importance", where importance can be measured in terms of population numbers (or density), economic status, or other variables. In the calculation of the gravity models, the gravitational force between each pair of places is proportional to the attraction (the importance variables) between the two places and inversely proportional to the geographical distance between them. In addition, the geographical distance acts as an impedance (or cost) and exerts a distance-decay effect on the gravitation force. The gravity model is a combination of the attractive effect and distance-decay effect.

In conclusion, by analyzing geospatial networks, this study examines the importance of the interaction between spatial features. Previous geospatial-network studies focused on the relative importance of geographical issues, the position or characteristics of connected spatial features, or the vulnerability or strength of a specific place within the network. Freeman's network-centrality metrics formed the foundation of the node's importance measurements in social networks, which were also frequently used in geospatial-network analysis, but it does not consider the transitive effect captured by the PR algorithm. These measurements were designed and developed in social-network and physics contexts; the distance-decay effect is not considered. Distance decay is extremely important in urban and economic geography. On the other hand, the gravity model is used in social and economic studies to describe certain behaviors and spatial processes. In other words, the distance-decay function and the gravity model are important geographical properties in the discussion of spatial interaction, but neither is considered in the recent network-analysis metrics.

1.2. Research objective

The objective of this paper is to develop two novel algorithms for measuring the importance of each node in geospatial networks that incorporate the distance-decay function and the gravity model. This study used degree centrality, which calculates the number of connected neighbors for each node to measure its power or ability to influence other nodes. Because the PR algorithm (Brin & Page, 1998) is an extension of the degree centrality, which can capture the transitive effect and measure the importance by the same definition, this study uses the calculation framework of PR as the basic structure of our novel algorithms. To capture the geographical properties in the geospatial-network analysis, this study integrated the distance-decay function and the gravity model in the algorithm.

The first algorithm proposed in this study can vary the effect of geographical distance between the nodes to capture the distance-decay effect in the PR algorithm. For this reason, it is called the Inverse-Distance PageRank (IDPR) algorithm. The second algorithm integrates the attractiveness function and the distance-decay function to incorporate the gravity model in a Geographical PageRank (GPR) algorithm.

To elucidate the potential of these novel algorithms, this study includes a case study with these algorithms in an inter-city network and used the data on the intensity of human activity for validation. Then, to understand how the measured importance is affected by changes in the topological structure of the network, this study added Taiwan's high-speed-rail system to the inter-city network.

The specific aims of this study are as follows:

1. To develop two novel measurements to integrate the distance-decay function and gravity model into the PR algorithm to analyze importance in geospatial networks.

2. To test the ability of the two measurements proposed to identify major cities in an inter-city network developed from the Taiwanese ground-transportation network (road layers and railway layer).

3. To add a new ground-transportation system to the inter-city network, namely Taiwan's high-speed rail system, to analyze the changes in the IDPR and GPR.

2. Related works

This chapter first discusses the characteristics of geospatial networks because different types of network employ different assumptions and would yield different results. Second, this study discusses centrality metrics. Third, this study discusses the existing PR algorithm.

2.1. The geospatial network

There are two types of geospatial network: a connection network and an interaction network. Theoretically these two types differ in ways that give rise to different interpretations.

2.1.1. Connection network

The links in a connection network exist physically, like airlines between airports (Reggiani et al., 2009), shipping lines between ports (Ducruet et al., 2010), or streets between places (Jiang, 2009; Wang et al., 2011). The links in the network can be binary, as for streets (Wang et al., 2011), and can have information capacity, as for ships (Ducruet et al., 2010). As the connection network is used to show the accessibility of connections between places, this type of network can provide information about the potential capability (or importance) of the nodes and the links within the geospatial network. In other words, a connection between nodes would provide the possibility of interaction, increasing the potential importance of the nodes and links. For example, if there is a route connecting city A and city B, then there is the possibility that people will move from city A to city B and vice versa; in contrast, if there is no route connecting

city A and city C, then there is zero possibility that people might move from city A to C or from city C to A directly. Therefore, the potential mutual importance of city A and city B might be greater than that of city A and city C. Jiang (2009) and Wang et al. (2011) used this logic and transformed streets into nodes to form a street network. By analyzing these networks, they found that the structure of the network can represent human movements or the intensity of human activity. Ducruet et al. (2010) analyzed the number of shipping vessels moving between the ports and found that the structure of the network can be used to measure the vulnerability of the ports. In summary, previous studies used connection networks to analyze the potential to attract human activity or the potential to become vulnerable using the network topology.

2.1.2. Interaction network

The links in an interaction network represent interactions between the nodes (for example, commuters; El-Geneidy & Levinson, 2011), the relationships between the headquarters and their subsidiaries (Alderson & Beckfield, 2004), and the total number of airline passengers (Derudder *et al.*, 2006). The interaction network also can be both binary (the relationship exists or does not) and weighted (a measure of flow through the links); for example, the relationships between headquarters and their subsidiaries (Alderson & Beckfield, 2004) are binary, but the number of commuting people and the total number of airline passengers (Derudder *et al.*, 2006; El-Geneidy & Levinson, 2011) are weighted. As the links are constructed based on observations, the relationship network shows and describes the recent relationships as the outcomes of interactions. For example, if 1000 persons move from city A to city B and 10 persons move from city B to city C, then human movement from city A to city B is greater than that from

city B to city C. This is not likely a result of the potential importance of a place as described in the connection network; this is more likely a result or current characteristic of an interaction between places. In another example, if we know that there is a high-speed railway connecting city A and city B, we can only say that people can move between city A and city B, but the exact number of people moving between the places is unknown. In contrast to the former example that measure the potential to move between places (as a result of the availability of physical facilities), this example shows the results of human movement, including human preference. Therefore, the relationships network can be used to show the observed relationships between places, and this network can be analyzed to understand the structure underlying the observed characteristics. For example, Alderson & Beckfield (2004) listed the locations of multinational enterprises and their subsidiaries and aggregated them into cities, with directional links from the headquarters city to the subsidiaries city. Then, they derived a global network of economic relationships. In summary, analysis using the interaction network uses a network to represent the observed relationships between nodes (places), which would give information about the status as a result of the interaction at a given time, and requires network data, as in previous such studies.

2.1.3. Summary

Theoretically, in the analysis framework, the connection network and the interaction network differ in the network-development stage and thus differ in the explanation stage. Analysis of different networks would yield different types of data. The connection network is constructed based on the availability of the connection, which would give information about the potential interaction of the nodes or links

within the network, while the interaction network is built based on the observed interaction and would give information about the status as a result of the interaction at a given time. Therefore, to explain the network analysis without over-interpretation, the type of the network should be known before analysis.

2.2. Network centrality metrics

Centrality is a series of metrics developed from network analysis and graph theory. Centrality holds that the most important node should be the center of the network (graph). Thus, the aim of centrality metrics is to measure the importance of each node to show the "position" of the node within the network. In this section, this study reviews the centrality metrics for node, including degree centrality, closeness centrality, betweenness centrality, and flow-betweenness centrality. These centrality metrics measure the importance of the nodes based on different concepts.

2.2.1. Degree centrality

Degree centrality is the most intuitive and simple of the three centrality metrics. The logic of degree centrality is very simple: the more linked neighbors a node has, the more important it is. Therefore, degree centrality calculates the number of links for each node (Equation 1). In an undirected network, the degree centrality of each node is computed by counting the number of links connecting the node; in a directed network, degree centrality can be separated into in-degree centrality, which is computed by counting the number of links that point in toward the node, and out-degree centrality, which is computed by counting the number of links that point out from the node.

$$\mathbf{C_{degree}}(\boldsymbol{i}) = \sum_{j \in G(i)}^{n} \boldsymbol{A_{ij}}$$

Equation 1

Where:

$\mathrm{C}_{\mathrm{degree}}(i)$: degree centrality of the node- i;

$\mathrm{G}(i)$: all other nodes in the graph, except the node-i;

A_{ij} : adjacency matrix of the node i, for undirected network, $A_{ij} = 1$ if there is a connection between node-*i* and node-*j*, else 0; for directed network, for calculating indegree, $A_{ij} = 1$ if node-*j* is pointing to node-*i*, else 0; for calculating outdegree, $A_{ij} = 1$ if node-*i* is pointing to node-*j*, else 0.

The degree centrality equation captures the basic logic of the degree centrality as noted above. For example, in undirected Facebook networks, people are friends with people they know; in directed networks, people might follow (like) their idols or be followed by their fans. The degree centrality of a given person is high if s/he has many friends; the in-degree centrality of the given person is high if s/he has many fans; the out-degree centrality of the given person is high if s/he is a fan of many people.

In a geospatial network, degree centrality shows the number of linked neighbors of a given place. As the number of linked neighbors is equal to the number of places that would directly interact or have relationships with a given place, degree centrality shows the number of places a place would loan its influence or power to (out-degree centrality), the number of places that would loan their influence to that place (in-degree centrality), or both (in an undirected network, degree centrality). Thus, degree centrality,

in-degree centrality, and out-degree centrality have different implications in different studies.

Degree-centrality metrics were used in recent geospatial analysis studies; for example, Alderson & Beckfield (2004) used out-degree centrality and in-degree centrality to show the power and the prestige of the city, respectively. The findings of this study show that larger cities rank higher in out-degree centrality and that cities in semi-peripheral countries are, on average, 304 ranks lower than cities in core countries, whereas cities in peripheral countries are 510 ranks lower. Moreover, the result for in-degree centrality, which is used to represent the prestige of each city, is also similar to that for out-degree centrality, implying that cities in semi-peripheral and peripheral countries are, on average, less prestigious than cities in core countries. To explain the global economic structure and the status of each city, Alderson & Beckfield (2004) integrated the network they built and the logic of the centralities in an example of how to include the phenomena or the network in network-analysis metrics; the idea is transformed from the network to the geographical scale. If the methods are used correctly, social-network analysis can be used to explore the geospatial network.

Degree centrality assumes that every node is equal in weight. Therefore, degree centrality can be calculated from the number of the nodes immediately connected to the given node. In the real world, however, not all neighbors are equivalent (some nodes might be more important, as discussed as the transitive effect), so eigenvector centrality is introduced.

Eigenvector centrality (Bonacich, 1987) is an extension of simple degree centrality. As Freeman's degree centrality implies that node's importance increased by having connection with more neighbors, eigenvector centrality implies that the node's

importance increased by having connection with other nodes that are themselves important (Newman, 2010). The results of eigenvector would also include the property that the result of a vertex is high if it has many neighbors, has important neighbors, or both. The purpose of eigenvector centrality is to distinguish the importance of neighbors, which can be understood as to capture the transitive effect. In addition, PR can be understood as a variant of the eigenvector centrality, which is reviewed in the next section (2.3).

2.2.2. Closeness centrality

The purpose of closeness centrality is to measure how close (mean distance) from each node connects to the rest of the network. The logic of closeness centrality can be summarized as follows: nodes rely on other nodes to relay messages through the network, so the closer a node to the rest of the world, the more important it is. Assume 2 separate Facebook user's networks formed from 2 communities, with the links representing the friendships within the communities. In one of the networks (network 1), there is a person (person A) who knows everyone directly, and in the other network (network 2), another person (person B) only knows a few people directly. In other words, person A can reach all other people within 1 step, and person B might need a few more steps to reach all other people in the network. Therefore, if these two persons share news on their user page, all other people in network 1 would immediately get the news on their page from person A; in network 2, most people will need more time to obtain the news that is shared by person B; further, if the direct friends of person B refuse to share the news, then other people will never get that news. In other words, person B needs to rely on people to spread the information through the community

network (less independent), whereas person A will need to rely on nobody to spread the news out (more independent). In network analysis terms, person A is said to be close to the rest of the network, or have high closeness centrality, whereas person B is said to be far from the rest of the network, or have low closeness centrality. Nodes that have greater closeness centrality within the network are more independent in getting or spreading information with the rest of the network. So, closeness centrality can be understood as a measure of how independent a node is on the information spreading process. To calculate closeness centrality, farness, which is the total steps needed from each node to reach all other nodes following the shortest path ($p\mathrm{ath}_{ij}$), must be calculated. Then, closeness centrality is the inverse of farness (Equation 2). As the calculation of closeness centrality would only consider the path between the nodes, the direction of links (directed or undirected link) would not change the equation but would change the shortest path. Figure 2.1 is a demonstration for calculating farness and closeness.

$$\mathrm{C}_{\mathrm{closeness}}(i) = \left(\mathrm{farness}(i)\right)^{-1} = \left(\sum_{j \in G(i)} path_{ij}\right)^{-1}$$

Equation 2

Where:

$\mathrm{C}_{\mathrm{closeness}}(\mathrm{i})$: closeness centrality of the node i;

$\mathrm{farness}(i)$: farness of the node i;

$\mathrm{G}(i)$: all nodes in the graph, except the node i;

path_{ij} : the number of steps of shortest path from node i to node j.

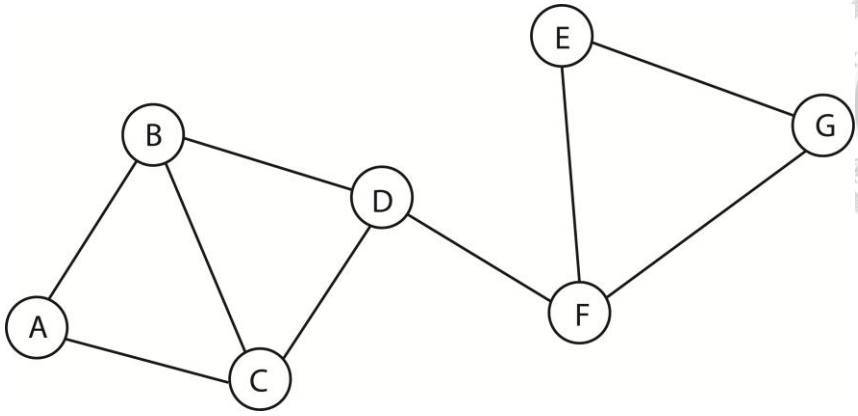


Figure 2.1 A demonstration for calculating farness and closeness.

For node-D, D-B, D-C, D-F need only 1 step, D-A, D-E, D-G need 2 steps, the farness is equal to the total steps needed from F to all nodes, which is $1 + 1 + 1 + 2 + 2 + 2 = 9$, and the closeness is equal to the inversed of farness, which is $1/9$; for node-G, G-E, G-F need 1 step, G-D need 2 steps, G-B, G-C need 3 steps, and G-A need 4 steps, by the same calculation, farness = 14, and closeness = $1/14$.

As closeness centrality can be understood as a measure of a node's independence of the information-spreading process, this idea can be translated into a resource-sending process: if the information is changed to resources and the friendship network is changed to a geospatial transportation network, the logic of closeness centrality is translated as the independence of the resource- or commodity-transporting process. That is, the place that is closer to the rest of the world (in the number of topological steps) is more independent and thus, more important.

In Alderson & Beckfield (2004), closeness centrality is used as an indicator of the independence of the world cities. Their results showed that the ranks of the closeness centrality and out-degree centrality (or in-degree centrality) are very different: in closeness centrality, many important developing cities fell into the top 50, whereas in out-degree centrality and in-degree centrality, the top 50 cities were all developed. As the network represents the structure of the world economic interaction, this result

implies that these cities have relatively independent positions within the network; thus, they are less vulnerable than other developing countries/cities.

Closeness centrality includes an assumption that information spreads to all other nodes through the fastest and most direct path, but in the real world, information might spread through a longer path or randomly. To address networks that do not fit the shortest-path assumption, modifications, including random walk and harmonic-mean length, were developed. While the original closeness centrality assumes that information spreads along the shortest path, random-walk closeness centrality (Noh & Rieger, 2004) assumes that information spreads by random walk. In other words, when information reaches a node, this node randomly chooses one linking node, and the mechanism repeats until all nodes have the information. On the other hand, information centrality (Stephenson & Zelen, 1989) measures the harmonic mean length of paths ending at a given node; this metric is smaller if there are many short paths connecting it to other nodes.

2.2.3. Betweenness centrality

Betweenness centrality, like closeness centrality, focuses on the positions of nodes within a network but measures how much control a node has over the flow of information or resources within the network (Figure 2.2). The purpose of betweenness centrality is to find the bridge or broker that controls information or resource flow between sub-networks.

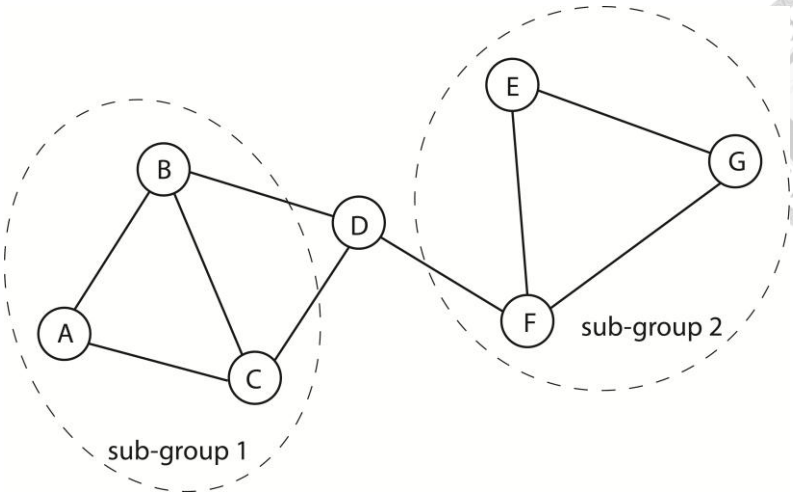


Figure 2.2 An illustration of the bridge position.

Node A ~ C formed subgroup 1, and node E~G formed subgroup 2, node D located between subgroup 1 and subgroup 2, so node D could control the information or resources flow between the subgroups, namely a bridge or a brokers.

In the calculations, betweenness centrality assumes information will flow through the shortest path, and it counts the number of times that the flow between each pair of other nodes (pair of *j*-*k*) passes through node *i* (∂_{jik}), then divides the number by the total number of other pairs of nodes (∂_{jk}; Equation 3).

$$C_{\text{Betweenness}}(i) = \frac{\partial_{jik}}{\partial_{jk}}$$

Equation 3

Where:

$C_{\text{Betweenness}}(i)$: betweenness centrality of the node i;

∂_{jk} : total number of other pairs of nodes;

∂_{jik} : the times of the shortest path between each pair of other nodes (pair of *j*-*k*) passing through the node *k*.

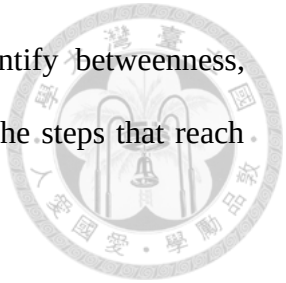
As betweenness centrality considers only the shortest path that passes through the node of interest, a change in the direction of the links would not change the equation but would change the shortest path.

The node that has a bridge or broker position in a network controls the flows between other nodes. This idea can be used to discuss the characteristics of a geospatial network. Betweenness centrality reflects the idea that a place has a structurally advantageous position when it falls between other places in the network (Scholz, 2011), giving it greater power in the sense that it brokers all exchanges (Alderson & Beckfield, 2004). For example, in an airline network, an airport with high betweenness centrality is a hub; most other airports nearby may need to connect with the hub to reach other airports.

Betweenness centrality is frequently used in airline networks because identifying the hub node or determining whether the network contains a hub node is important in air-transportation systems. For example, Reggiani *et al.*, (2009), Scholz (2011), and Ducruet *et al.*, (2011) used betweenness centrality to understand the structure of airline networks and the business strategies of different airlines. On the other hand, betweenness centrality is used in city networks; for example, in Alderson & Beckfield (2004), the cities' rank of betweenness centrality is highly correlated with out-degree centrality but only modestly correlated with closeness centrality.

As for closeness centrality, in the calculation of betweenness centrality, only the shortest path is assumed to control the flows between nodes, but alternative bridge nodes should also have high betweenness centrality. To address this problem, extensions of betweenness centrality were developed, including flow betweenness (Freeman *et al.*, 1991), and random-walk betweenness (Newman, 2005). Flow

betweenness (Freeman *et al.*, 1991) uses maximum flow to identify betweenness, whereas random-walk betweenness replaces the shortest path with the steps that reach each node first by random walk.



2.2.4. Flow Betweenness centrality

As an extension of betweenness centrality, flow betweenness centrality (Freeman et al., 1991) also measures the probability that a node lies on the path connecting a pair of nodes. Unlike other centrality metrics, flow betweenness centrality can be used to analyze both binary and weighted networks. The "flow" in this centrality metric is an idea that is proposed by Ford & Fulkerson (1956, 1957, 1962). The key idea in flow betweenness is that the maximum flow from a source node to a target node is equal to the minimum capacity of the cut set (all of the possible paths that connect the given pair of nodes).

$$C_{flow\ betweenness}(i) = \frac{\sum_j \sum_k \left(M_{jk}(i)\right)}{\sum_j \sum_k M_{jk}};\ j,k \in G(i), j < k$$

Equation 4

Where:

$C_{flow\ betweenness}(i)$: flow betweenness centrality of the node i;

$M_{jk}(i)$: maximum possible flow from node-j to node-k that passes through node-i;

M_{jk} : maximum possible flow from node-j to node-k.

Flow-betweenness centrality is more complex than the previous centrality metrics. Roughly speaking, flow betweenness centrality calculates the sum of the

minimum flows that must pass through a given node between a pair of nodes to reach the maximum flow between them.

First, the cut set for each pair of nodes must be identified. For example, in Figure 2.3 (a demonstration network from Freeman *et al.* (1991), the cut set between node A to node C includes the paths A-C, A-B-C, and A-D-C, and the maximum flow between node A and node C is equal to 6 (the sum of the minimum capacity of each path). For another example, the cut set between node A to node D includes the paths A-D, A-C-D, and A-B-C-D, and the maximum flow between node A and node D is equal to 4 (for the same reason). This step considers all possible paths and calculates the maximum flow between all pairs of nodes. This maximum is the target number for the next step.

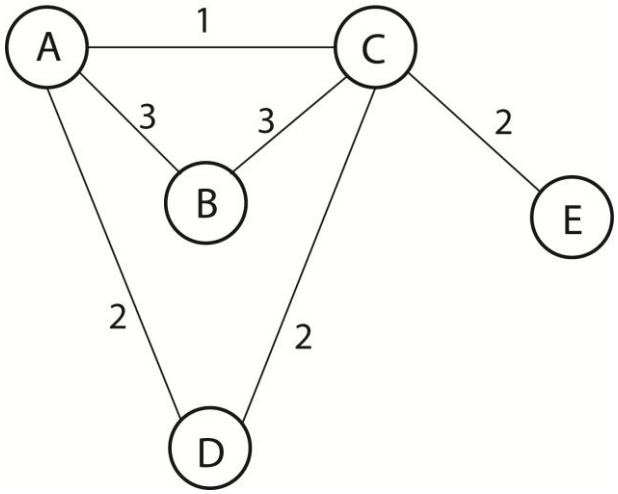


Figure 2.3 A demonstration network for flow betweenness.
The number beside links indicates the flow capacity of the correspondence link.

Second, the minimum flow for each path must be computed. For example, from node A to node E, the maximum flow is 2. To reach this target, if we calculate the minimum flow that must pass node D, then 1 unit can flow through path A-C-E, and the other unit can flow through path A-B-E; therefore, node D is not required to achieve

maximum flow. If we calculates the minimum flow that must pass node B, then 1 unit can flow through path A-C-E, and the other unit can flow in path A-D-E; therefore, node B is not required (Table 2.1). Flow betweenness thus tries to capture the position that is required to achieve the maximum flow between nodes. In other words, the greater the flow betweenness, the more it is required to pass between a pair of nodes. Flow betweenness is calculated as shown in (Equation 4).

For the example in Figure 2.3, the calculation result is shown in Table 2.1. The node that has the highest flow betweenness is node C, and node A is the second highest, then node D, then node B, and node E is the lowest.

Table 2.1 The calculation of flow betweenness for each node.

Node-1	Node-2	max possible flow	direct flow	the minimum flow that must pass				
				node-A	node-B	node-C	node-D	node-E
A	B	6	3			3	2	0
A	C	6	1		3		2	0
A	D	4	2		1	2		0
A	E	2	0		0	2	0	
B	C	6	3	3			2	0
B	D	4	0	2		2		0
B	E	2	0	0		2	0	
C	D	4	2	2	1			0
C	E	2	2	0	0		0	
D	E	2	0	0	0	2		
Total flow that must pass:				7	5	13	6	0
Flow betweenness:				0.35	0.25	0.65	0.25	0

The results for reversed arrangement of first and second node are same for undirected network.

The flow-betweenness centrality indicates the flow that must pass through each node. In a geospatial network of traffic flow, this metric can be understood as the minimum transportation flows that must pass through each city. In other words, if the

flow betweenness of a given city is high, then this city is most likely a hub in the transportation network. Flow betweenness uses the max-flow min-cut theorem (Ford & Fulkerson, 1956) to measure the flows that would pass through a given node and uses this information to represent the importance of a node. The flow betweenness would never overestimate the importance of each node because it considers only the minimum flow. Therefore, flow betweenness can be considered necessary for the nodes to reach maximum flow. This concept can be useful in an analysis of transportation-traffic flow.

2.2.5. Summary

Derived from network analysis, topological structure is the key to understanding the position and power of nodes derived from linking relationships. Centrality metrics help us to understand topological structure. From geospatial-network analysis, in recent geospatial network studies, centrality metrics were used frequently to analyze different types of importance; for example, out-degree centrality is a measure of power; in-degree centrality is a measure of prestige; closeness is a measure of independence; betweenness centrality is a measure of the potential to be a hub in the air-transportation network, and the flow betweenness is a measure of the importance of the place to maximizing air-traffic flow.

2.3. The Existing PageRank algorithms

With the rapid development of the Internet, the retrieval of important information from a large database has become an important issue in information science. Algorithms, including PR, were developed to retrieve information from large databases according to hyperlink structure (Brin & Page, 1998). PR can be considered a variant of the eigenvector centrality, which distinguishes the importance of neighbors. WPR (Xing and Ghorbani, 2004) is a variant of PR that modifies the weight in the algorithm.

This section briefly introduces the basic PR algorithm (Brin & Page, 1998) and a modified PR algorithm named WPR (Xing and Ghorbani, 2004).

2.3.1. PageRank (PR)

The hyperlink network can be used to conceptually organize the Internet. Thus, analysis of the hyperlink network can provide clues to retrieving information from webpages. PR is a famous, successful algorithm that addresses the hyperlink network. This algorithm, which is used by Google to rank search results, measures the importance of a webpage based on the hyperlink network (Brin & Page, 1998). The hyperlink network is composed of webpages (as nodes) and hyperlinks (as directed links). Like other network-centrality measurements, PR aims to identify important webpages from large, complex networks.

PR is based conceptually on a simulated person who surfs the Internet at random from a webpage's hyperlink network, which has many webpages and many directed links (hyperlinks) that point to other pages. The simulation begins with many random surfers on randomly selected webpages and then lets them select a hyperlink and move

on in a random walk. All webpages record when they are reached by the random surfers. After enough iterations, the simulation stops. The probability of a page being reached can be calculated by dividing the number of times it is reached by the product of the total iteration times and the total number of random surfers. Thus, the webpages with high probabilities of being reached are important because they are pointed to from many webpages and from other important webpages. The probability for each webpage is the PR score. In agent-based-modeling terms, the PR score is determined by the chances of being selected by the random surfers of each hyperlink, which can be calculated by the proportion equation.

Therefore, PR calculation (Equation 5) can be understood as an iterative voting process (Figure 2.4) within the hyperlinks network. At the beginning, all of the nodes have an equal number of votes (initial PR score). Then, all of the nodes send their votes to their outgoing link-neighbors and receive votes from their incoming link-neighbors. This voting process will be repeated until equilibrium is reached (final PR score).

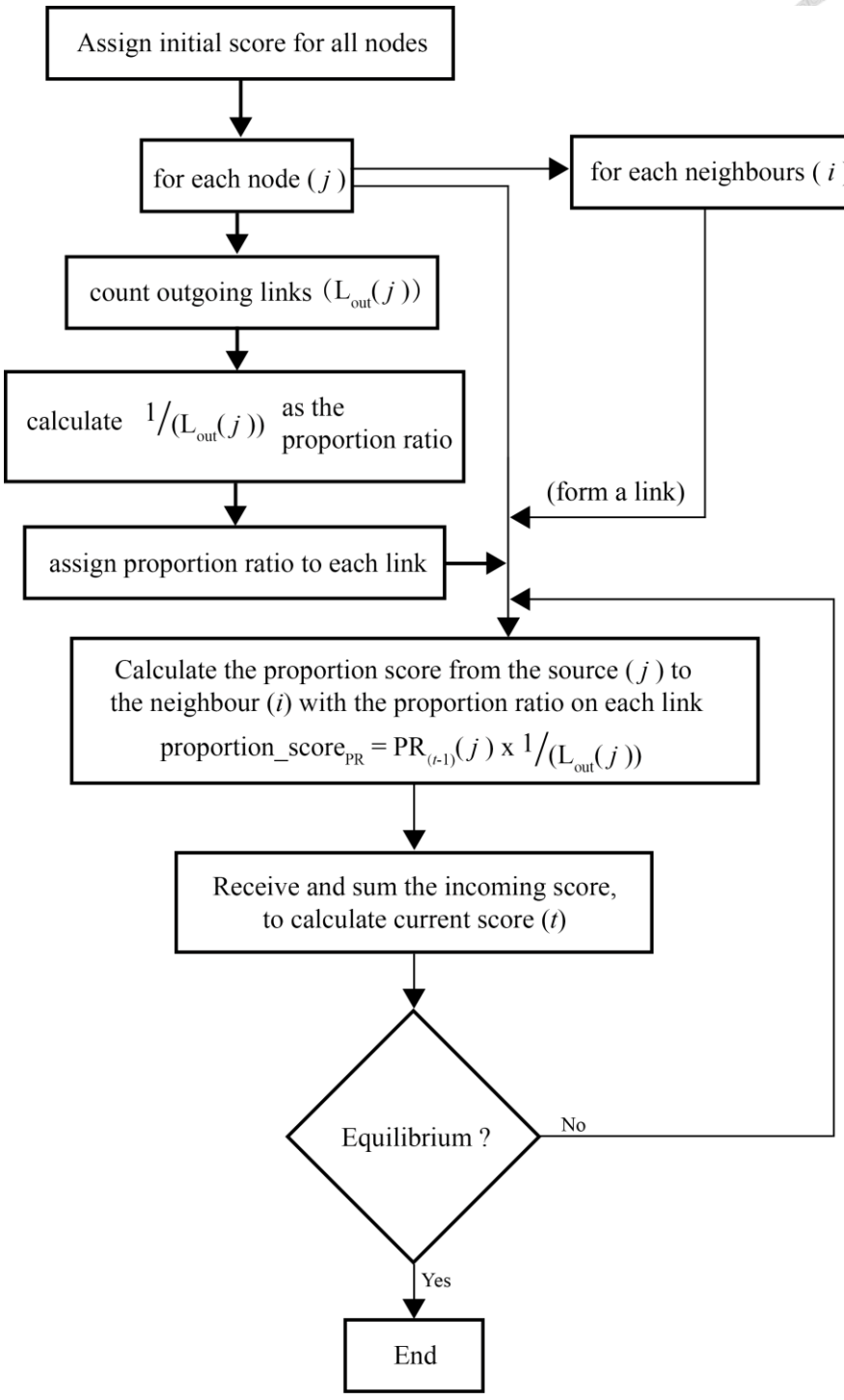
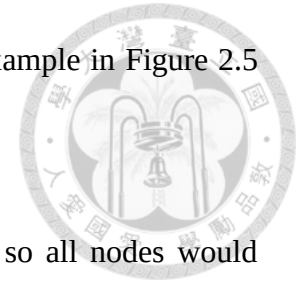


Figure 2.4 The calculation procedure of PR algorithm.



According to the calculation procedure in Figure 2.4, the example in Figure 2.5 shows how to calculate the PR, using nodes 2 and 5 as follows:

1. Assigning initial score: There are 5 nodes in the network, so all nodes would have $\frac{1}{5} = 0.2$ PR at the beginning.
2. Sending outgoing score: Node-2 has 3 outgoing link-neighbors, node 1, node 3, and node 5, so node 2 would send $\frac{1}{3} \times PR(node\ 2)$ to node 1, node 3, and node 5 (Table 2.2).
3. Receiving incoming score: On the other hand, node 5 has 2 incoming link-neighbors, node 2 and node 3; node 2 has 3 outgoing link-neighbors, and node 3 has 2 outgoing link-neighbors. Therefore, the PR score of node 5 at the end of this iteration is equal to $= \frac{1}{3} PR(node\ 2) + \frac{1}{2}\ PR(node\ 3)$. (Table 2.2)
4. After this iteration finishes, the next iteration repeats step 2 and step 3.
5. The calculation process is stopped when the PR score of each node reaches equilibrium.

Table 2.2 A demonstration of the calculation for the PR proportion.

Link *j*-*k*	$L_{out}(j)$	$1/L_{out}(j)$
2-1	3	$1/3$
2-3	3	$1/3$
2-5	3	$1/3$
3-1	2	$1/2$
3-5	2	$1/2$

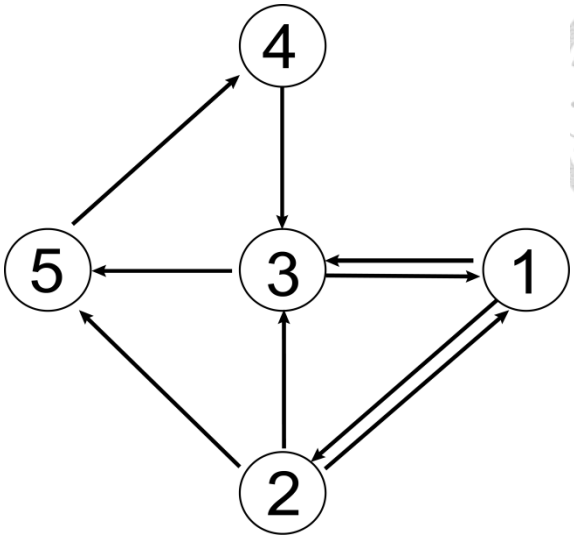


Figure 2.5 An example network.

The equation of PR, for $j \in \mathrm{IN}(i)$ (as shown in Figure 2.7):

$$\mathrm{PR}_t(i) = \sum \left(\mathrm{PR}_{t-1}(j) \times \frac{1}{\mathrm{L}_{\mathrm{out}}(j)} \right)$$

Equation 5

Where:

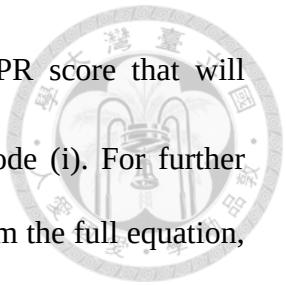
$\mathrm{PR}_t(i)$: PR score of the node i in iteration t;

$\mathrm{PR}_{t-1}(j)$: PR score of the node j in iteration t-1;

$\mathrm{L}_{\mathrm{out}}(j)$: The number of outgoing links of the node j;

$\mathrm{IN}(i)$: The set of incoming link-neighbours of the node i.

In summary, the PR equation is shown above (Equation 5). The PR score of a given node is determined by the scores of its incoming links. This is a dynamic system in which the nodes are stocks and the links represent flow between stocks. A stock ($\mathrm{PR_t(i)}$) is the PR score of a node i at iteration t with the same structure for the stock



$(\mathrm{PR}_{t-1}(j))$. A flow $(\mathrm{PR}_{t-1}(j) \times \frac{1}{\mathrm{L}_{out}(j)})$ is the proportion of the PR score that will transfer through the links from the source node (j) to the target node (i). For further discussion and development, the proportion equation is separated from the full equation, as shown below (Equation 6 and Equation 7).

$$xPR_t(i) = \sum (xPR_{t-1}(j) \times Prop_{xPR})$$

Equation 6

$$Prop_{PR} = \frac{1}{\mathrm{L}_{out}(j)}$$

Equation 7

On the other hand, PR can be understood as a Markov-chain process in which the states are nodes (or stocks) and the transition probabilities are the links between nodes (or flows between stocks). Because of the Markov-chain process, this calculation would reach an equilibrium status after enough iteration. Because PR calculation is also a Markov-chain process, after enough iterations, the distribution of the PR score would reach equilibrium. This characteristic is the same as that used in the calculation of eigenvector centrality: given a network, if a single greatest eigenvalue does exist, the correspondence eigenvector is the eigenvector centrality of the nodes in the network, but if there are two (or more) greatest eigenvalues, the score oscillates between them. Because the geospatial network is normally complex, the latter outcome is rare. The PR calculation is consistent with the eigenvector, so it normally reaches equilibrium.

PR calculation is similar to in-degree centrality in that both use the incoming link-neighbors of each node to calculate their importance. Furthermore, PR can also be understood as a variant of eigenvector centrality. Like eigenvector centrality, PR calculation uses an iterative voting process to differentiate the importance of incoming

link-neighbors. PR calculation modifies the adjacency matrix by dividing it by the number of outgoing links ($\frac{A_{ij}}{L_{out}(j)}$). PR is a network reputation system; the PR score is sent to the outgoing node, and nodes that receive relatively high PR scores from other high-scoring nodes (analogous to a recommendation from an important person) are important. As a result, PR captures the network's transitive effect.

As the summed PR scores of all nodes would equal 1, the calculation results of PR score would equal to the probability of a surfer stopping at each node. More surfers stop at nodes that have better positions in the network, so the probability of the surfer stopping at a given node can be understood as the relative importance of the node.

The PR is a network metric that helps to elucidate the topological structure and physical logic of a network. Therefore, it can be used to analyze other networks that are similar to the hyperlink network, which is binary, random-walk-based, and considers the transitive effect. As the PR score also represents the probability of the surfer stopping at the correspondence node after enough random-walk iteration, the PR score can also be understood as the probability of people or resources stopping at the correspondence place to that of a long random walk within the network. PR has been used in studies to analyze geospatial networks (Jiang, 2009; Jiang *et al.*, 2009; and Jiang & Jia, 2011). An analysis of the street-street topological network with PR and other network metrics found that the PR can predict human movement better than the centrality metrics, but WPR is the best indicator.

2.3.2. Weighted PageRank (WPR)

WPR is a modified version of PR. When sending scores to outgoing link-neighbors, PR can differentiate only the importance of the source nodes. PR cannot differentiate the attractiveness of each node because in the proportion equation of the original PR, the PR score of a source node is equally divided by the number of its outgoing link-neighbors to yield the sending probability for each outgoing link-neighbor. Therefore, WPR is introduced (Xing & Ghorbani, 2004).

Xing and Ghorbani (2004) thought that the attractiveness of the target node should influence the proportion of PR score that will be transferred from the source node and proposed the WPR, using the number of incoming links (L_{in}) to represent the attractiveness and dividing the proportion equation accordingly (Equation 8).

The proportion equation in WPR considers the characteristics of the target node rather than only the source. The proportion ratio is the sending proportion from the sources to their targets and acts as the transition probability in the Markov-chain process. The equation divides the number of incoming links of node *i*, which is an outgoing link-neighbor of node *j*, by the total number of incoming links of the all outgoing link-neighbors of node *j* to decide the proportion of the WPR score node *j* would send to node *i*.

In the random-surfer simulation, the WPR algorithm assumes that the random surfers would select the webpage with relatively greater attractiveness. This change in behavior would also change the probability of each webpage being reached, as shown by the WPR proportion equation (Equation 8).

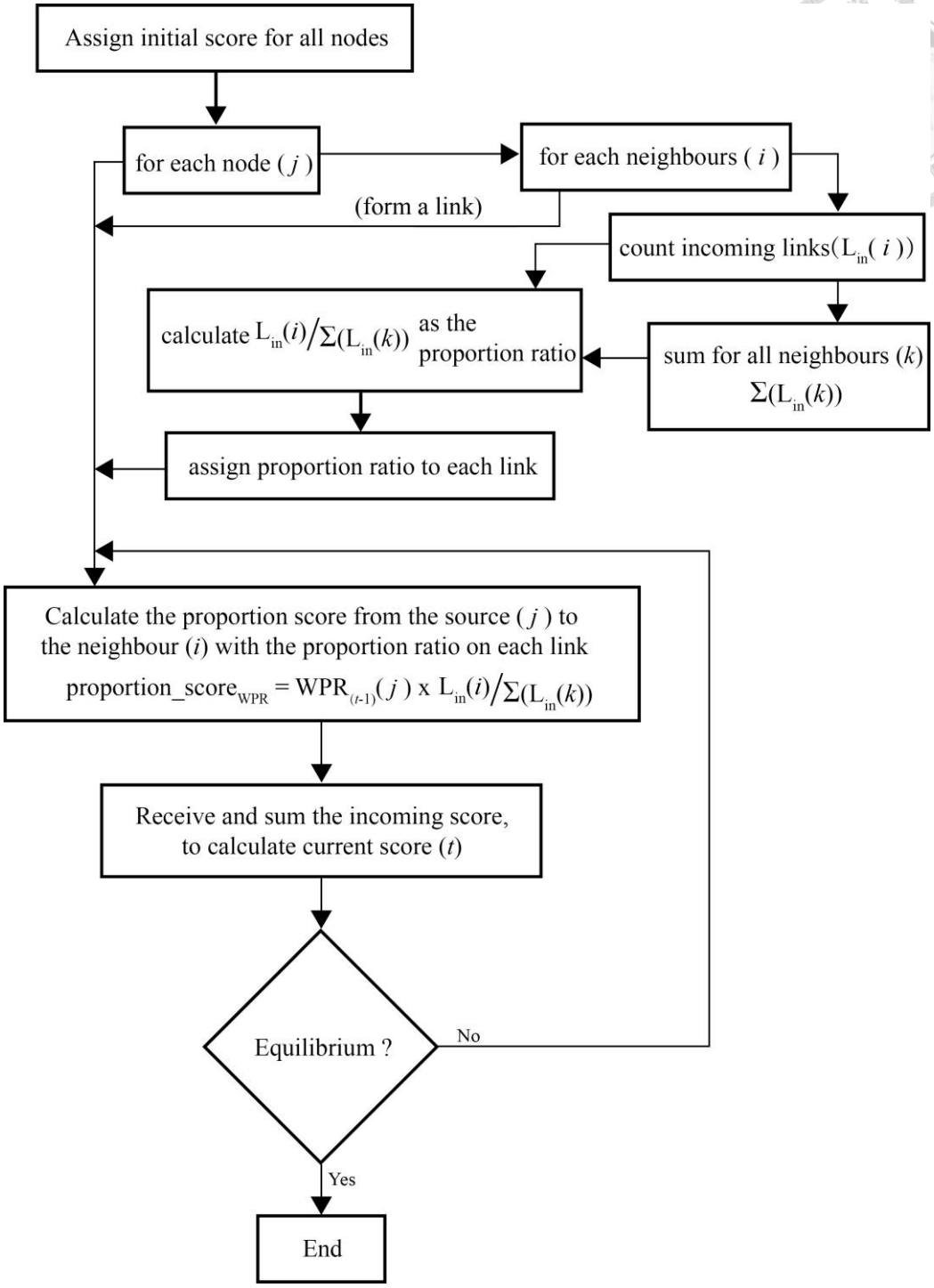


Figure 2.6 The calculation procedure of WPR algorithm.

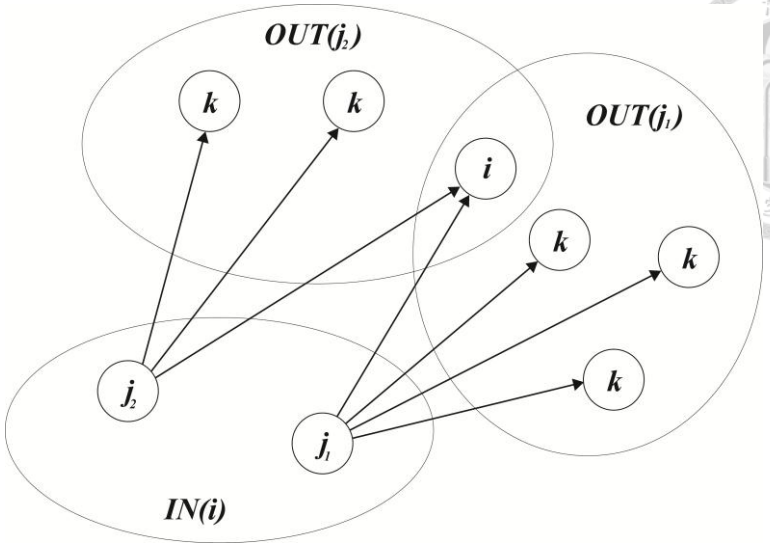


Figure 2.7 An illustration of the relationships between nodes.

The nodes in the set IN(*i*) (the nodes of *j*) are the supplier of scores for the node *i*, whereas the other nodes in the set OUT(*j*) (the nodes of *k*) are the competitors of the node *i*.

For $j \in \mathrm{IN}(i)$, $i \in k \in \mathrm{OUT}(j)$ (as shown in Figure 2.7), the proportion equation of WPR is shown below:

$$\mathrm{Prop}_{WPR} = \frac{L_{in}(i)}{\sum L_{in}(k)}$$

Equation 8

Where:

$L_{in}(i)$: indegree of node *i*;

$\sum L_{in}(k)$: sum of indegree of all node in set $\mathrm{OUT}(j)$;

$\mathrm{OUT}(j)$: The set of outgoing link-neighbours of the node *j*.

For example, in Figure 2.5, node 2 will send its score to node 1, node 3, and node 5; the calculation process is shown in Table 2.3:

1. Assign an initial score for all nodes.
2. For node 2 sending a score out: the number of incoming links for node 1 is 2; for node 3, 3; for node 5, 5; so, the sum is 7. Then, node 1 would get $\frac{2}{7}\times$ PR(node 2) from node 2; node 3 would get $\frac{3}{7}\times$ WPR(node 2) from node 2, and node 5 would get $\frac{2}{7}\times$ WPR(node 2) from node 2.
3. Node 5 would receive a score from node 2 and node 3. The proportion from node 2 is shown above, which is $\frac{2}{7}\times$ WPR(node 2). By the same calculations, node 5 would get $\frac{2}{4}\times$ WPR(node 3) from node 3.
4. Therefore, the PR score of node 5 at the end of this iteration is equal to $=$ $\frac{2}{7}WPR(node\ 2)+\frac{2}{4}\ WPR(node\ 3)$.

Table 2.3 A demonstration of the calculation for the WPR proportion.

Link *j*-*i*	$L_{in}(i)$	$\sum L_{in}(k)$	$\frac{L_{in}(i)}{\sum L_{in}(k)}$
2-1	2	7	$^2/_7$
2-3	3		$^3/_7$
2-5	2		$^2/_7$
3-1	2	4	$^2/_4$
3-5	2		$^2/_4$

Technically, WPR changes only the proportion equation or sending proportion (or transition probability in the Markov-chain process). This change would not affect the structure of the PR algorithm; it is still a Markov-chain process and captures the attractive effect without collapsing the algorithm. In other words, most of the properties of the original PR can be applied to the WPR.

That WPR changes the weight without changing the whole structure of the PR algorithm is an important inspiration for this study, in which the weighting captures different effects of the geospatial network.

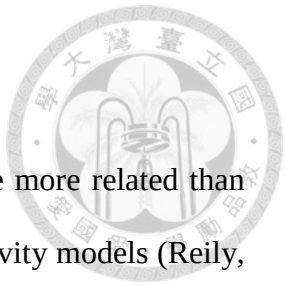
The difference between PR and WPR is that the WPR emphasizes the difference in the attractiveness between the outgoing link-neighbors, capturing the network's transitive effect and emphasizing the importance of nodes with greater attractiveness. More attractive nodes get high scores from source nodes and, thus, become more important.

Jiang (2009) used PR and WPR to analyze the street-street topology network and proved that WPR has a stronger correlation with human movement. Because PR and WPR focus only on the network's topological structures, they cannot capture the differences between incoming link-neighbors at different distances.

2.3.3. Summary

PR and WPR are both methods of analyzing the network's topological structure. To capture the transitive effect, PR uses an iterative voting process to differentiate the source node's importance but does not calculate the attractiveness of each target node. WPR uses a weighted proportion ratio calculated based on the in-degree centrality of each target node. The overall structure does not change. The advantages of the PR algorithm, including the ability to capture the random-surfer concept, the transitive effect, the Markov-chain process that brings the score distribution to equilibrium, and the constant total score of all nodes, are retained in the WPR algorithm.

3. Geographically modified PageRank algorithms



"Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970) This concept is observed from the gravity models (Reily, 1931; Stewart, 1950; Tinbergen, 1963) of the distance-decay effect in the interaction between spaces. To analyze a geospatial network, the PR is modified to include the distance-decay weight to produce the IDPR; the WPR is modified to include distance decay (a gravity model) to produce the GPR.

3.1. Inverse-Distance PageRank (IDPR)

To capture the distance-decay effect in PR, this study considers a distance-decay function as the sending proportion, and develops a modified version of PR, namely Inverse-Distance PageRank. In IDPR, the proportion equation differentiates the geographical distance (geometric distance) from the source node (node *j*) to the given target node (node *i*), compared to other target nodes (nodes-*k*) from the same source node (node *j*). Thus, the source node can send its score out to the target nodes according to the geographical proximity level -- the nearer the distance, the higher the proportion.

In terms of random-surfer simulation, in a geospatial network with the nodes located at fixed points in a geographical space, the IDPR assumes the random surfers would select the nearer neighbors with a higher probability and that the farther away the neighbors are, the lower the probability of being chosen. According to the distance-decay effect on a person moving between places (nodes), the random surfers would have higher probabilities of choosing a relatively nearer place to visit rather than selecting a place randomly without considering the distance.

The modification in the proportion equation of IDPR (Equation 9) is like WPR but with geographical properties (Figure 3.1): the equation compares $D_{ji}{}^{-\beta}$, the inverse distance from the source node (*j*) to the target node (*i*), to the sum of the inverse distance from the source node (*j*) to all outgoing neighbors (*k*), $\sum(D_{jk}{}^{-\beta})$. Before calculating the proportion equation, a distance factor (beta), the power of the inverse distance, must be assigned. Moreover, this power would affect the distance-decay curve in the interaction between the nodes. With this proportion equation, more PR scores will be sent to the nearer nodes, and fewer PR scores will be sent to the distant nodes. Therefore, the interaction between shorter links is emphasized.

For $j \in \mathrm{IN}(i)$, $i \in k \in \mathrm{OUT}(j)$ (as shown in Figure 2.7), the proportion equation of IDPR is shown below.

$$\mathrm{Prop}_{IDPR} = \frac{D_{ji}{}^{-\beta}}{\sum(D_{jk}{}^{-\beta})} \ , i \in k \ , k \in \mathrm{OUT}(j)$$

Equation 9

Where:

$D_{ji}{}^{-\beta}$: inverse geographical distance between node *j-i* to the power of beta;

$\sum(D_{jk}{}^{-\beta})$: sum of inverse geographical distance between node *j* and all node *k* to the power of beta.

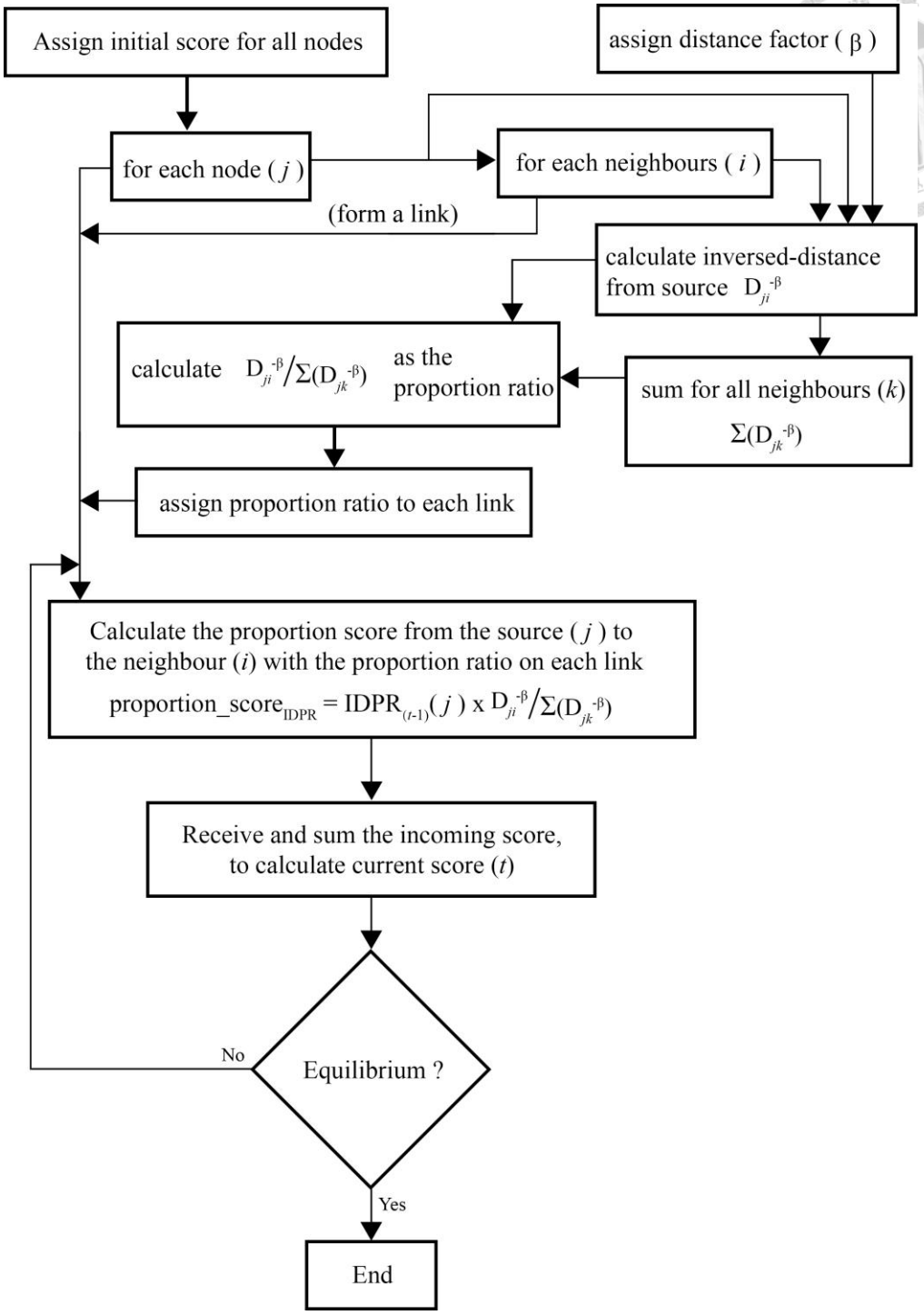


Figure 3.1 The calculation procedure of IDPR algorithm.

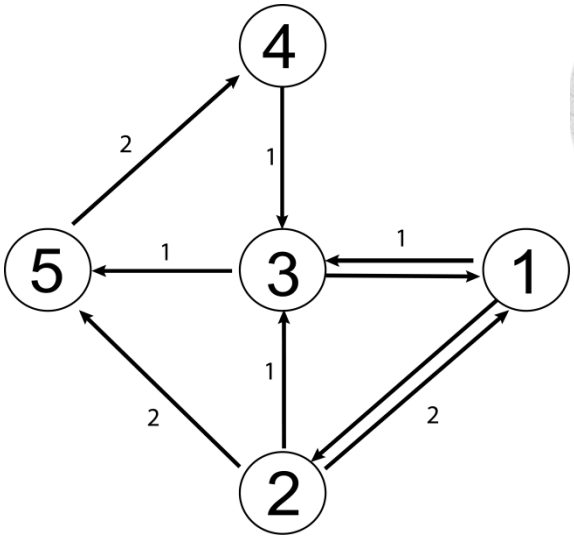


Figure 3.2 An example geospatial network with geographical distance (meter)

In Figure 3.2, node 2 will send its score to node 1, node 3, and node 5, and node 5 would receive a score from node 2 and node 3; the calculation of the IDPR proportion is shown in Table 3.1. The results follow:

1. Assign initial scores for all nodes.
2. The inverse distance for link 2-1 is equal to $^{1}/_{2\ m}$; that for link 2-3 is $^{1}/_{1\ m}$; and that for link 2-5 is $^{1}/_{2\ m}$, so the sum is $^{4}/_{2\ m}$.
3. Then, node 1 would get $^{2}/_{8} \times \text{IDPR(node 2)}$ from node 2; node 3 would get $^{4}/_{8} \times \text{IDPR(node 2)}$ from node 2, and node 5 would get $^{2}/_{8} \times \text{IDPR(node 2)}$ from node 2.
4. Node 5 would receive scores from node 2 and node 3. The proportion from node 2 is equal to $^{4}/_{8} \times \text{IDPR(node 2)}$. By the same process, node 5 would get $^{1}/_{2} \times \text{IDPR(node 3)}$ from node 3.
5. Therefore, the PR score of node 5 at the end of this iteration is equal to $^{4}/_{8} \times \text{IDPR(node 2)} + {}^{1}/_{2} \times \text{IDPR(node 3)}$.

Table 3.1 A demonstration of the calculation for the IDPR proportion ($beta = 1$).

Link *j*-*i*	$D_{ji}{}^{-1}$	$\sum(D_{jk}{}^{-1})$	$\frac{D_{ji}{}^{-1}}{\sum(D_{jk}{}^{-1})}$
2-1	$\frac{1}{(2\,m)}$	$\frac{4}{(2\,m)}$	$\frac{1/2}{4/2}=\frac{2}{8}$
2-3	$\frac{1}{(1\,m)}$		$\frac{2/2}{4/2}=\frac{4}{8}$
2-5	$\frac{1}{(2\,m)}$		$\frac{1/2}{4/2}=\frac{2}{8}$
3-1	$\frac{1}{(1\,m)}$	$\frac{2}{(1\,m)}$	$\frac{1/1}{2/1}=\frac{1}{2}$
3-5	$\frac{1}{(1\,m)}$		$\frac{1/1}{2/1}=\frac{1}{2}$

IDPR considers the inverse-distance (D^{-beta}) function as a distance-decay curve. The distance factor (beta) would influence the proportion allotted each target node. Theoretically, the distance-decay curve represents how the relative importance of the interaction changes with distance (the impedance effect of distance). Therefore, the beta should be chosen to fit with the network type. In transportation studies, there are regulations about the selection of distance factor.

One of the key features of the PR is the way it assigned the proportion of the score to a node's outgoing link-neighbors. In the IDPR proposed above, the nearer the neighbor is, the greater the proportion it has. Therefore, it emphasizes the interaction between near nodes in local clusters.

3.2. Geographical PageRank (GPR)

The modified GPR is based on gravity models, also called spatial-interaction models. Based on this concept, the relative importance of interaction between the source node and target node should be proportional to the attraction between them and inversely proportional to the distance between them. Therefore, the gravity model (GPR) is an integration of the attractive effect and distance-decay effect (WPR and IDPR).

Therefore, to integrate attractiveness and the distance-decay effect, this study integrates the attractiveness variable ($L_{in}(i)$ and $L_{in}(k)$) in WPR and inverse distance ($D_{ji}{}^{-\beta}$ and $D_{jk}{}^{-\beta}$) in IDPR to capture the gravity model and obtain the geographically modified proportion equation (Equation 10). As in IDPR, a distance factor must be assigned. With this proportion equation, the difference in the attractiveness of the target node and the distance-decay effect on the interaction are integrated, so the nearer and more attractive target node would get a higher PR score, while the more distant and less attractive target node would get a lower PR score. The calculation integrates the WPR and IDPR (Figure 3.3), considering both the topological characteristic of the neighbors and the distance between the source (j) and the target (i).

In terms of the random-surfer simulation, GPR can be understood to change the behavior of random surfers selecting the target. According to the gravity model, the gravitational force across shorter distances and to higher-mass things should be greater. For the geographical random surfer, the probability of choosing a nearer and more attractive place should be higher. Therefore, the proportion allotted each link should be negatively proportional to its length and positively proportional to the attractiveness of the target node, as in the proportion equation of the novel GPR.

For $j \in \mathrm{IN}(i)$, $i \in k \in \mathrm{OUT}(j)$ (as shown in Figure 2.7), the proportion of the equation for GPR is shown below.

$$\mathrm{Prop}_{GPR} = \frac{G(j,i)}{\sum(G(j,k))} = \frac{{L_{in}(i)}/{D_{ji}{}^{\beta}}}{\sum({L_{in}(k)}/{D_{jk}{}^{\beta}})} \ , i \in k \ , k \in \mathrm{OUT}(j)$$

Equation 10

Where:

$G(j,i)$: gravity between node *j*-*i* ;

$\sum(G(j,k))$: sum of gravity between node *j* to all node in set $\mathrm{OUT}(j)$;

${L_{in}(x)}/{D_{jx}{}^{\beta}}$: incoming link-neighbours of a given node x divided by the geographical distance between node *j* to node *x* to the power of beta.

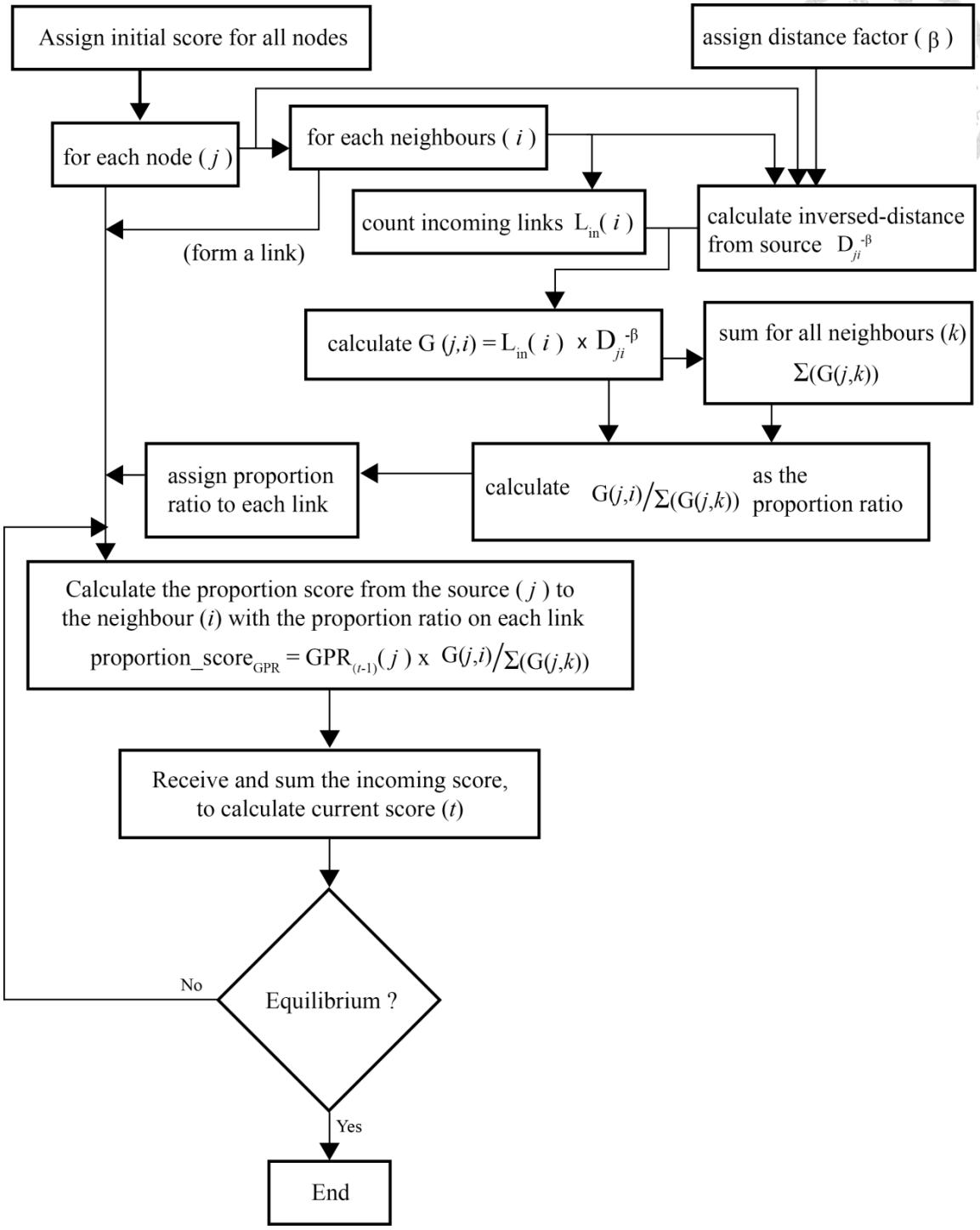


Figure 3.3 The calculation procedure of GPR algorithm.

Table 3.2 A demonstration of the calculation for the GPR proportion ($beta = 1$).

Link *j*-*i*	$L_{in}(i)$	${D_{ji}}^{-1}$	$G(j,i) = L_{in}(i)/D_{ji}$	$\sum(G(j,k)) = \sum(L_{in}(k)/D_{jk})$	$\frac{G(j,i)}{\sum(G(j,k))}$
2-1	2	$\frac{1}{(2\,m)}$	$\frac{2}{(2\,m)}$	$\frac{10}{(2\,m)}$	$\frac{2/2}{10/2} = \frac{4}{20}$
2-3	3	$\frac{1}{(1\,m)}$	$\frac{3}{(1\,m)}$		$\frac{3/1}{10/2} = \frac{6}{10}$
2-5	2	$\frac{1}{(2\,m)}$	$\frac{2}{(2\,m)}$		$\frac{2/2}{10/2} = \frac{4}{20}$
3-1	2	$\frac{1}{(1\,m)}$	$\frac{2}{(1\,m)}$	$\frac{4}{(1\,m)}$	$\frac{2/1}{4/1} = \frac{2}{4}$
3-5	2	$\frac{1}{(1\,m)}$	$\frac{2}{(1\,m)}$		$\frac{2/1}{4/1} = \frac{2}{4}$

For example in Figure 3.2, where node 2 will send its score to node 1, node 3, and node 5, and node 5 would receive score from node 2 and node 3, the calculation of the GPR proportion is shown Table 3.2. The calculation process follows:

1. Assign initial scores for all nodes.
2. For node 2, by combining the results from Table 2.3 and Table 3.1, the gravity between link 2-1 is found to be $^{2}/_{2\,m}$, link 2-3 is $^{3}/_{1\,m}$, and link 2-5 is $^{2}/_{2\,m}$. Therefore, the total is $^{10}/_{2\,m}$.
3. From node 2, node 1 would get $^{4}/_{20} \times GPR(\text{node 2})$; node 3 would get $^{6}/_{10} \times GPR(\text{node 2})$; and node 5 would get $^{4}/_{20} \times GPR(\text{node 2})$.
4. For node 5, it would receive scores from node 2 and node 3. The proportion from node 2 is shown above. By the same calculation, node 5 would get $^{2}/_{4} \times GPR(\text{node 3})$ from node 3.

5. Therefore, the PR score of node 5 at the end of this iteration is equal to $= {}^{4}/_{20}\, GPR(node\ 2) + {}^{2}/_{4}\ GPR(node\ 3)$.

Formerly, the gravity model was a way to measure the relative importance of the relationships between places. GPR uses the gravity model to differentiate the relative importance of relationships between the target nodes of a given source node. This study assumes the gravity model applies to network interactions; thus, the flows of importance should be proportional to the gravity between the nodes.

3.3. Summary

To capture the distance-decay effect in interactions within the network, this study proposes two algorithms, IDPR and GPR. Both retain the advantages of the PR algorithm as the WPR did, including the random-surfer concept, the ability to capture the transitive effect, the calculation properties of the Markov-chain process that brings the score distribution to equilibrium, and the constant total score of all nodes because both the IDPR and GPR change the way the proportion ratio is assigned to the target.

By integrating the distance-decay function into the PR calculation, the distance between the nodes would influence the proportion ratio, and this implies that the interaction between the nodes would decay with distance. On the other hand, GPR changes the proportion equation to capture the gravity model, which is also called the spatial-interaction model. By integrating the attractiveness function and the distance-decay function, the target node's attractiveness and the distance between the target node and source node are considered, implying that the interaction between the nodes decays with distance and increases with attractiveness.

4. Experiment 1 – Inter-city network

4.1. Preface

To test the four PageRank algorithms (PRs), this study tested their ability to identify the major cities from a transportation network. The background and assumptions are shown in Figure 4.1. Accessibility has a long tradition in the transport-economics literature and is defined as the potential for spatial interaction (Reggiani et al., 2011). Recent studies used street networks to calculate network centrality to represent accessibility (Jiang, 2009; Wang et al., 2011, Reggiani et al., 2011), and concluded that street networks can capture accessibility. Furthermore, they found that the connectivity of transportation networks is correlated with human activity; for example, Jiang (2009) found that human movement is correlated with the connectivity of the street network (by using the Weighted PageRank as the network's connectivity index); Wang et al. (2011) found that the land-use intensity (calculated by the local employment density and local population density) is correlated with the street centrality. Transportation networks and the intensity of human activity are correlated, which means it is possible to locate the places with high-intensity human activity by analyzing the transportation network.

In this test, this study conceptualized the accessibility of transportation network as the importance factor and defined the major cities as the places with high importance. Thus, this study intended to locate the places with high importance from network connectivity via the PRs.

From the above description, the demography data is a way to represent the intensity of human activity. For example, Wang et al. (2011) used demography data (the

urbanization level) to represent the land-use density (the intensity of human activity). Because the urbanization data can also be used to locate major cities, this study used the urbanization data to check the results of the importance calculated from the network.

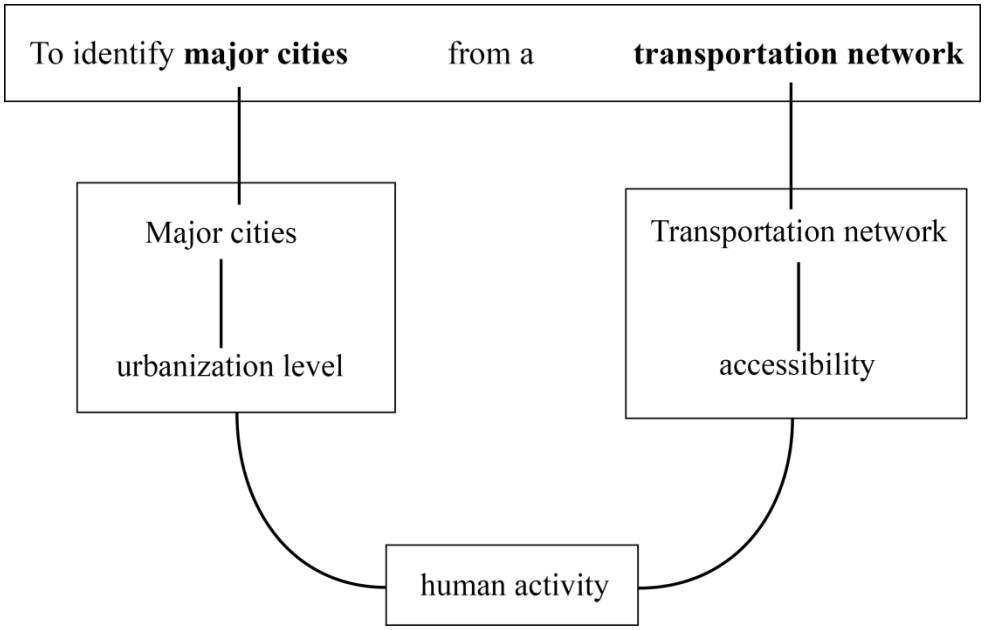


Figure 4.1 The background of using transportation network to identify the major cities.

In summary, this experiment is designed to locate major cities based only on the ground-transportation data, using an inter-city network developed from Taiwan's ground-transportation data, including roads and railway data, to identify major cities based on the above considerations. Then, the correlation between the importance (calculated from the network indices) and the intensities data were calculated by Spearman rank correlation to check the ability of the network indices to predict the urbanization level. Finally, a sensitivity analysis checked the sensitivity of three key parameters.

4.2. Datasets preparation

4.2.1. Geospatial network

This study developed a geospatial network based on the Taiwanese roads layer, in which the nodes were cities and the links were the 1-hour transportation connectivity (the city networks). This study defined the cities as the mean centers of the spaces where people live. For example, a city includes the places where people go to live, work, study, eat, and go other places. Because this definition is inadequate for administrative regions, they were defined by several criteria, including urban planning, historical outcomes, and management needs. Therefore, the centroid area of the administrative regions does not capture where people live. Because the location of nodes in a geospatial network should represent the center of interaction with other places, which is the basis for calculating the relative distance from other cities, this study did not use the centroids of the administrative regions as the nodes in our city networks. To show where human activity is, this study calculated the mean centers for the road junctions extracted from the roads layer[1]. This study used k-mean clustering to calculate 500 mean centers from 391,446 road junctions, and these mean centers (the location of the mean centers) were the nodes in the geospatial network.

Links represented the connectivity between nodes. Batten (1995) suggested that the average urban travelling time is about one hour per day and remained unchanged from 1800 to 1995. This study used one hour as the threshold to decide whether the links should be connected or removed. To do this, this study calculated the time needed

[1] The roads layer data was extracted in year 2007, which include general roads, county highways, and

to travel from each node to every other node with the underlying road layers[2] and railway layer [3](Figure 4.2) at different speeds (Table 4.1).

Table 4.1 Speed settings of road layers and railway layer

Layer	speed (km/hr)
Road (general and county)	20
Highway	60
Railway	90

Using network-analysis tools, this study calculated the time needed to travel between nodes, with the speed and the distance from the underlying transportation layers. Then, this study removed links for which the traveling time is longer than 1 hour. Finally, a transportation network is constructed in which the nodes are cities and the links have travelling times shorter than 1 hour.

4.2.2. Data of intensity of human activity

From the assumptions for this test as mentioned above, this study used the data of intensity of human activity for representing the urbanization level. The intensity of human activity relative importance data were collected from the real world, which was prepared for validation and is shown in section 4.4. To represent the intensity of human activity, this study used two different sets of data, including the demographic data and the human-flows data. For the demographic data, this study used the township-level population data from 2005 [4]and the population density based on the population and the township's area. Because each measure captures a different aspect of urbanization, this

[2] Same as the roads layer that was used to locate the nodes' locations.
[3] The railway layer data was extracted in year 2008.
[4] The population data was got from the website of Directorate-General of Budget, Accounting and Statistics, Executive Yuan, R.O.C. (Taiwan).

study included both in the validation. For the human-flows data (IOT, 2009), this study used the inter-township daily car-flows data[5], which is a OD-matrix of inter-township car flow (per day) from 2005. This study processed the OD-matrix data into township-level data, including total flow (sum of the incoming and outgoing flows for each township), inflow (sum of the incoming flows for each township), and flow betweenness centrality (using flow betweenness centrality to conceptualize the car-flow data into the minimum requirement of each township to reach the maximum flow between nodes).

Using these 5 characteristics, this study tries to capture the spatial distribution of the intensity of human activity in the spatial pattern of urbanization. This study used these data for validation.

4.3. Data visualization

4.3.1. Geospatial network

In this part, this study observed the geospatial network and the underlying landscape. Railways were built according to the landscape barriers, which were also surrounding the whole Taiwan (Figure 4.2). There are three things should be mentioned: first, at the middle west part of Taiwan (Miaoli County (苗栗縣) and Taichung City (台中市)), there are two railway surrounding the area; second, at the east coast of Taiwan (Hualian county (花蓮縣) and Taitung county (台東縣)), the railway goes between the Central Mountain Range (中央山脈) and the Coastal Range (海岸山脈), and the coast side was separated from the East Rift Valley (花東縱谷) by the Coastal Range; third,

[5] The human flow data was got from the Institute of Transportation, Ministry of Transportation and Communications.

the railway did not go to the Hengchun Peninsula (恆春半島), but it connect the east coast and the west coast at the north side of the peninsula. This section examines the geospatial network and the underlying landscape. Railways were built around the landscape barriers, which also surround Taiwan (Figure 4.2). Three things should be noted: first, midwestern Taiwan (Miaoli County and Taichung City) has two surrounding railways; second, on the east coast of Taiwan (Hualian county and Taitung county), the railway goes between the Central Mountain Range and the Coastal Range, and the coast is separated from the East Rift Valley by the Coastal Range; third, the railway does not go to the Hengchun Peninsula but connects the east coast and the west coast at the north side of the peninsula.

Figure 4.3 shows the distribution of nodes (cities; Figure 4.3a) and how the nodes are distributed over the road system (Figure 4.3b, Figure 4.3c) at a smaller scale. The nodes are distributed around natural landscape barriers (Figure 4.2, Figure 4.3a), which is also a restriction for road junctions. For example, the nodes were separated by the Central Mountain Range in the middle of Taiwan, by the Coastal Range in the east, and by the Tatun volcano group (大屯火山群) in the north. The nodes are also distributed around road junctions, where the road junctions cluster (Figure 4.3b for Yilan county (宜蘭縣), Figure 4.3c for Taichung city and Changhua county (彰化縣)).

Because of the landscape barriers, the links do not cross the Central Mountain Range and are distributed around Taiwan, like the railway network (Figure 4.4). Because the road network and the railway network differ in shape and travelling speed, the links of different types differ in length. Figure 4.4 shows one-hour links between nodes. Short links, which are established in street networks, are distributed regularly between close nodes. Because the railway is relatively straight and its travelling speed is

relatively fast, longer links between farther nodes can be constructed in the same time threshold. Therefore, nodes that are close to railway stations can connect with each other even if they are far from each other. As a result, those concentrated links are distributed between places with railway stations. Because Miaoli County and Taichung City are surrounded by railways (Figure 4.2), the nodes there have better connectivity. In contrast, the connectivity at Hengchun Peninsula and the coast side of the Coastal Range is relatively weak.

Traditional social-network indicators were computed. This network had an average degree of 27.5, a diameter of 12 steps, and an average shortest path length of 4.3 steps, which means that any node in this network (a total of 487 nodes; 13 nodes were removed because they did not connect with any other nodes) can reach any other node within 12 steps, and the average path length between any 2 nodes is 4.3 steps. These 2 sets of social-network indicators imply the existence of long links that connect sub-networks. These long links also form hubs on nodes near railway stations, making the world “smaller”.

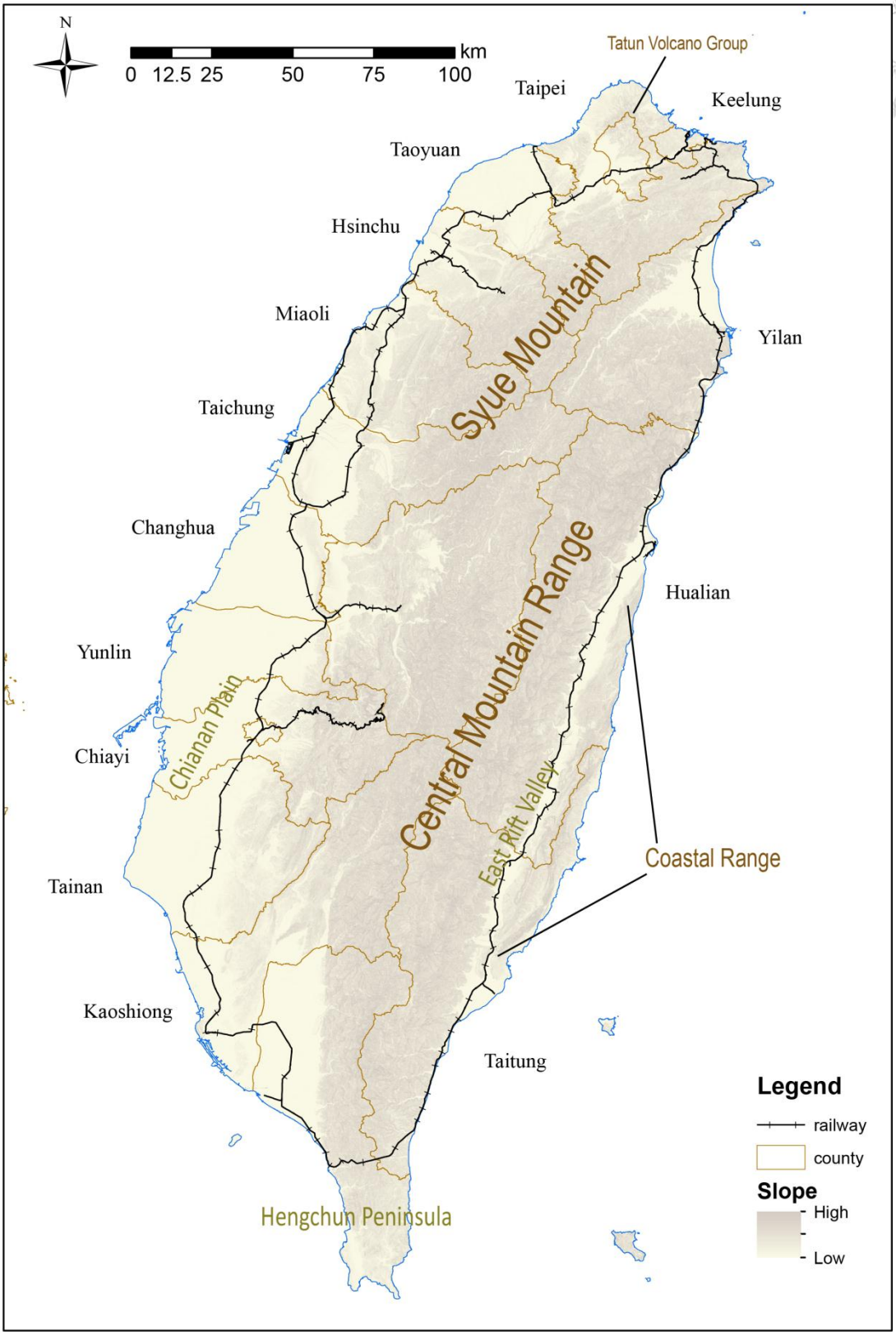


Figure 4.2 Digital Terain Map (slope) of Taiwan and the railway network

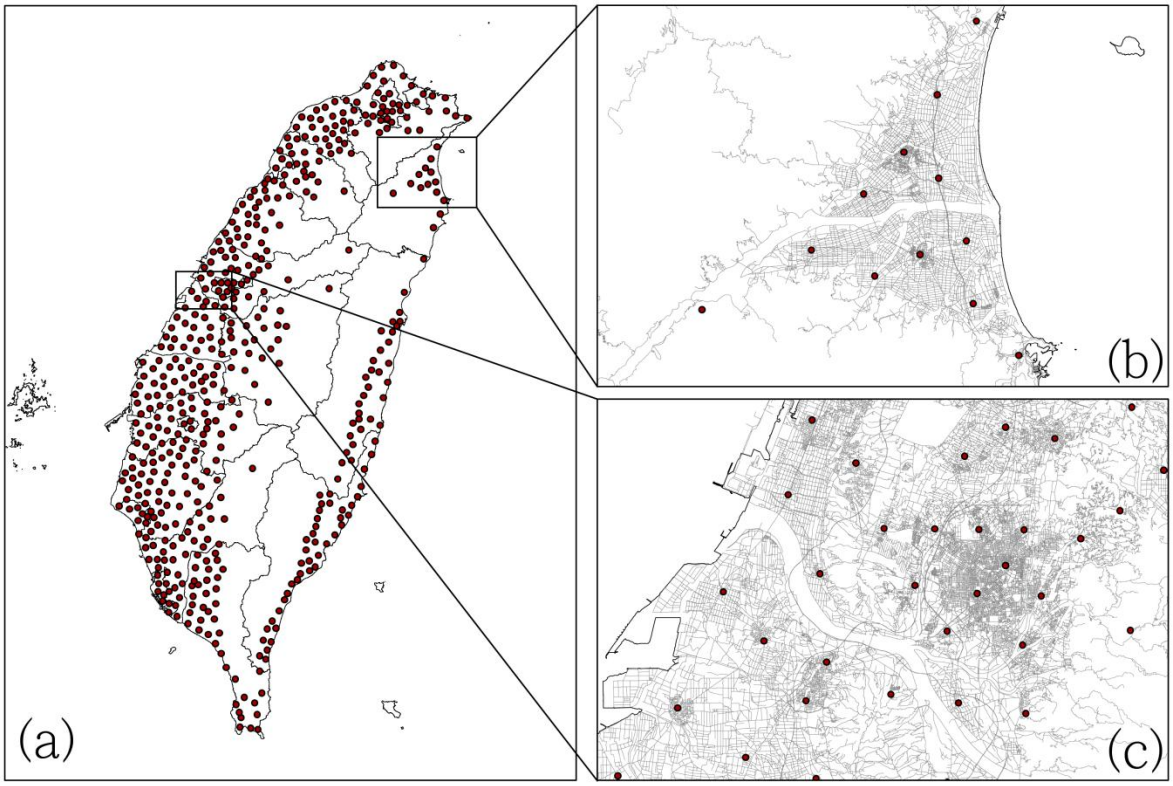


Figure 4.3 Nodes distribution, (a) study area, (b) Yilan and (c) Taichung

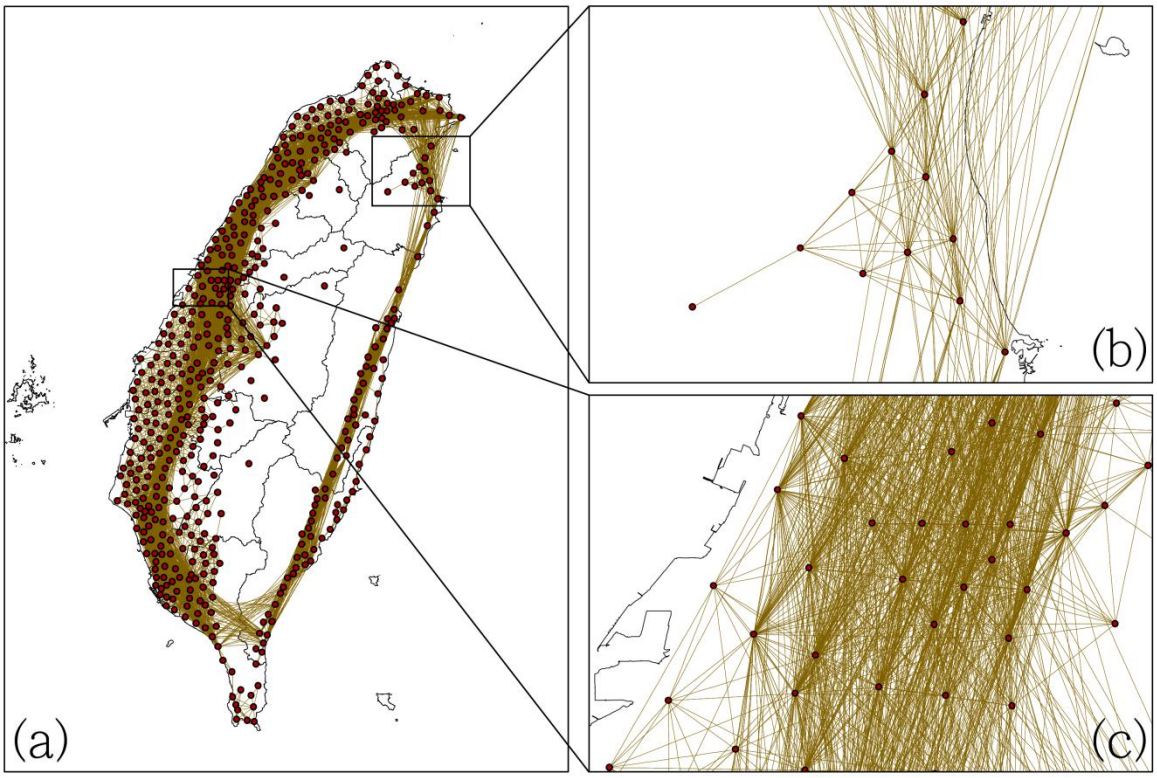


Figure 4.4 One hour links distribution, (a) study area, (b) Yilan and (c) Taichung

4.3.2. Spatial distribution of the intensity of human activity

In this part, this study observed the spatial distribution of urbanization data to elucidate the characteristics of each city at different intensities of human activity (Figure 4.5, Figure 4.6, Figure 4.7, Figure 4.8, and Figure 4.9). Generally speaking, five of the urbanization levels had similar patterns. The area including Taipei city (台北市), New Taipei city (新北市), Taoyuan county (桃園縣) had high value in all indices. The middle area, Taichung city (台中市), includes several high-value townships in all indices. In southern Taiwan (including Tainan city (台南市) and Kaoshiung city (高雄市)), there are high-value townships in all indices. Three counties on the east coast (Yilan county (宜蘭縣), Hualian county (花蓮縣), and Taitung county (台東縣)) were relatively lower in all indices than were counties in the north and west.

The population-density data (Figure 4.6) show greater disparities than the population-size data because the population is more concentrated in the urban area, as in Taipei city, Taichung city center[6] (formerly Taichung city, 改制前台中市地區), and Tainan city center[7] (formerly Tainan city, 改制前台南市地區). The population-size data (Figure 4.5) were less concentrated than the population density data, in terms of the areas near to the city area or city center that were also considered high-value areas.

Within the human-flows data, the total flows (Figure 4.7) and the incoming flows data (Figure 4.8) looked alike in spatial distribution. This implies that the outgoing flows might be correlated with the incoming flows, which meant that the sum of the incoming and outgoing flows did not change the spatial distributions substantially. Although they looked alike, this study used both for validation because the total

[6] The area formerly named Taichung city before the City-County Consolidation in 2010.
[7] The area formerly named Tainan city before the City-County Consolidation in 2010.

represents the number of cars that passed and the incoming flow represents the attractiveness, while the flow betweenness centrality of the flow's OD-matrix data has a slightly different pattern (Figure 4.9). The greater value of flow betweenness centrality is observed in central and southern Taiwan; northern Taiwan also had a high value of this index, but the number of townships of high value is relatively lower than in central and southern Taiwan.

From these results, it is clear that the urbanization levels were influenced by the natural barrier of the Syue Mountains (雪山山脈) and Central Mountain Range (中央山脈), which separate the east coast (the 3 counties) from the west area (from Taipei city) to Kaoshiung city), as shown in Figure 4.2.

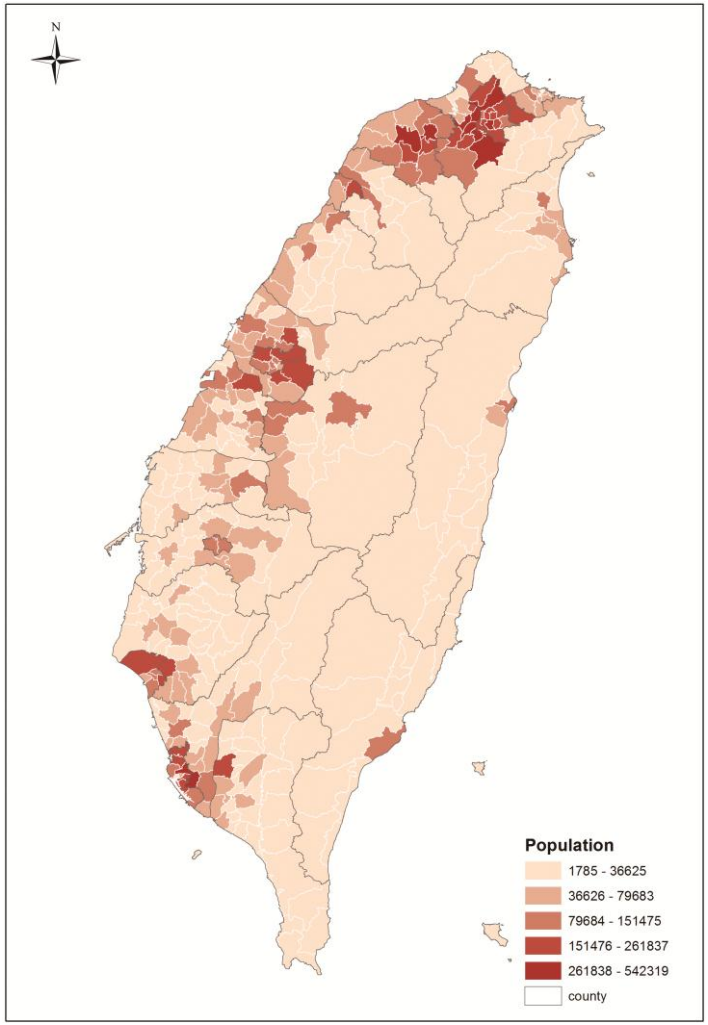


Figure 4.5 Spatial distribution of population size

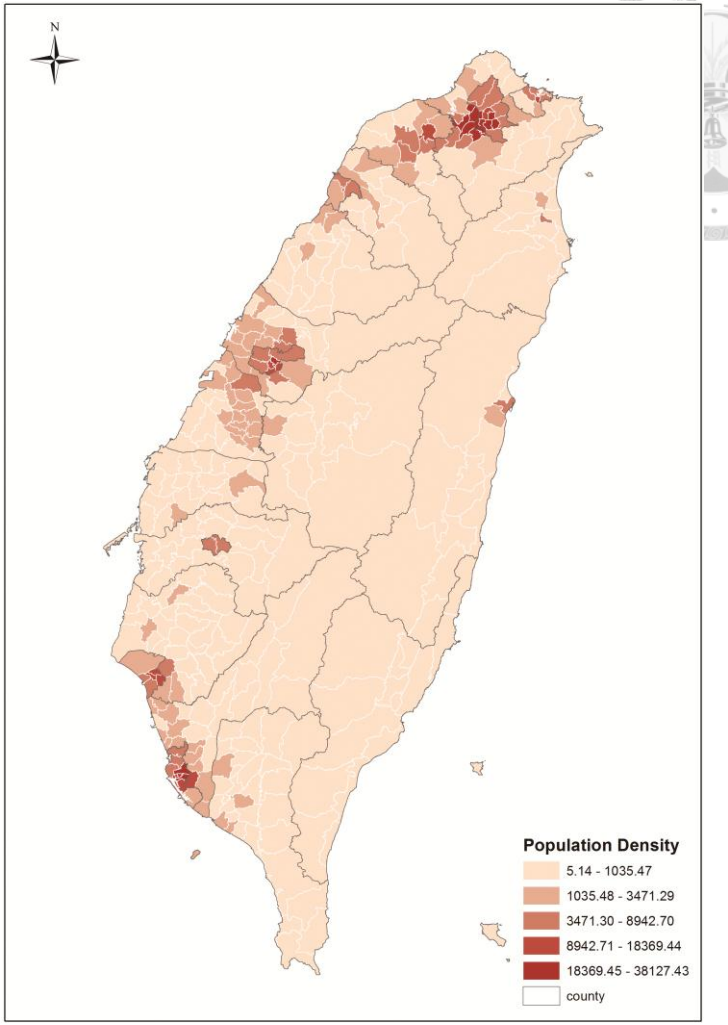


Figure 4.6 Spatial distribution of population density

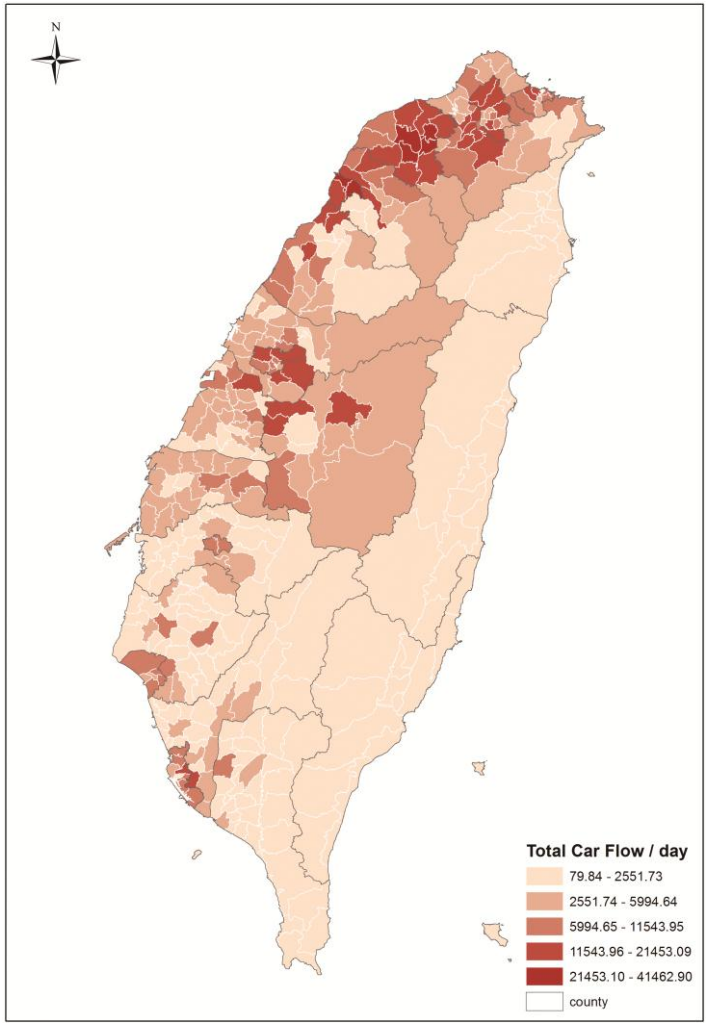


Figure 4.7 Spatial distribution of total car flow per day

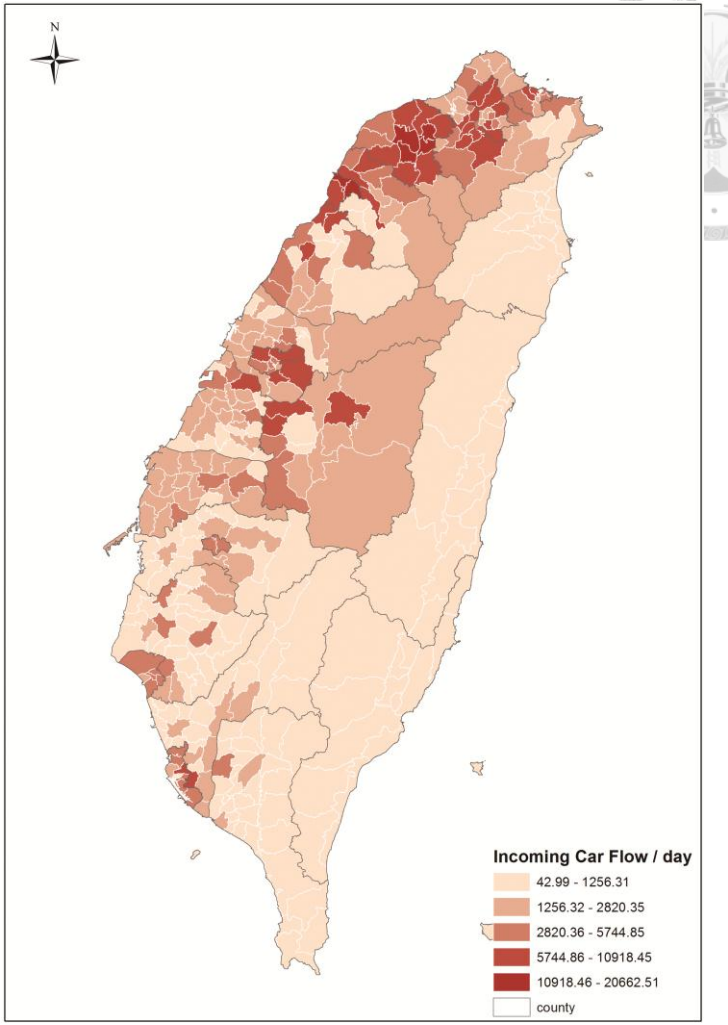


Figure 4.8 Spatial distribution of incoming car flow per day

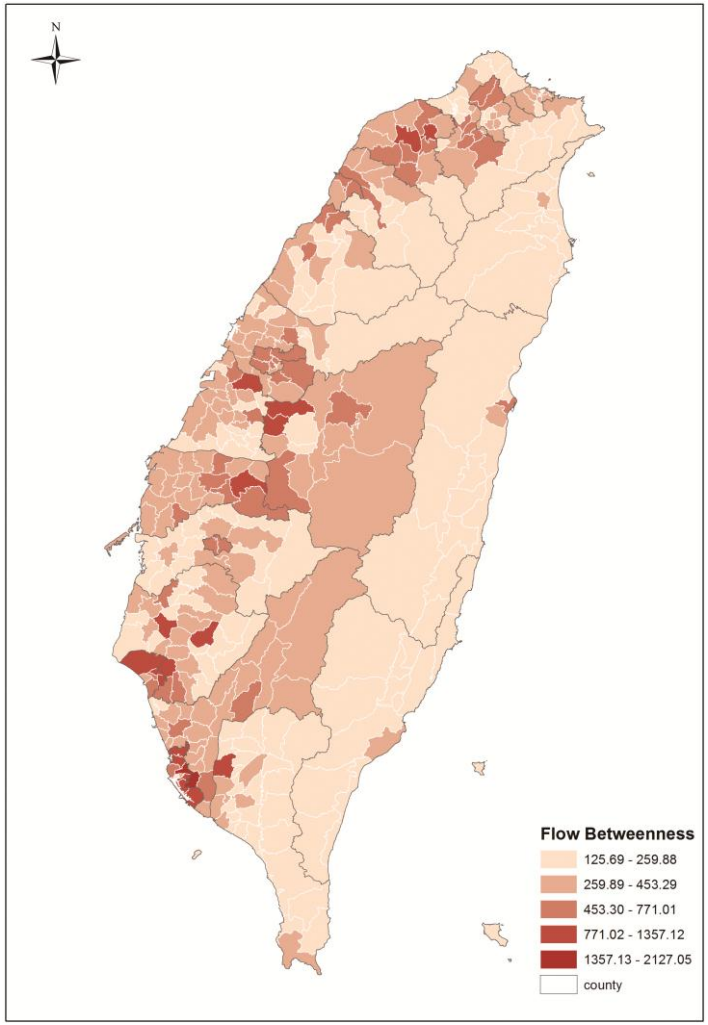


Figure 4.9 Spatial distribution of the car flow's flow betweenness

4.4. Calculation results

The frequency distributions differed for the four PR algorithms (Figure 4.10). The frequency distribution of WPR and GPR were scale-free, with many nodes scoring below 0.00145, and the number of nodes dropped dramatically as the score increased. On the other hand, the frequency distribution of PR and IDPR show that the numbers of nodes in different score categories were relatively balanced. This might be because the PR & IDPR consider only the transitive effect and the distance-decay effect, which do not influence the frequency distribution. The WPR and GPR's consideration of the attractive effect (emphasizing the differences in the target node) makes the frequency distribution more like a scale-free distribution.

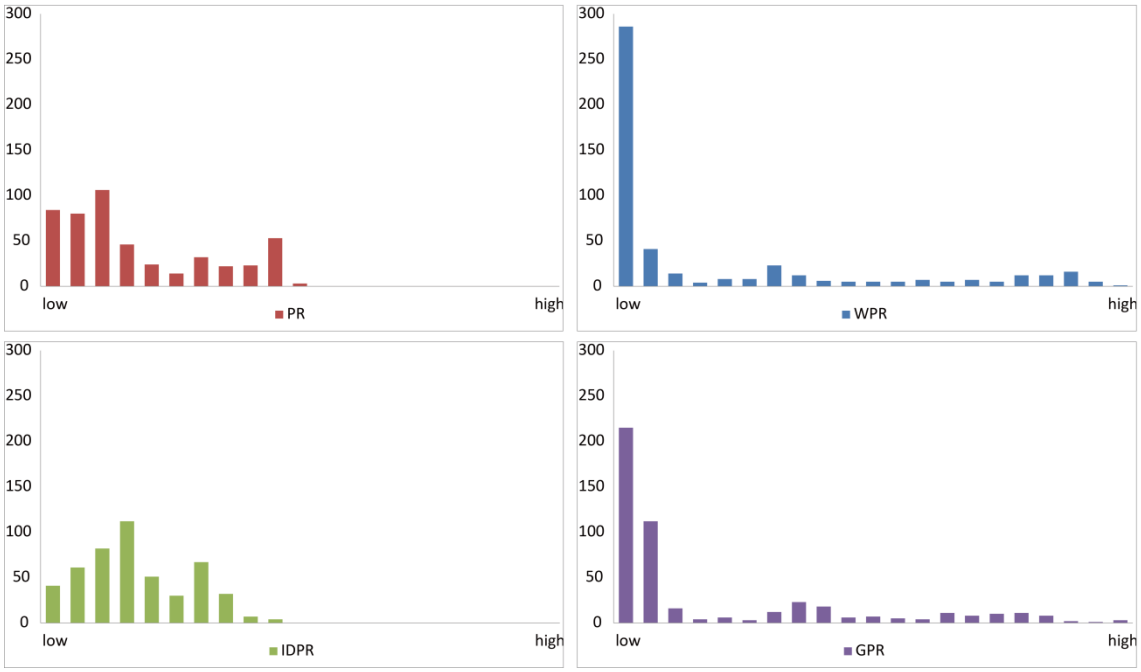


Figure 4.10 Frequency distribution of the PRs

horizontal: PR score, vertical: frequency

Although the frequency distribution of the PRs differed, they were ranked by score and grouped into three categories with similar spatial distributions. Figure 4.11,

Figure 4.12, Figure 4.13, and Figure 4.14 show results of the calculation of four PRs in three categories (by quantile): low (blue), medium (yellow), and high (red). There are common results for four of the PRs: the nodes in the high category were all located near the railway; as a result of the railway's distribution, a cluster of nodes with high PR scores is located in Miaoli County (苗栗縣) and Taichung City (台中市), and a cluster of nodes with low PR scores is located on Hengchun Peninsula (恆春半島); the nodes in the east of the Coastal Range (中央山脈) all had low values, as did the nodes on the west coast (coast of Changhua County (彰化縣), Yunlin County (雲林縣), Chiayi County (嘉義縣), and Tainan City (台南市)).

The result of the WPR (Figure 4.12) is similar to that of the PR (Figure 4.11) in most of the network. Significant differences between them were observed in the nodes on the west side of the Central Mountain Range, where the result of the Weighted PageRank changed gradually from high-value nodes at the railway to low-value nodes on the coast. Both PR and WPR yielded high-value nodes in the East Rift Valley (花東縱谷).

The result of the IDPR (Figure 4.13) differed from those of PR and WPR: high-value nodes were distant from the railway in Yunlin County; all of the nodes on the east coast (from Yilan County (宜蘭縣) to Taitung County (台東縣)) were low-value, including the nodes in East Rift Valley; the nodes in the midwest area (Chi-Chi railway line, 鐵路集集線) did not have high values.

The result of the GPR (Figure 4.14) is similar to the result of the WPR, including the gradual change on the west side of the Central Mountain Range.

Additionally, it is similar to the result of the IDPR on the east coast in that all nodes at the east were in the low-value category, including the nodes in the East Rift Valley.

This spatial distribution of the ranked score categories implies that the characteristics of the algorithms have been captured by the modified PR algorithms.

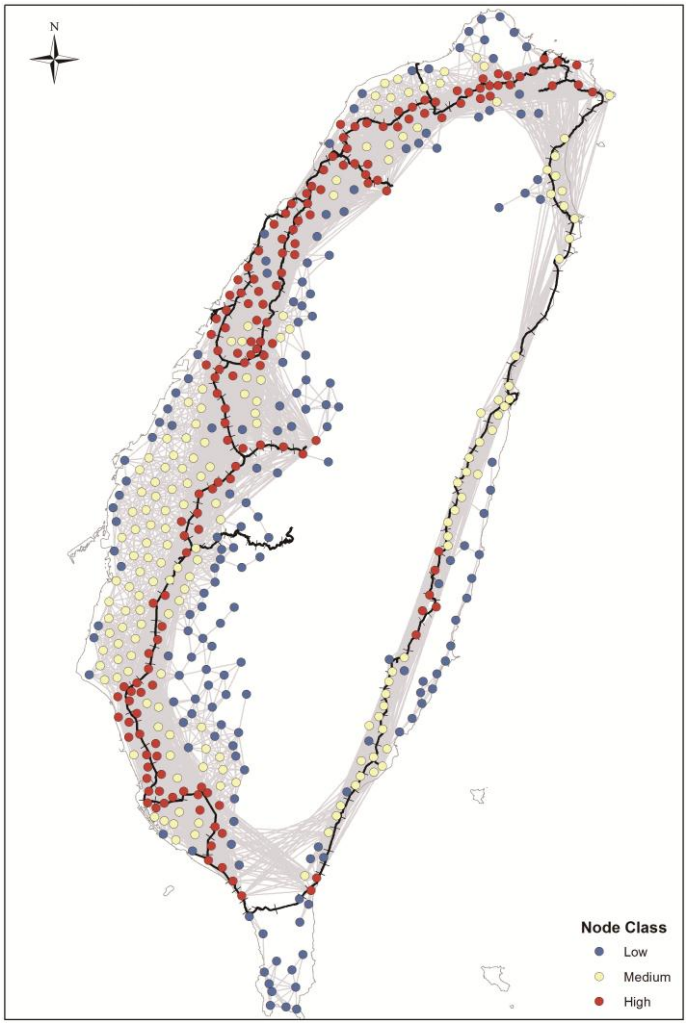


Figure 4.11 The calculation results of PR

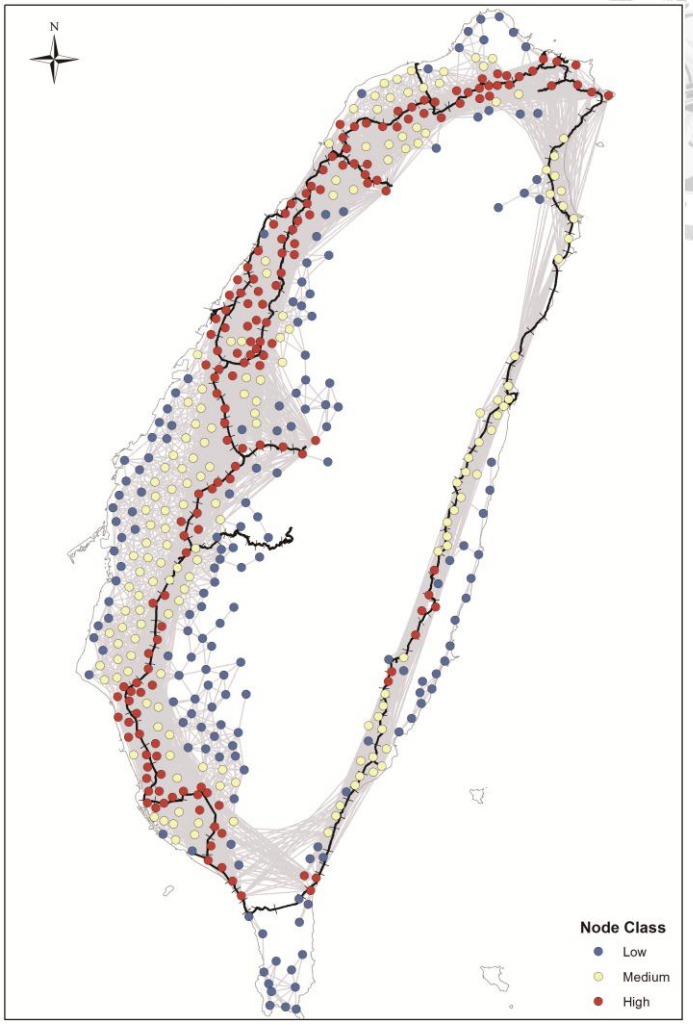


Figure 4.12 The calculation results of WPR

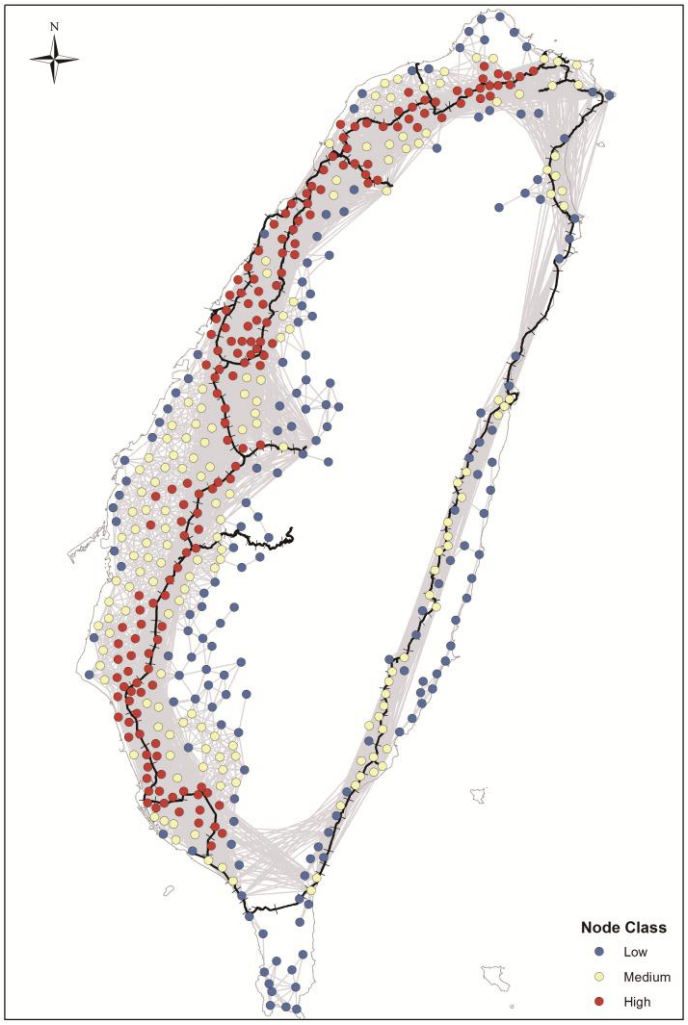


Figure 4.13 The calculation results of IDPR

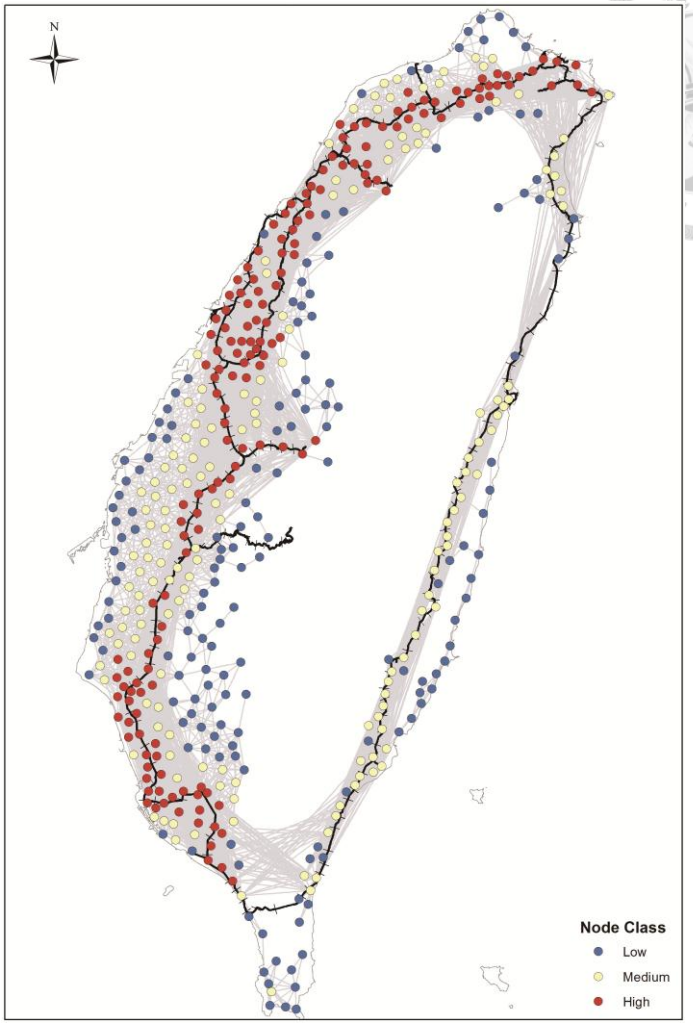


Figure 4.14 The calculation results of GPR

4.5. Correlation analysis

To know the capacity of each PR to predict the intensity of human activity, this study used the 5 indices of human activity relative importance to check their correlation with the network indices. The intensity data included population size (raw count), population density, totals flow of cars (the total number of cars entering and leaving each township per day), incoming flow (the number of cars entering each city), and flow betweenness. The network indices used in this correlation analysis include degree centrality, closeness centrality, betweenness centrality, PR, WPR, IDPR, and GPR.

Table 4.2 shows the Spearman Rank Correlation (rho) to compare the measurements and the human-activity intensity index. The result shows that four of the PRs have greater rho then the closeness and betweenness centrality metrics. The IDPR is the most strongly correlated with all intensity levels among the PR algorithms, and GPR is second.

Table 4.2 Spearman Rank Correlation (rho) between importance level and intensities indices

	Degree	Closeness	Betweeness	PR	WPR	IDPR	GPR
Population (size)	0.4941	0.4141	0.1893	0.4932	0.4978	0.6431	0.5475
Population (density)	0.5660	0.5015	0.2263	0.5655	0.5724	0.7112	0.625
Flow-Total (raw)	0.3493	0.3241	-0.0081	0.3476	0.3724	0.5033	0.446
Flow-In (raw)	0.3434	0.3199	-0.011	0.3417	0.3662	0.4962	0.4396
Flow (FB)	0.3353	0.3718	0.0572	0.3347	0.3316	0.5038	0.3954

Of the network-centrality metrics, closeness and betweenness centralities both had a lower rho with the human-activity index, possibly because they do not correspond to the spatial distribution of the human activity level. Thus, the distributions of population and human movement between cities do not follow closeness, which is defined based on the independence level (on resource/information flow), and betweenness, which is defined based on the potential to bridge to other nodes. In contrast, the result shows that the distribution of population and human movement pattern has a better correlation with the importance calculated from the degree centrality, which is defined based on the number of links (namely the neighbors in the network). This implies that the importance calculated from the number of neighbors is a better indicator of the intensity of human activity than that calculated from the independence and the potential to bridge to other nodes.

Overall, the PRs are strongly correlated with the distribution of the population and human movement. Thus, the probability of each city in the network being reached is related to human movement in the area.

As the network in this study is undirected, the PR-score distribution is proportional to the degree distribution. Therefore, the PR is very closely correlated to the degree centrality. The PR algorithm in an undirected network thus cannot capture the transitive effect in network. On the other hand, WPR captures the transitive effect, as described in the previous section, where the important cities in PageRank become more important in PR and the low-ranked cities become lower-ranked. Nevertheless, the correlation from WPR is close to that from PR. In other words, the transitive effect does not lead to a huge difference from PR in the correlation result, possibly because the Spearman Rank correlation is calculated based on the rank of each city, implying that

the transitive effect causes the difference in score distribution but does not change the rank distribution.

The correlation results showed that IDPR has the highest correlation with the intensity of human activity, with GPR second. Thus, the score distribution of IDPR is closer to reality than other network measurements, as observed in the previous section. Moreover, considering the distance as an impedance factor can capture the population distribution and human-movement patterns. In other words, the distance-decay effect is the most important effect for predicting the intensity of human activity. On the other hand, the correlation from the GPR is markedly different from that from the WPR, implying that the distance-decay effect did change the rank distribution in GPR, as in the IDPR and not in WPR. The result showed that GPR, which is built from the gravity model, is also a better indicator than the WPR and PR. The gravity model (or spatial-interaction model) is also useful for predicting the intensity of human activity.

4.6. Sensitivity analysis

The Spearman Rank correlation results showed that the IDPR and GPR were both better indicators for locating the major cities within the geospatial network. These results also suggest that the framework of this experiment can be used to locate the major cities from the raw ground-transportation data (including the road-network layers and railway-network layers), but the framework has key parameters with the potential to influence the correlation results, including one key parameter in the calculation of IDPR and GPR (the distance factor), and two key parameters in the network preparation process -- the number of cities extracted from the junctions distribution data (the number of mean centers or clusters in the k-means clustering) and the time threshold for constructing the links. These parameters might influence the correlation result, so the results might change if other parameters are chosen. Therefore, we performed a sensitivity analysis with these key parameters to check the Spearman Rank correlation in connection with the intensity of human activity.

The distance factor (beta) is the key to the distance-decay effect in the IDPR and GPR. Beta represents the power of the distance as an impedance. Previously, we used beta = 1. In this section, we used beta = 0.5, beta = 1, and beta = 2 for the sensitivity analysis. The ability to influence the moving probability of different distance factors (beta) is shown in Figure 4.15. Comparing a beta of 1 to a beta of less than 1, the influence of distance becomes less significant as the distance increases; when beta is greater than 1, the influence of distance becomes more significant as the distance increases. Therefore, this study aimed to test how the correlation results change as the betas change.

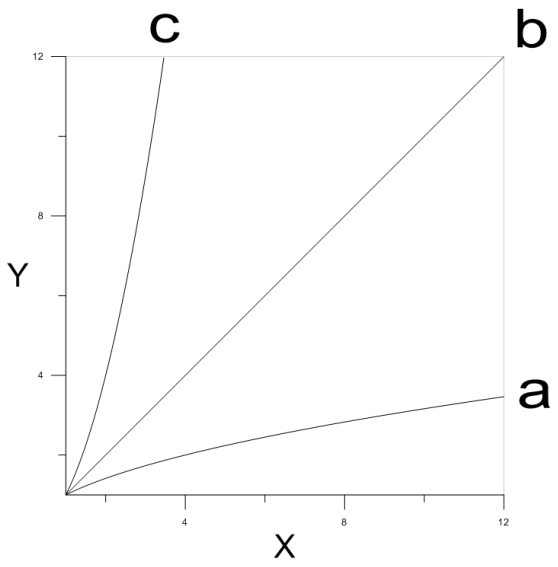


Figure 4.15 The ability to influence moving probability (Y) of the distance (X) with different beta.

$Y = X^{\beta}$, where (a) $\beta = 0.5$; (b) $\beta = 1$; (c) $\beta = 2$.

Different numbers of cities and time-threshold parameters would change the complexity of the geospatial network. Therefore, we used 100-to-750 cities with an increment of 50 (14 settings) and time thresholds from 30 min to 300 min (0.5 hour to 5 hour) with an increment of 30 min for the sensitivity analysis (10 settings); a total of 140 sets of parameters were used to calculate PRs at each beta (0.5, 1, and 2). Then, 140 x 3 calculations of each PR were prepared for the Spearman rank correlation with the intensity indices.

Five intensity indices were used in the previous section. As the population size and population density have similar distributions and correlations, and the total flow is similar to the incoming flow, this study takes only one of each pair above for the sensitivity analysis. Therefore, the intensity indices used in this section included population density, total flow, and flow betweenness.

Four PRs were used in this study. The results showed that the IDPR and GPR appeared to have the first- and second-highest correlation results (the sensitivity is always greater than those of the PR and WPR). Therefore, this study presents only the IDPR and GPR for the following discussion, focusing on the sensitivity calculated from the correlations from different parameter settings.

The sensitivity analysis results were shown in different intensity-index categories: Figure 4.16 shows the sensitivity plots of the correlation with population density; Figure 4.17 shows the sensitivity plots of the correlation with total flow, and Figure 4.18 shows the sensitivity plots of the correlation with flow betweenness. In each sensitivity figure, the results of the correlation with IDPR are shown on the left column (Plot a - Plot c), with GPR on the right (Plot d - Plot f), arranged according to the beta value. Each 3-D plot includes 140 points to that of the correlation (rho), with intensity indices for each set of parameters (X-axis for the number of cities, Y-axis for time threshold, and Z-axis for correlation), and an interpolated surface to show the sensitivity pattern.

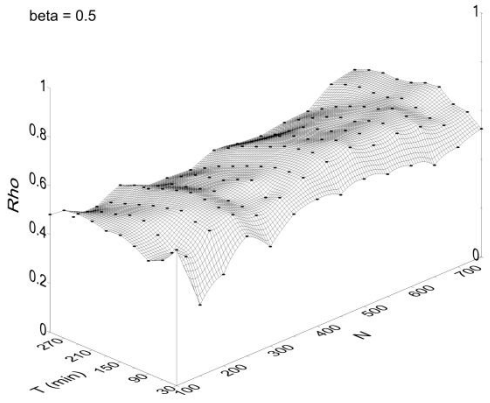
- Correlation with population density

The correlation results using the population density (Figure 4.16) were greater than other 2 intensity indices, similar to the result in Table 4.2. For the correlation using the population density with IDPR, while the beta is set to 0.5 (Figure 4.16a), the IDPR becomes stable when the number of nodes is greater than 350; when the time threshold is greater than 150 min and the number of nodes is more than 350, the correlation slowly decreases as the time threshold increases. When the beta is 1.0 (Figure 4.16b), the trend while the number of nodes increased is stable; but the decrease trend while the

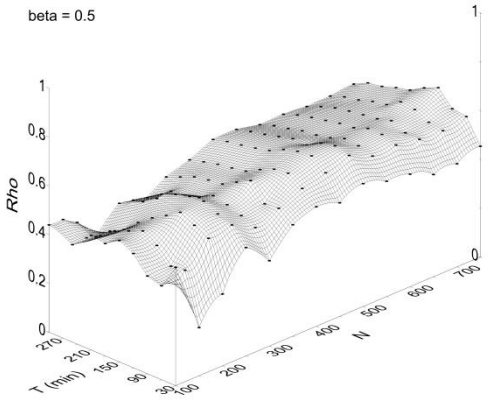
time threshold is greater than 150 min and the number of nodes is more than 350 became more significant than the previous beta setting. When the beta is 2.0 (Figure 4.16c), the trend is smooth and stable while the time threshold parameter setting is greater than 30 min; the number of node parameters has a wave-like trend, whereas a decreasing trend appears when the number of nodes is greater than 300, an increasing trend appears when the number of nodes is greater than 400, another decreasing trend appears when the number of nodes is greater than 550, and the trend increases again when the number of nodes is greater than 700; however, the increase or decrease extent (rho) is not more than 0.05, and the overall correlation is the highest compared to the results in Figure 4.16a and Figure 4.16b. For GPR, when beta is 0.5 (Figure 4.16d), there is a similar result in that the GPR is stable when the number of nodes is more than 350; when the time threshold is greater than 150 min, the correlation sensitivity surface becomes flat, which means the number of nodes and time threshold do not have sensitive effects on the correlation results. When beta is 1.0 (Figure 4.16e), the sensitivity surface pattern is similar to that in Figure 4.16d, but the overall correlation is greater in Figure 4.16e. When the beta is 2.0 (Figure 4.16f), the correlation sensitivity surface becomes flat at most settings, and it is the highest compared to the result from the other two beta settings.

IDPR vs. population density

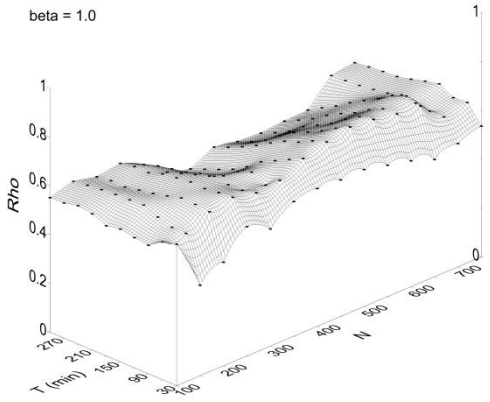
GPR vs. population density



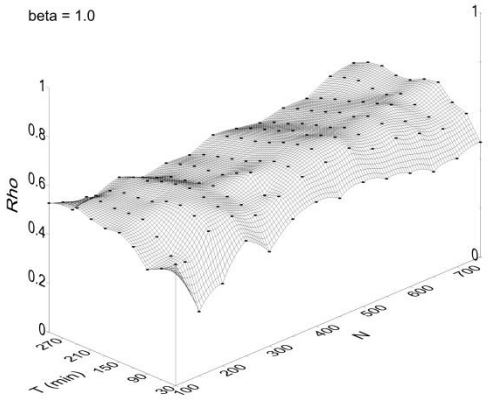
(a) beta = 0.5



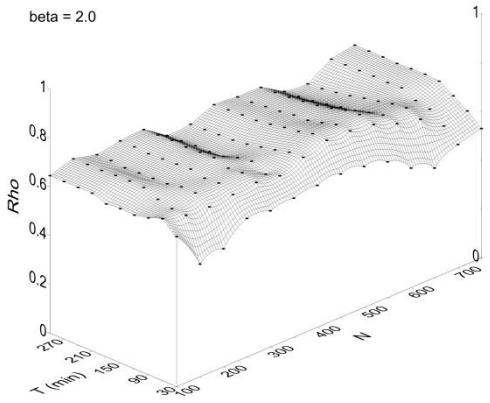
(d) beta = 0.5



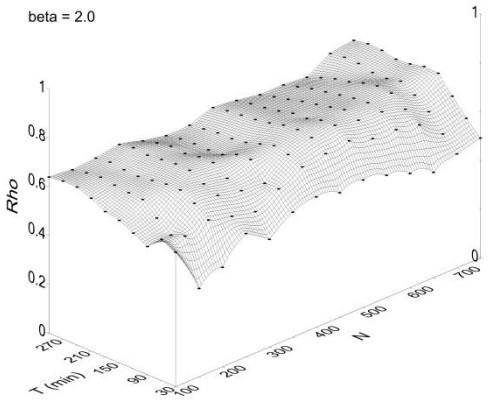
(b) beta = 1.0



(e) beta = 1.0



(c) beta = 2.0



(f) beta = 2.0

Figure 4.16 Sensitivity of no. clusters and time threshold for constructing links on correlation between IDPR (or GPR) with population density

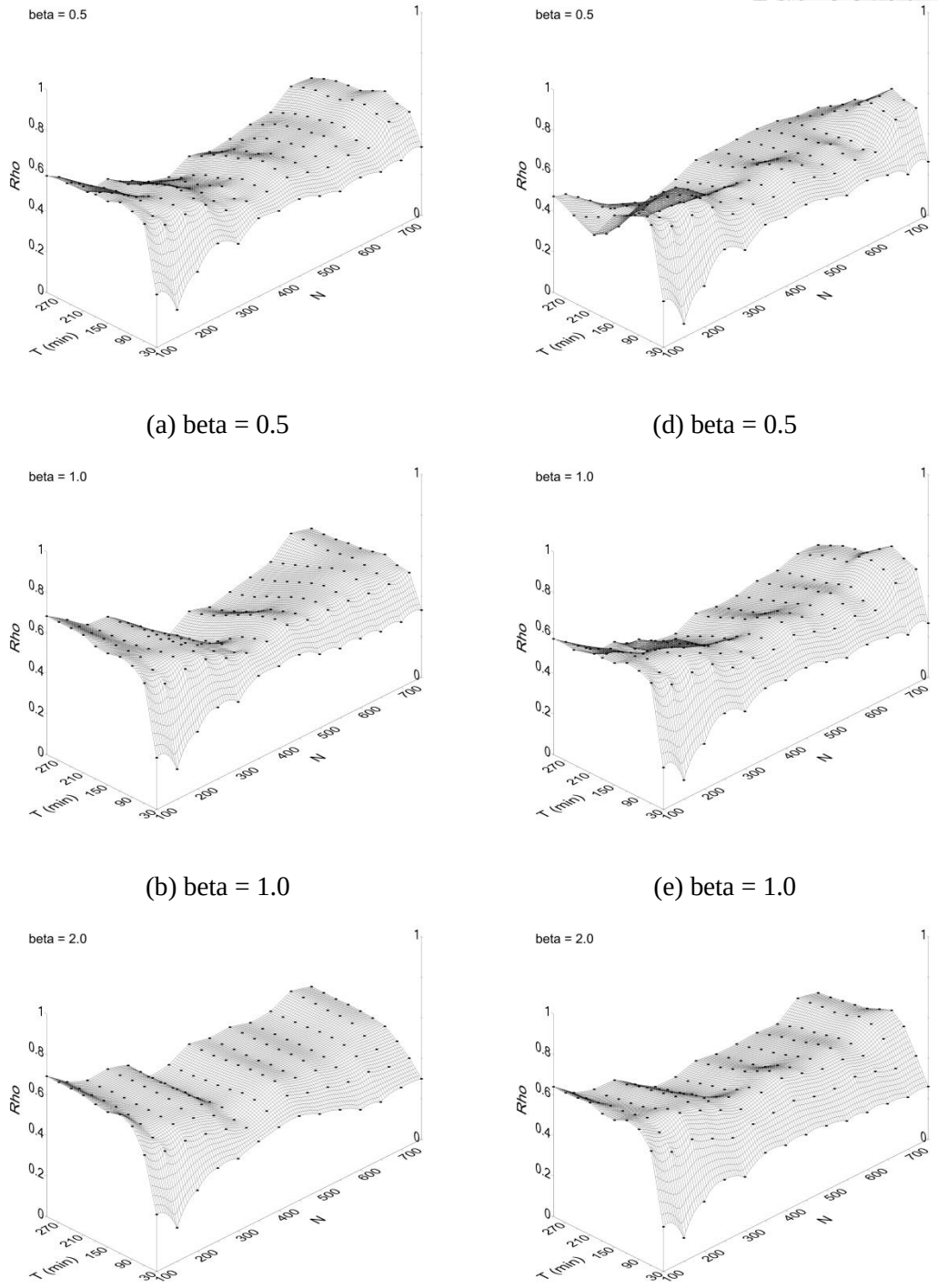
X: N, no. clusters in k-means clustering process; Y: T, time threshold (min) for constructing links with the underlying ground transportation system; Z: Rho, Spearman rank correlation of population density with IDPR (left) and GPR (right)

- Correlation with total flow

The correlation results using the total flow (Figure 4.17) had significant changes when the number of nodes is lower than 300 compared to other two relative importance indices. For IDPR, the time threshold did not have a significant effect on the surface in all beta settings' surface; when the beta is 0.5 (Figure 4.17a), the overall sensitivity surface appears to be flat, and small waves can be observed while the number of nodes are lower than 400. When beta is 1.0 (Figure 4.17b), the correlations were greater at the beginning (100 nodes), decreased significantly between 300 nodes and 400 nodes, and then the correlation increased again from 400 to 450, and became stable while the number of nodes is greater than 450. When the beta is 2.0 (Figure 4.17c), the wave-like pattern appeared, with a high correlation at the beginning, and a significant decrease from 300 to 350, and a little increase from 400 to 450, becoming stable when the number of nodes is greater than 450. For GPR, the changes in the number of nodes are similar to the IDPR in 3 sets of beta settings; the changes in the time threshold had a significant effect on the correlation results before the number of nodes reaches 450. When the beta is 0.5 (Figure 4.17d), a significant increase in correlation appeared from 30 min to 120 min, and then, the correlation decreased until 240 min, followed by a slower increase until 300 min. When the beta is 1.0 (Figure 4.17e), there was a significant increase in correlation from 30 min to 120 min, and then a decrease until 180 min, at which point, it became stable until 300 min. When the beta is 2.0 (Figure 4.17f), the sensitivity surface of correlation became similar to the IDPR that had a wave-like pattern; overall, the correlation between GPR and total flow is lower than the correlation between IDPR and total flow.

IDPR vs. total flow

GPR vs. total flow



(a) beta = 0.5

(d) beta = 0.5

(b) beta = 1.0

(e) beta = 1.0

(c) beta = 2.0

(f) beta = 2.0

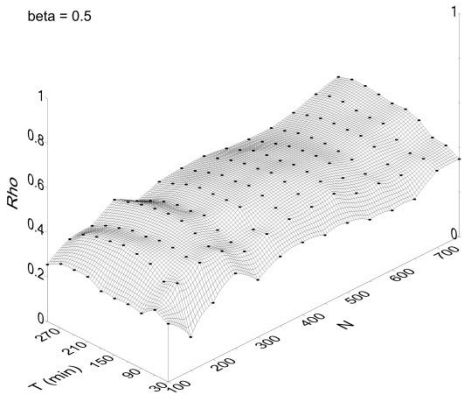
Figure 4.17 Sensitivity of no. clusters and time threshold for constructing links on correlation between IDPR (or GPR) with total flow

X: N, no. clusters in k-means clustering process; Y: T, time threshold (min) for constructing links with the underlying ground transportation system; Z: Rho, Spearman rank correlation of population density with IDPR (left) and GPR (right)

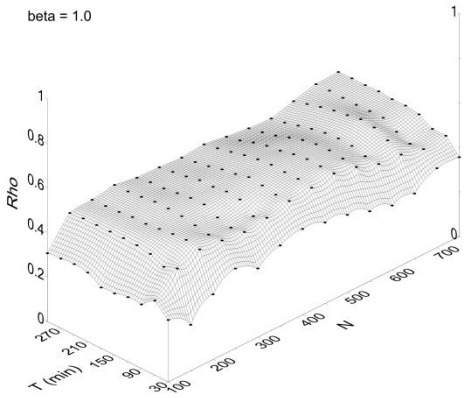
- Correlation with flow betweenness

The correlation results using the flow betweenness (Figure 4.18) appeared to be lower and relatively more stable than the previous results using the relative importance indices. For IDPR, when beta is 0.5 (Figure 4.18a), a flat surface appears after the number of nodes reaches 350, while the changes in time threshold and number of nodes are not significant. When the beta is 1.0 (Figure 4.18b), the flat surface appeared when the number of nodes reached 150, where the changes in time threshold and number of nodes also did not have a sensitive effect. When the beta is 2.0 (Figure 4.18c), the flat surface appeared between nodes in the range from 150 to 550, and a little wave-like pattern appeared, which had a small waving extent, there was a decrease from 550 to 600, and the surface became flat from 600 to 750. For GPR, when the beta is 0.5 (Figure 4.18d), the changes in time threshold had a significantly unstable pattern when the number of nodes was lower than 350, which suggested that the changes in time threshold had a very sensitive effect on the correlation. There was an increasing trend from 30 min to 120 min after 350; the flat surface appeared after the number of nodes reached 350, and the time threshold was greater than 120 min, which suggested that the number of nodes and time threshold at these ranges would not have a significant effect on the correlation with the flow betweenness. When the beta is 1.0 (Figure 4.18e), the sensitivity surface becomes relatively stable and the number of nodes reaches 350; when the number reaches 350 and the time threshold is greater than 120 min, the flat surface is similar to that in previous results (Figure 4.18d); an increasing trend appears from 30 min to 120 min, as in Figure 4.18d. When beta is 2.0 (Figure 4.18f), the sensitivity surface is nearly identical to the result in Figure 4.18c in the range from 100 to 550; the surface is relatively stable compared to Figure 4.18(c) after the number reaches 550.

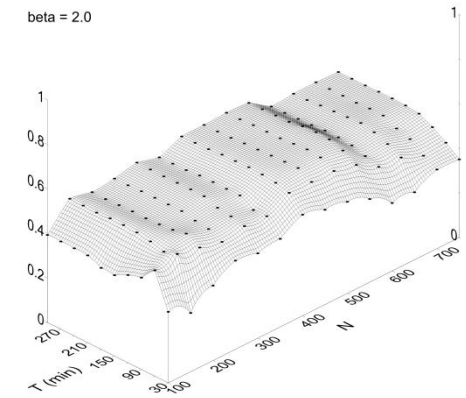
IDPR vs. flow betweenness



(a) beta = 0.5

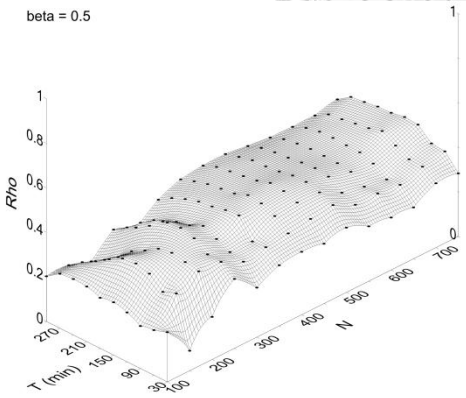


(b) beta = 1.0

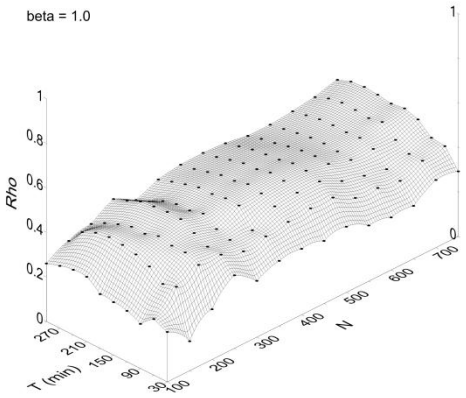


(c) beta = 2.0

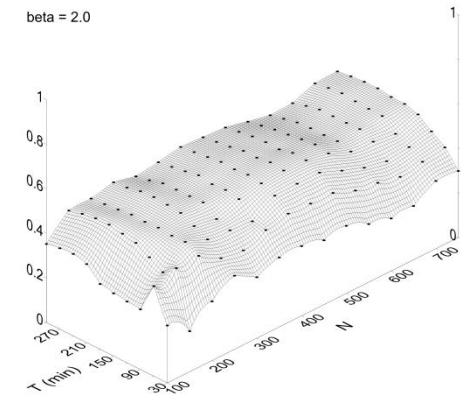
GPR vs. flow betweenness



(d) beta = 0.5



(e) beta = 1.0



(f) beta = 2.0

Figure 4.18 Sensitivity of no. clusters and time threshold for constructing links on correlation between IDPR (or GPR) with flow betweenness

X: N, no. clusters in k-means clustering process; Y: T, time threshold (min) for constructing links with the underlying ground transportation system; Z: Rho, Spearman rank correlation of population density with IDPR (left) and GPR (right)

Overall, the sensitivity surface of the correlation between IDPR and GPR with the intensity indices appears stable when the number of nodes is set between 400 and 600; within this range, the time threshold does not have a significant effect on the correlation results after 120 min and increases from 30 min to 120 min. Therefore, the correlation with the IDPR and GPR on the network, at 500 nodes and 60 min, is similar for different numbers of nodes and is greater when the time threshold increases to 120 min.

As the beta is the key to capturing the distance-decay effect, the calculation and correlation results with different beta settings have similar surface patterns that differ in the degree of increase or decrease (waving). Overall, the correlation is greater when the beta equals 2.0 and lower when the beta equals 0.5.

In conclusion, the settings of three key parameters have low sensitivity to the number of nodes and the time threshold, with similar trends. Additionally, the correlation result for 500 nodes and 1 hour predicts the intensity of human activity. When the time threshold is 120 min, the prediction is better than that in Section 4.5.

5. Experiment 2 – Taiwan High Speed Rail System

5.1. Preface

Taiwan's High Speed Rail (HSR) system was built and started services in the west plain of Taiwan since 2007. THSR is an inter-regional transport system connecting northern Taiwan (Taipei city, 台北市) and southern Taiwan (Kaoshiung city, 高雄市) in about 90 minutes, with 345 km of railway.

In network-analysis terms, HSR made longer-distance connection possible, directly changing the transportation-network connectivity, especially for the places that have a HSR station (Figure 5.1). By the transitive effect, "importance" should be transferred from the source to the target. In other words, those nodes that are not directly influenced by the HSR station but are near such nodes should rise in importance. Given the previous results (that the transitive effect, attractive effect, and distance-decay effect exist in the inter-city transportation network), would the influencing area also experience the transitive effect? Furthermore, which one of the four PRs is the best choice for capturing the transitive effect and representing the changes in importance?

As distance is defined as an impedance in this study, a long-distance connection might have relatively lower weight in the calculation of IDPR and GPR compared to PR and WPR. In other words, the calculation of IDPR and GPR would assign less weight to the long-distance links that were connected by the HSR, so the changes in the connection structure might not strongly influence the overall importance-score distribution, generating a question: do the long-distance links have strong effects on the importance shifts?

On the other hand, the HSR stations were built far from each other, and the cities that surround the HSR cities (the cities near HSR stations) would normally connect the HSR cities by the local roads. These short links might gain importance from the changes in the connection structure. What, then, is the effect of the short links that connect the HSR cities?

The transportation network in this study is undirected, which means that the changes in the original PR results are similar to the changes in degree. In other words, the transitive effect would not be represented in the original PR results. While the WPR, IDPR, and GPR consider the network and geography, these algorithms were expected to capture the transitive effect in the shifting network.

Therefore, this study aimed to check the shift in importance using the four PR algorithms and to evaluate the ability of the four PR algorithms to capture the network transitive effect after HSR.

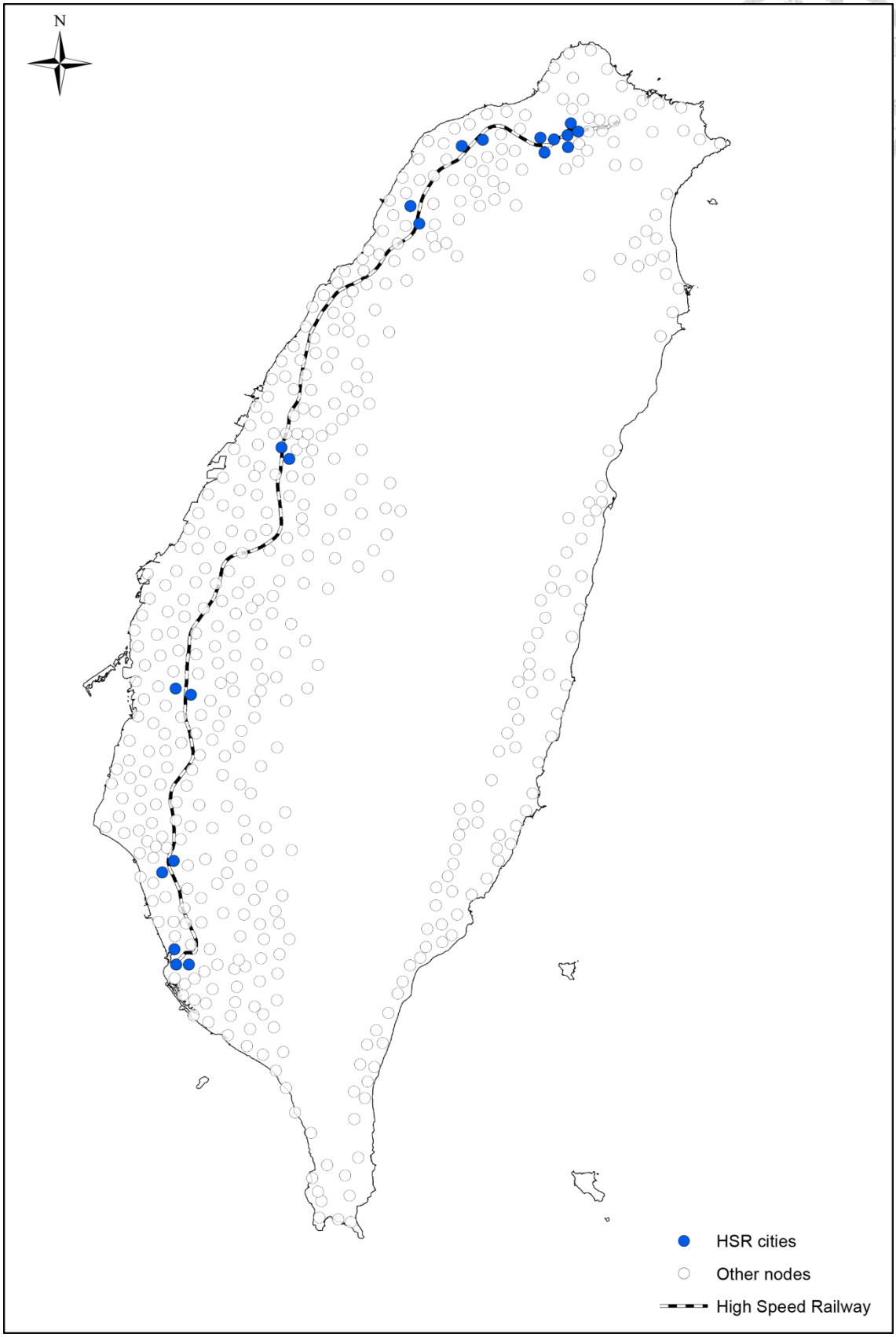


Figure 5.1 The Taiwan's High Speed Rail System and the HSR cities

5.2. Data preparations and calculations

This experiment used 2 networks: the “before” network (before HSR was built) and the “after” network (after HSR was built). This study used the one-hour accessible network from the previous experiment as the “before” network, including the road and railway in the network, then integrated the road, railway, and HSR, redoing the links as in the previous test to build the “after” network.

This study separately calculated the four PR algorithms and the degree centrality for the networks. Comparing the results, this study identified the cities with increased scores then compared the results from each algorithm for these cities to check whether the scores for the other algorithms also increased (Figure 5.3, Figure 5.3).

Finally, this study creates standard deviational ellipses to show the area of significant influence (Figure 5.4).

5.3. Results

After the calculations, this study compares the before network to the after network, and classifies the nodes as increasing or decreasing. As the sum of all nodes would always equal 1, when the score of one node increases, another node must decrease. This test focuses on the increased scores; the nodes with decreased scores are not discussed in this section.

Figure 5.2 shows the number of nodes that increased in different metrics. The 20 HSR cities (blue nodes in Figure 5.3) increased in PR, IDPR and degree centrality. In addition to the HSR cities, 30 other cities (yellow nodes in Figure 5.3) increased in WPR and in GPR. A further 9 cities (red nodes in Figure 5.3) increased in GPR.

The results showed that HSR cities that are directly influenced by the stations (the blue nodes in Figure 5.1) directly increased in degree centrality and in all four PRs (Figure 5.2, Figure 5.3). The trends in PR and IDPR are identical; besides the HSR cities, no other cities showed an increased PR or IDPR. The PR is expected to be similar to the degree centrality because the network is undirected. On the other hand, the IDPR did not differ significantly from the degree centrality, possibly because distance decay is not strong enough to get more scores from nearby HSR cities.

On the other hand, cities other than the HSR cities have increased WPR scores. These nodes have no direct changes in network topological structure, but they are all connected to nearby HSR cities. In other words, because of the connection to a city experiencing increased activity (in network-topological terms), their importance also increased. Thus, because the cities are close and connect to the HSR cities, they become

more accessible. The WPR better captures the transitive effect after considering the attractive effect in an undirected network.

In addition, the result shows that more cities increase in GPR than in WPR. Besides the changes in network topological structure that make nodes more accessible, the distance-decay effect also makes the other cities close to these cities (including the HSR cities and the surrounding cities that became more accessible) increase in importance. In other words, the distance-decay effect can also cause another type of transitivity.

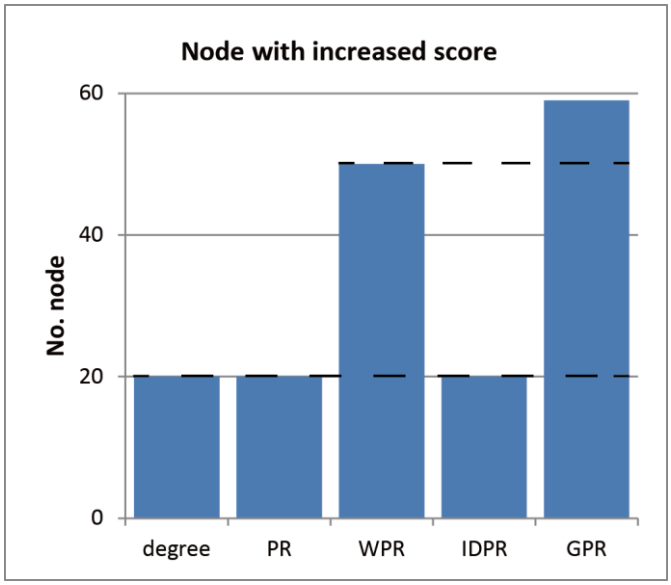


Figure 5.2 The number of nodes in 5 network metrics that had increased scores.

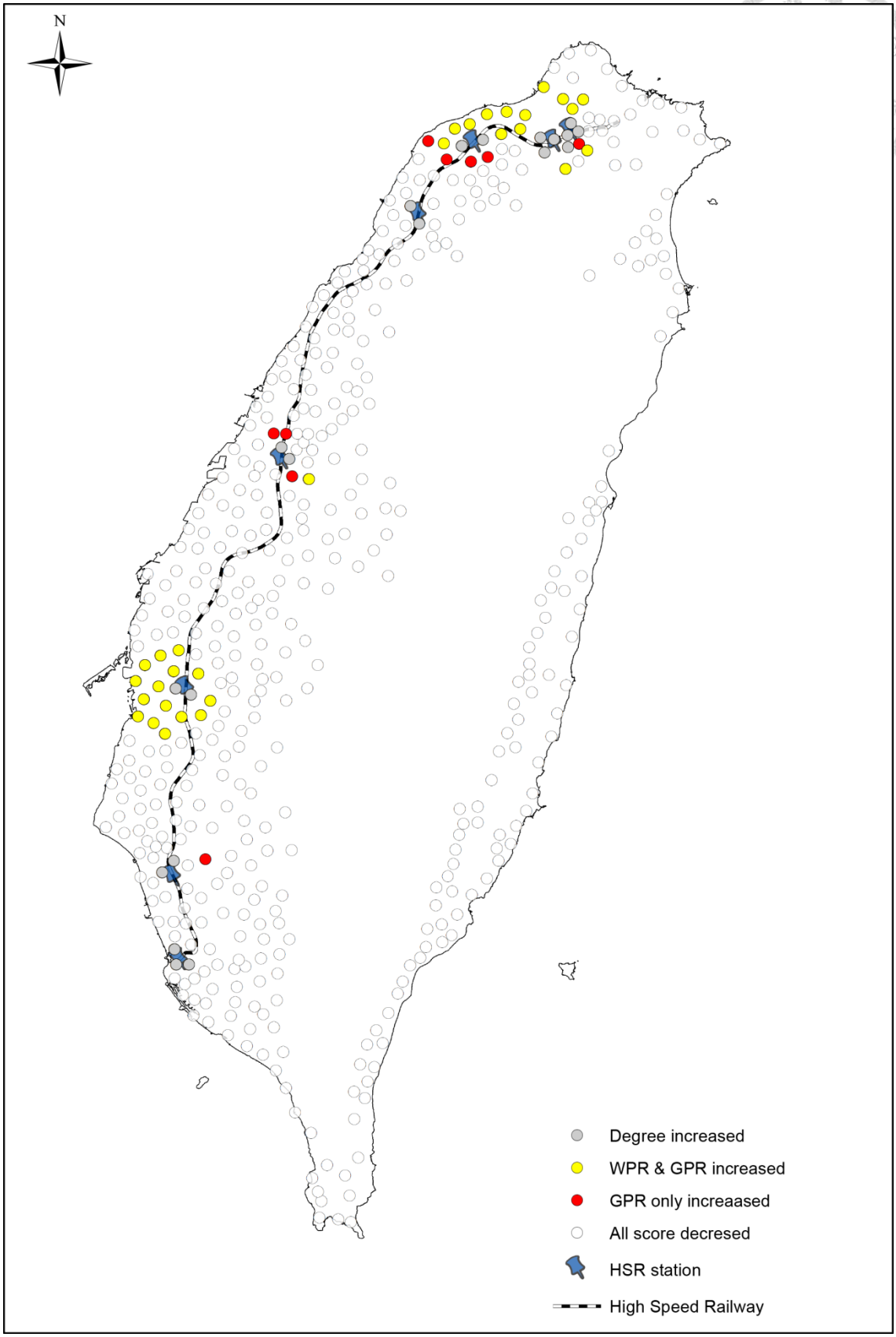


Figure 5.3 The changing status of nodes.

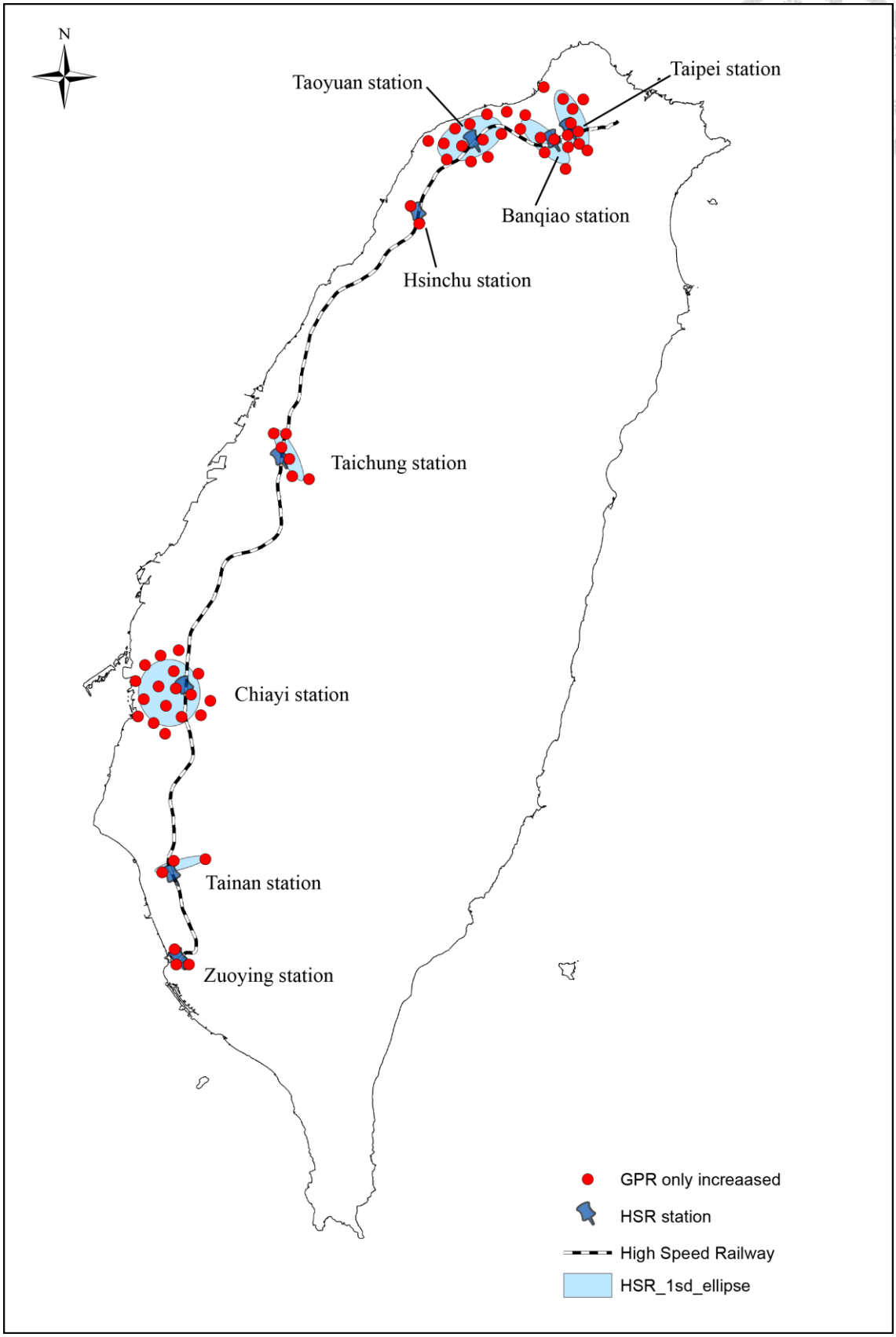


Figure 5.4 The influence ellipse of 1 standard deviation

Figure 5.4 shows the nodes that increased in GPR and their influencing area by using the ellipse of 1 standard deviation. Because there are more HSR stations in northern Taiwan, the 27 red nodes formed a cluster, and 3 ellipses showed the influencing area of the 3 HSR stations in northern Taiwan. Beside the north cluster, the cities near the HSR Chiayi station (嘉義站) formed an area of broad influence. Based on the above results and observations, this finding might have arisen because the HSR Chiayi station is far from the Taichung station (台中站) on the north side and from the Tainan station (台南站) on the south side; the HSR cities of Chiayi station can thus assign greater proportions of scores to nearby cities (connected by the local transportation system), which is a result of the distance-decay effect. In other words, this finding is a result of the interaction between the change in the network topological structure and the distance-decay effect in the GPR calculation.

Although the Taichung station has similar characteristics to the Chiayi station, it did not form an area of wide influence, possibly because before HSR, the transportation connectivity of the Taichung area was already fully developed. Therefore, the operation of the HSR system did not have a large impact on the local accessibility and network connectivity.

In summary, first, the GPR might be the best choice for capturing the transitive effect of the network changes; second, even though long-distance links have less weight, the changes in the topological structure would also influence the node's score; third, the short links that surround the HSR cities would influence the surrounding cities' scores. Finally, this test gave an example of the effect of the integrated network topological structure and its geographical properties.

6. Discussions

6.1. The key features of the PageRanks algorithms

As the PR algorithm came from a random-surfer simulation, it differs from the traditional centrality metrics (including degree centrality, closeness, and betweenness centrality), which are defined based on the node alone. In addition, the random-surfer simulation also provides the PR algorithm with more opportunities for modification. By changing the behavior of the random surfers, this study extends the PR algorithm to the IDPR and GPR to analyze geospatial networks.

The distance factor (beta function) is a power of the distance, and it is the key to forming the distance-decay curve, so it would affect the behavior of the random surfers selecting the next target. In other words, a high distance factor would keep a greater proportion of the scores in a local cluster (a local set of nodes that formed a circle route). The distance factor should be assigned according to other theories or concepts in transportation geography, to capture different issues with different distance-decay curves. Therefore, the distance-decay effect can be captured by adding the inverse distance with a distance factor as the power in the PR algorithm.

This study adopted the idea from the gravity model, integrating attractiveness and distance decay to modify the proportion equation of PR. For attractiveness, this study did not multiply the mass of cities (on the geographical scope, e.g., the population, economic level, etc.) of two ends of the link, which was done in spatial interaction model (Tobler, 1970), instead using the in-degree centrality of the target node and calculating its proportion compared to the other nodes with the same source node. The in-degree centrality is an index from the network topological structure. In other words,

this study uses the attractiveness derived from the network topological scope. Then, this study integrates the attractiveness with the distance-decay function, which is a concept from the geographical scope. Therefore, the GPR is a special version of the PR that integrates both the topological and the geographical characteristics.

6.2. Experiment 1

The inter-city network in this study is a connection network, with links making movement between the places possible in under an hour. Therefore, the importance in this network represented accessibility, which was defined as the potential for opportunities for spatial interaction (Reggiani *et al.*, 2011). As mentioned in the random-surfer simulation, the results of the IDPR and GPR also represent the probability of each place being reached by the people (random surfers). Therefore, in this case study, the calculation results can also be explained as the accessibility of the cities.

The results in Section 4.4 show that the PR and WPR have high-category nodes on the east coast and in the mountains (suburbs), a result of the long links formed by the railway. In network topology, these long links are global bridges that connect different subgroups and thus make these subgroups more accessible, as represented by higher scores in PR and WPR, but these characteristics are not known in the real world because the accessibility of those areas is relatively lower. The ability of IDPR to keep score in the local clusters make the score in the western plain, and the same reason also make the accessibility of those suburbs area lower, which is closer to the real situation.

As in the result in Section 4.5, the Spearman rank correlation showed that all the PRs had greater correlation than the traditional centrality metrics. Thus, the probabilities, which were calculated based on the random-surfers concept to capture the transitive effect, were better indicators than the centrality metrics, which were based on the characteristics of each node. Within the PR algorithms, IDPR has the highest correlation and the GPR has the second-highest correlation. Thus, the distance between the connected places matters in the spatial interaction, which follows a distance-decay curve.

A limitation in this experiment is that the data used to analyze the Spearman rank's correlation are township-level data. As our nodes in the inter-city network represent the cities (or settlements) as the mean centers of where people live, they did not follow the township boundaries. Therefore, this study must conduct a spatial joining and summarizing process for matching the cities and the township level before the Spearman's rank-correlation analysis. This process might influence the result, but no better data was found for this analysis. On the other hand, these characteristics eliminate the problem of places with heavy traffic but no inhabitants; data on the intensity of human activity (demography data and car flow data) cannot capture such characteristics because the results were spatially joined into townships.

6.3. Experiment 2

The results of experiment 2 suggest that the changes in the links only influence the cities surrounding the THSR cities in the distribution of the WPR and GPR. The changes in IDPR are not large enough to increase the scores of cities surrounding the THSR cities, which might be why distance is an impedance in the calculation of the

IDPR. Therefore, the proportion of the new long links is relatively lower and gives rise to this characteristic. The changes in PR are the same as the degree expected.

These changes suggest that adding long links in the network would significantly influence the results, except for the IDPR, as discussed above. GPR has more nodes with increased scores than does WPR, suggesting that the distance-decay effect not only impedes the new long links, as in IDPR, but also forms a gravitational force between the surrounding cities and the THSR cities, increasing the scores of the surrounding cities.

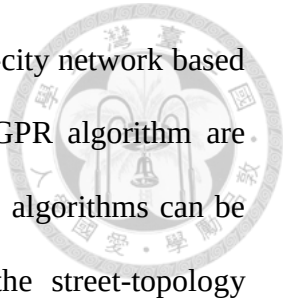
The results also suggest that the farther apart the new hubs are (in this case, the THSR cities), the larger the influencing area. Thus, if new THSR stations are built on the east coast, they might strongly impact the local area, but the spatial distribution of the local nodes and their former scores and connectivities must also be considered. If the local nodes do not form a local cluster or the number of nodes is limited by the landscape, it is hard to form a large influencing area; if the local nodes' former scores were high or their connectivities were dense, the impact would differ from the prediction.

7. Conclusions and suggestions for future studies

PR can be viewed as a random-surfer simulation in which the random surfers move among webpages by selecting hyperlinks at random. The PR algorithm assumes that people would move between webpages randomly and calculates the probability of each webpage being reached. This idea is extended in this study into geography; the random surfers visit places randomly within the geospatial network by choosing the next visit target according to the distance-decay function and gravity model. Thus, this study proposed the IDPR, under which people choose the next visit target according to the distance-decay curve. This study proposed the GPR, which implies that people choose the next visit target by the gravity model. Therefore, the algorithms this study proposes integrate the network topology and geography. Using these algorithms, the accessibility of each node in the geospatial networks can be calculated with the implication of the random surfers, who act according to the distance-decay function or gravity model.

Our framework in Chapter 4 is designed to locate the major cities based only on ground transportation data. The analysis results showed that the IDPR is the best indicator of the intensity of human activity. In other words, together, the study framework and the IDPR can be used to identify the major cities from only the ground transportation data.

This study is the first to use the proposed study framework, and the proposed algorithms (IDPR and GPR) are improvements, but to prove that this framework and these novel algorithms can be used to identify major cities, other countries and scales should be tested with the framework.



The network in this study is a case study on forming an inter-city network based on the ground transportation data. The IDPR algorithm and the GPR algorithm are designed to analyze the geospatial networks, which mean that these algorithms can be used on other geospatial networks, e.g., the airline networks, the street-topology network, and the shipping network. Indeed, these algorithms should be tested on more types of geospatial network to elucidate the characteristics and potential of these algorithms. The inter-city network in this study is undirected, and the PR algorithms would consider the links as two-way links, so the calculation results of the PR algorithm become proportional to the degree centrality and the transitive effect cannot be emphasized. On the other hand, the calculation of other PR algorithms might be limited by an undirected network. Therefore, this study suggests that these PR algorithms should be tested in other types of network with directed links.

The novel algorithms (including IDPR and GPR) considered the distance-decay curve of the inverse distance with a distance factor (beta function). No other type of distance-decay curve was considered. As other types of spatial interaction might involve other types of distance-decay function, the distance-decay function should not be limited to the inverse distance. Therefore, the distance-decay function can be modified according to the spatial-interaction settings.

Reference

Alderson, A. S., and J. Beckfield. 2004. Power and position in the world city system. *American Journal of Sociology* 109(4): 811-851.

Batten, D. F. 1995. Network cities: creative urban agglomerations for the 21st century. *Urban Studies* 32(2): 313-327.

Bonacich, P. 1987. Power and centrality: A family of measures. *The American Journal of Sociology* 92(5): 1170-1182.

Brin, S., and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30: 107-117.

Commoner, M. 1971. The Closing Circle: Nature, Man, and Technology. New York.

Derudder, B., L. Devriendt, and F. Witlox. 2007. Flying where you don't want to go: An empirical analysis of hubs in the global airline network. *Tijdschrift voor Economische en Sociale Geografie* 98(3): 307-324.

Ducruet, C., D. Ietri, and C. Rozenblat. 2011. Cities in worldwide air and sea flows: A multiple networks analysis. *Cybergeo : European Journal of Geography: Systems, Modeling, Geostatistics* (article 528).

Ducruet, C., S. W. Lee, A. K. Y. Ng. 2010. Centrality and vulnerability in liner shipping networks: revisiting the Northeast Asian port hierarchy. *Maritime Policy & Management* 37(1): 17-36.

El-Geneidy, A., and D. Levinson. 2011. Place Rank: Valueing spatial interactions. *Networks and Spatial Economics* 11(4): 643-659.

Freeman, L. C. 1978. Centrality in social networks conceptual clarification. *Social Network* 1(3): 215-239.

Fotheringham, S. 1981. Spatial structure and distance-decay parameters. Annals of the Association of American Geographers 71(3): 425-436.

Gargiulo, F., M. Lenormand, S. Huet, and O. Baqueiro Espinosa. 2012. Commuting network models: Getting the essentials. Journal of Artificial Societies and Social Simulation 15(2): 13

Guimera, R., S. Mossa, A. Turtschi, and I. A. N. Amaral. 2005. The worldwide air transportation network: Anoalous centrality, community structure, and cities' global roles. *PNAS* 103(22): 7794-7799.

Institute of Transportation, Ministry of Transportation and Communications. 2009. The Demand model of intecity transportation systems under national sustainable development in Taiwan (4/4). ISBN: 978-986-01-8085-5.

Jiang, B. 2009. Ranking spaces for predicting human movement in an urban environment. *International Journal of Geographical Information Science* 23(7): 823-837.

Jiang, B., J. Yin, S. Zhao. 2009. Characterizing human mobility patterns in a large street network. *Physical Review E* 80(2).

Jiang, B., T. Jia. 2011. Agent-based simulation of human movement shaped by the underlying street structure. *International Journal of Geographical Information Science* 25(1): 51-64.

Jøsang, A., E. Gray, M. Kinateder. 2006. Simplification and analysis of transitive trust networks. *Web Intelligence and Agent Systems* 4(2): 139-161.

Jøsang, A., R. Ismail, C. Boyd. 2007. A survey of trust and reputation systems for online service provision. *Decision Support Systems* 43(2): 618-644.

Newman, M. 2010. Networks: An Introduction. Oxford University Press.

Noh, J. D., and H. Rieger. 2004. Random walks on complex networks. Physical Review Letters 92(11): 118701-1-118701-4.

Reggiani, A., P. Bucci, G. Russo. 2011. Accessibility and network structures in the German commuting. *Networks and Spatial Economics* 11(4): 621-641.

Reggiani, A.,S. Signoretti, P. Nijkamp, A. Cento. 2009. Network measures in civil air transport: A case study of Lufthansa. Tinbergen Institute Discussion Paper, Tinbergen Institute, Amsterdam, The Netherlands.

Reily, W. J. 1931. The law of retail gravitation. New York.

Scholz, A. B. 2011. Spatial network configurations of cargo airlines. *Working Paper Series in Economics* 20.

Stephenson, K., and M. Zelen. 1989. Rethinking centrality: Methods and examples. *Social Networ*ks 11: 1-37.

Stewart, J. Q. 1950. The development of social physics. *American Journal of Physics* 18: 239-253.

Symeonidis, P., E. Tiakas, and Y. Manolopoulos. 2010. Transitive node similarity for link prediction in social networks with positive and negative links. Paper

presented at the 4th ACM conference on Recommender Systems, NY, United States of America.

Tinbergen, J. 1963. Shaping the world economy. *The International Executive* 5(1):27-30.

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2): 234-210.

Tobler, W. 2004. On the first law of geography: A reply. *Annals of the Association of American Geographers* 94(2): 304-310.

Wang F., A. Antipova, S. Porta. 2011. Street centrality and land use intensity in Baton Rouge, Louisiana. *Journal of Transport Geography* 19(2): 285-293.

Xing, W., and A. Ghorbani. 2004. Weighted PageRank Algorithm. Paper presented at the 2nd Annual Conference on Communication Networks and Services Research, NB, Canada.