Traffic & Transportation

Data Source Importance Evaluation for Highway Networks: A Complex Network-Based Approach

Huangqin HUANG[1], Jianhua GUO[2], Xiangyu SHI[3], Leixiao SHEN[4]

Original Scientific Paper
Submitted: 12 Nov. 2023
Accepted: 18 Mar. 2024

[1] hhq5527@163.com, Southeast University, Intelligent Transportation System Research Centre

[2] Corresponding author, seugjh@163.com, Southeast University, Intelligent Transportation System Research Centre; Ministry of Transport, Key Laboratory of Transport Industry of Comprehensive Transportation Theory (Nanjing Modern Multimodal Transportation Laboratory)

[3] 13341416139@163.com, Southeast University, Intelligent Transportation System Research Centre; BYD Company Limited

[4] 106163580@qq.com, Xuzhou Highway Management Agency

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT

Data collection technologies or data sources are critical for highway network management. However, due to the limitations on available management resources, determining the importance of these data sources is necessary to allocate these resources reasonably. This study proposes a complex network based method for evaluating the importance of multiple data sources in highway networks. This method includes mainly three steps. First, the business-data source relation will be identified and formulated for the highway network. Second, a business data source complex network is built from the previously identified business-data relationship. Third, an entropy weight method is used to compute and rank the importance of data source nodes by combining three indexes of degree centrality (DC), closeness centrality (CC) and structural holes (SC) computed based on the complex network. The proposed method is applied and illustrated using the highway network of Xuzhou City, Jiangsu Province, China. The results show that among the data sources, the most important data source is the continuous traffic survey station, followed by an automatic gantry-based station and vehicle detectors-based system. Discussions on the limitations, applications and future studies are provided for the proposed approach.

KEYWORDS

highway network operations; data source; complex network; centrality; entropy weight method.

1. INTRODUCTION

The highway network is a critical infrastructure system supporting the development of national society and economy. To this end, effective management of highway networks is one of the essential tasks of highway management agencies around the world. For this purpose, highway management agencies have developed a variety of business models to fulfil the needs of highway network operations. The business models refer to the management tasks that highway management agencies have to perform to operate the highway network, such as performance measurement, incident detection or traffic volume prediction [1]. All these business models play a vital role in the improvement of highway network operations.

The highway management business models rely heavily on the data collected in the highway network [2, 3]. To this end, many data collection technologies, or data sources, have been applied to highway networks around the world to collect and supply a large amount of data to support highway network management tasks [4, 5]. There has been a growing consensus that the application of such data will be inevitable for enhancing the effectiveness of highway network operations.

However, constructing and maintaining these data collection facilities, or data sources, requires a significant amount of resources [6, 7]. To this end, there is a need to allocate limited resources across these data

collection facilities appropriately in terms of supporting highway network management tasks. Consequently, it is necessary to identify and hence rank the importance of these data sources [8]. In doing so, highway management agencies can allocate limited resources based on this importance ranking order, and well-maintained data collection facilities will be able to generate normal highway network data that are beneficial for supporting highway network management tasks.

Therefore, the purpose of this paper is to provide a complex network based data source importance evaluation approach. In this approach, the critical part is to develop a complex network connecting the data sources and the business processes or tasks of the highway network. Then based on such a complex network, the importance of these data sources can be evaluated and ranked. For this purpose, three steps are included in the proposed approach. First, the business-data source relation will be identified and formulated. Second, based on the identified business-data source relationship, the business-data source complex network will be constructed by integrating these business-data source relations. Third, based on this complex network, the importance of data sources will be evaluated by integrating three conventional importance assessment indexes [9 – 11], namely, degree centrality (DC), closeness centrality (CC), and structural holes (SC) [12, 13], using an entropy weight method [14, 15].

The rest of the paper is organised as follows. After a literature review section, the proposed method is presented in detail. Then, the proposed method is illustrated in a case study using the highway network of Xuzhou City, Jiangsu Province, China. Finally, discussions and conclusions are provided.

2. LITERATURE REVIEW

In this section, studies concerning data source importance evaluation and the relationship between businesses and data sources are reviewed and summarised.

Data source importance evaluation is not new and has been investigated in many fields. In social sciences, Hanson-DeFusco evaluated the importance of data sources relating to triangulation in social sciences in decision-making and planning evaluations [16]. Also, Park et al. proposed a method to solve the problem of estimating node importance in knowledge graphs [17]. In the medical area, Narayan reviewed and evaluated the national mortality data from the Civil Registration System (CRS), Medical Certificate of Death (MCCD), and Sample Registration System (SRS) in India, showing that the SRS data source scored the highest [18]. In addition, Price and Burley assessed the choice of primary and secondary information sources and identified the main information sources for the potential use of current awareness in occupational diseases [19]. When evaluating secondary data sources in epidemiology, Sorensen et al. believe that if the data source is very relevant to a specific research question, then this data source is very important and cost-effective [20]. Hjørland explained why it is necessary to compare information sources with research frontiers. If findings on the research front change, the evaluation of information sources may also change [21]. In information science, Hjørland discussed 12 different methods for evaluating information sources [22]. In addition, in terms of the Worldwide Governance Indicators (WGI), Kaufmann et al. argued that the data source of business information providers is more important than other types of data sources because it provides more sample data [23]. In transportation, Wood and Regehr tested the effectiveness of different axle load data sources in road design through a series of hierarchical analyses [24]. In addition, Broach et al. affirmed the importance of the old "small" data sources in estimating the annual average daily bicycle traffic [25]. Jiang et al. tested various combinations of electronic smart card data and global positioning system data in terms of bus travel time prediction [26]. Despite these studies, research is still limited on the importance evaluation of data sources in transportation. In particular, the operation of highway networks has many data collection facilities, which makes it necessary to evaluate the importance of these data sources.

The relationship between businesses and data sources is critical for evaluating the importance of data sources, and many studies have been conducted over the decades. Yang et al. analysed the connection between the data-driven business of circular economy and data sources and used scatter plots to represent this connection [27]. Khorashadizadeh et al. reviewed the knowledge map of COVID-19 management, constructing

the relationship between various COVID-19 data and the concerns of the people [28]. Wan et al. used knowledge graphs and big data analysis to integrate various maritime business and navigation service data [29]. Kam et al. identified key supply chain locations and industry relationships through existing data sources of goods [30]. Nguyen and Cao proposed a new sorting function to effectively measure the correlation between the XML data sources and a given query [31]. Tok et al. developed a California freight data repository, connecting freight businesses with data sources and providing lookup tables for each data source for the convenience of the users [32]. Tijssen et al. discussed the INDSCAL statistical model, revealing the structure of the relationship between multiple data sources and scientific entities [33]. Through the above review, it can be found that there are studies exploring the relationship between businesses and data sources in many fields. However, as also can be seen, the studies are still limited when it comes to transportation.

In summary, data source importance evaluation has been investigated widely in many fields, while in transportation, in particular in highway network operations which involves a rich amount of data sources and business models, such studies are still limited. Therefore, this paper will construct a complex network connecting business models and multiple data sources, based on which a complex network based method is proposed to evaluate the importance of data sources.

3. PROPOSED METHOD

In this section, the proposed method is presented, including an overview and detailed descriptions of each step of the method.

3.1 Overview

The method framework is shown in *Figure 1*. According to *Figure 1*, the relationship will be first identified through analysing the businesses and data sources in the highway network, to construct the business-data source relation. After that, the business-data source complex network will be constructed after defining the nodes and edges of the network. Finally, three indexes are computed and integrated to evaluate and rank the data source's importance.

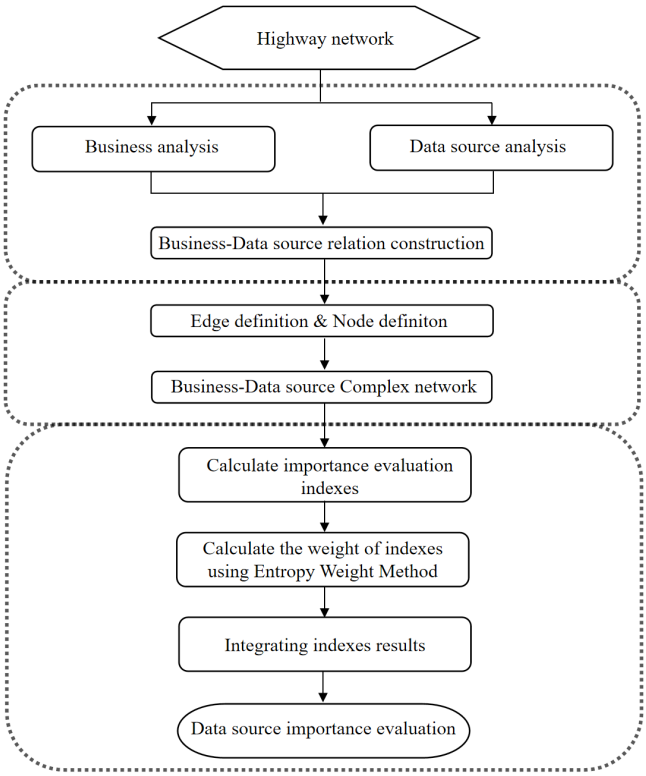

Figure 1 – Method framework

3.2 Business-data source relation construction

Recall that business models such as incident detection are essential for highway network management, and data items collected from each data source are vital for supporting these business models. Therefore, it is straightforward to capture the relationship between the business models and the data sources, i.e. the business-data source relation, such that the data sources can be investigated in the context of highway network management business models.

Business-data source relation is the connection that relates business processes to data sources in the highway networks. In this relationship, the businesses refer to the management tasks that highway management agencies have to perform to operate the highway network. For each business process, the process can be decomposed into business data items which are the inputs required to implement the business process. Naturally, a business process may have multiple business data items. For example, detecting an incident might need traffic volume and speed simultaneously. In contrast, the data sources refer to the detecting techniques or facilities that can provide operational data on the highway network. Also, a data source might include multiple data source fields, and different data sources can generate the same data source field. For example, traffic volume can be collected from continuous traffic count stations or electronic toll collection systems, and a count station can also collect volume, speed, or vehicle types at the same time.

After breaking the business processes into business data items and the data sources into data source fields, business data items to data source fields can be related, formulating a relationship as depicted in *Figure 2*. There will be a multitude amount of such relationships, and combined, all the relationships constitute the foundation for structuring the complex network to be used for data source importance evaluation. It is worthwhile to mention that these relationships are critical for the proposed approach in that they make it possible to evaluate the importance of different data sources in terms of highway network management tasks.

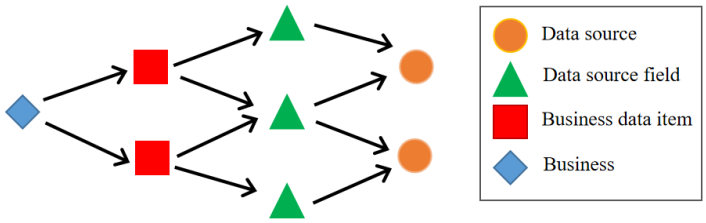

Figure 2 – Diagram of business-data source relation

3.3 Business-data source complex network construction

As mentioned previously, establishing the business-data source relations provides a possibility of investigating the data sources in the context of business models. However, as can be seen in the previous section, for each business model, this paper can develop an individual business-data source relation diagram, and with the increase of business models, the diagrams will increase too, yielding the analysis almost intractable. Therefore, under this circumstance, this paper proposes to use a complex network to integrate these business-data source diagrams, i.e. to construct a business-data source complex network, to make the analysis of data source importance trackable.

Upon the above discussion, after obtaining the business-data source relations, this paper constructs the business-data source complex network through a process of integration. For a complex network, the nodes and the edges need to be defined first. In this paper, the node definition is straightforward, that is, 4 types of nodes, i.e. business node, business data item node, data source field node, and data source node, are defined according to the business-data source relation. Upon definition of the nodes, the connections in the business-data source relation diagram are then naturally defined as the edge, and the direction of the edges starts in an order of data source node, data source field node, business data item node and business node. Note that at this time all the edges have the same weight. Therefore, in order to simplify the complex network, this paper keeps only one edge between two nodes and defines the weight of the edge as the total number of edges between two nodes before the simplification. Upon this integration process, the schematic diagram of

the complex network is shown in *Figure 3*. Intuitively, from a connectivity point of view, if the weight of an edge is high, then the nodes connected by the edge will likely be more important.

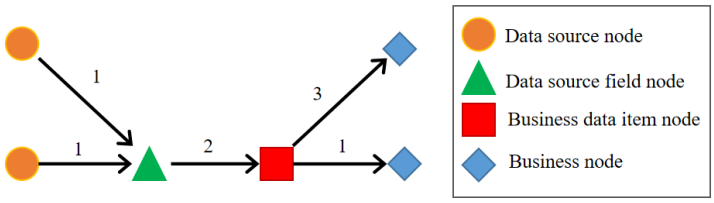

Figure 3 – Schematic diagram of the complex network

3.4 Centrality index computation

Upon construction of the business-data source complex network, instruments offered by the complex network theory can then be applied to evaluate the importance of the nodes. In this paper, three centrality indexes can be used to evaluate the importance of data source nodes, including degree centrality (DC), closeness centrality (CC) and structural holes (SC). These three indexes are defined in *Equation 1*.

$$\begin{cases} C_D(v_i) = \dfrac{k_i}{n-1} \\ C_c(v_i) = \dfrac{n-1}{\sum_{i \neq j} d_{ij}} \\ S_i = \sum_j (p_{ij} + \sum_q p_{iq} \cdot p_{qj})^2 \end{cases} \quad (1)$$

where $C_D(v_i)$ is the degree centrality of node v_i, n is the number of nodes in the network, k_i is the degree value of node v_i , indicating the number of edges associated with the node, $C_c(v_i)$ is the closeness centrality of node v_i, d_{ij} is the shortest distance from node v_i to node v_j, S_i is the network constraint coefficient concerning the structure hole, p_{ij} is the proportion of the number of connections between node v_j and node v_i to all connections of node v_i, and q is the common neighbours of nodes v_i and j and is not equal to i and j.

In a complex network, degree refers to the number of adjacent edges of a node, and degree centrality of a node refers to its centrality level among neighbouring nodes directly connected to it. If the degree centrality of a node is higher, the importance of the node is higher. Closeness centrality considers the average length of the shortest path from a node to other nodes. If the closeness centrality of a node is greater, the importance of the node is greater, indicating that it is located at the centre of the network. A structural hole indicates a disconnection between the nodes in the complex network. For node v_i structural hole index can be obtained by calculating the network constraint coefficient. The smaller the constraint coefficient, the more likely it is to form a structural hole [34].

3.5 Calculate index weight

The previous section provides three indexes for evaluating the importance of data sources, and each of these indexes can be applied individually. However, current literature shows that evaluation using multi-indexes integration is more effective than a single index [35, 36]. In this sense, determining the weights becomes important for integrating these indexes.

Therefore, in this section, this paper computes the weights for integrating the centrality indexes computed previously. For this purpose, the centrality indexes need to be normalised first to eliminate the impact of different dimensions. For normalisation, the node set definition is $V = \{v_1, v_2, \cdots, v_n\}$, and the centrality indexes set is $S = \{s_1, s_2, ..., s_m\}$, and a decision matrix X is formulated as *Equation 2*. Note that in X, DC and CC are regarded as benefit indexes, and SC is regarded as cost index. The formulas for normalising benefit indexes and cost indexes are shown in *Equations 3 and 4*, respectively.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1m} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2m} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{nm} \end{bmatrix} \tag{2}$$

$$f_{ij} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}} \tag{3}$$

$$f_{ij} = \frac{x_j^{\max} - x_{ij}}{x_j^{\max} - x_j^{\min}} \tag{4}$$

where x_{ij} is the index value of node v_i under index s_j, f_{ij} is the normalised value of node v_i under index s_j, x_j^{min} is the minimum value among all nodes under index s_j, and x_j^{max} is the maximum value among all nodes under index s_j.

Then, entropy is used as the weight for index integration. For this purpose, the information entropy E_j of each index is computed as *Equations 5 and 6*.

$$r_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n} f_{ij}} \tag{5}$$

$$E_j = -\frac{\sum_{i=1}^{n} r_{ij} \cdot \ln r_{ij}}{\ln n} \tag{6}$$

where r_{ij} is the proportion of the standardised index value of node v_i to the total standardised index values of all nodes under index s_j.

Finally, the weight of each index is computed using the information entropy as *Equation 7*.

$$\omega_j = \frac{1 - E}{m - \sum E_j} \tag{7}$$

where ω_j is the weights of each index s_j, and m is the number of indexes, i.e. m=3. Note that for each index, the weight indicates the dispersion of the index, and greater weight indicates greater importance for the index in integration.

3.6 Indexes integration and importance evaluation

In this section, this paper first uses the index weights calculated in the previous step to calculate the relative distances between different data source nodes and the most important data source nodes, as in *Equation 8*. The smaller the relative distance, the more important the node is.

$$\begin{cases} H_i = \sum_{j=1}^{m} \dfrac{\omega_j \cdot (f_j^+ - f_{ij})}{f_j^+ - f_j^-} \\ R_i = \max\left[\dfrac{\omega_j \cdot (f_j^+ - f_{ij})}{f_j^+ - f_j^-}\right] \end{cases} \tag{8}$$

where H_i is the weighted sum of the relative distances, R_i is the maximum weighted relative distance, f_j^+ is the most important data source node under index s_j, and f_j^- is the least important data source node under index s_j. Note that $\begin{cases} f_j^+ = \max(f_{ij}) \\ f_j^- = \min(f_{ij}) \end{cases}$ is for benefit indexes and $\begin{cases} f_j^+ = \min(f_{ij}) \\ f_j^- = \max(f_{ij}) \end{cases}$ is for cost indexes.

Then the integrated index Q_i, or the comprehensive relative distance, is computed as *Equations 9 and 10*.

$$\begin{cases} H^- = \min H_i \\ H^+ = \max H_i \\ R^- = \min R_i \\ R^+ = \max R_i \end{cases} \tag{9}$$

$$Q_i = \sigma \cdot \left[\frac{H_i - H^-}{H^+ - H^-} \right] + (1-\sigma) \cdot \left[\frac{R_i - R^-}{R^+ - R^-} \right] \tag{10}$$

where H^+ is the largest weighted sum of the relative distances, H^- is the smallest weighted sum of the relative distances, and σ is the adjustment coefficient set as 0.5.

Finally, the importance of data source nodes is ranked according to Q as *Equations 11 and 12*. Note that smaller T_i' indicates a closer relative distance between the node and the most important node, meaning that the node is more important.

$$\boldsymbol{T} = (Q_1, Q_2, ..., Q_n) \tag{11}$$

$$\boldsymbol{T'} = \text{sort}(\boldsymbol{T}) = (T_l', ..., T_i', ..., T_n') \tag{12}$$

where $\boldsymbol{T}$ is the node importance vector and $T_i' \in [Q_1, \ldots, Q_i, \ldots, Q_n]$ with $T_i' \leqslant T_{i+1}'$, $i \in [1, n]$.

4. CASE STUDY

This paper uses the highway network in Xuzhou City, Jiangsu Province, China for the case study.

4.1 Highway network introduction

The selected highway network in Xuzhou City has 8 national roads and 20 provincial roads, and by the year 2022, the total mileage was 1,590.6 kilometres. The density of the highway network is 141.29 kilometres per 100 square kilometres or 17.62 kilometres per 10000 people. This highway network constitutes the backbone of the surface transportation systems in Xuzhou city.

On this highway network, 16 major business processes or business tasks, with the business ID (identification) denoted by D_1, …, D_{16}, are conducted shown and explained in *Table 1*. To support these businesses, 6 major data collection methods or data sources under evaluation, with the data source ID denoted by A_1, …, A_6, have been established as shown in *Table 2*. According to *Table 2*, a continuous traffic survey station collects real-time vehicle flow information to form a large amount of traffic data. A vehicle detector-based system collects data such as vehicle speed, occupancy rate, and axle load through vehicle detection equipment. A manual toll collection system collects vehicle information when collecting the toll manually for a vehicle. An automatic gantry-based station collects and stores the vehicle data when collecting the toll automatically without stopping. Video-based vehicle detection system collects traffic data through real-time video analysis. Basic information on emergency bases mainly includes the type and amount of emergency resources, emergency base identification number, and geographical location of the emergency base, for supporting the operation of emergency resource scheduling.

4.2 Business-data source relation construction

As shown in *Tables 1* and *2*, after analysing the business processes and data sources in the Xuzhou highway network, the business-data source relation will be constructed. In doing so, this paper breaks the business processes into business data items first and breaks the data sources into data source fields, as shown in *Tables 3* and *4*. Note that the business data item indicates the conceptual or logical data item that will be used in the business model which can be found by examining the data flow of conducting the business model, and the data source fields indicate the physical or actual data field that can be collected from each data source. A business data item and data sources field are different in that a business data item might be directed to

Table 1 – Business Models Summary

Business ID	Business name	Explanation
D_1	Traffic volume measurement	Measure the number of vehicles within a time interval
D_2	Standardised traffic flow measurement	Measure the number of passenger cars within a time interval
D_3	Average speed measurement	Measure the average speed of the vehicles within a time interval
D_4	Truck flow measurement	Measure the number of trucks within a time interval
D_5	Flow saturation rate measurement	Measure the ratio of traffic flow to the capacity of a road section
D_6	Road congestion index measurement	Measure the index reflecting the level of road congestion
D_7	Average travel time measurement	Measure the average travel time of vehicle within a time interval
D_8	Travel time index measurement	Measure the travel time index based on travel times
D_9	Travel time dispersion interval measurement	Measure the dispersion interval for vehicle travel times
D_{10}	Travel time variation coefficient measurement	Measure the dispersion coefficient of vehicle travel times.
D_{11}	Traffic density measurement	Measure the number of vehicles on a road section
D_{12}	Incident detection	Detect the occurrence of incident on a road section
D_{13}	Traffic flow prediction	Anticipate traffic flow in the near future for a road section
D_{14}	Incident impact range prediction	Predict the influence range of an incident
D_{15}	Emergency resource scheduling	Allocate emergency management resources
D_{16}	Travel path planning	Plan route for vehicles

Table 2 – Data source summary

Data source ID	Data source name
A_1	Continuous traffic survey station
A_2	Vehicle detector-based system
A_3	Manual toll collection system
A_4	Automatic gantry-based station
A_5	Video-based vehicle detection system
A_6	Basic information of emergency bases

multiple data source fields, i.e. a business data item might be supported by multiple data sources. After the breaking process, the definitions of the categories and the items in each category are listed for data source fields and business data items in *Table 3*, with the items denoted by B_1, …, B_{87} and C_1, …, C_{24}, respectively. As shown in *Table 3*, there are 24 business data items and 87 data source fields in total. In addition, the explanations of these items are listed in *Table 4*.

Table 3 – Definition of business data item and data source field

Category	Items	
Data source field	B_1: data collection time B_2: data record ID B_3: device ID code B_4: lane ID B_5: road section ID code B_6: road section name B_7: administrative jurisdiction code B_8: site ID code B_9: upstream small truck volume B_{10}: downstream small truck volume B_{11}: upstream medium truck volume B_{12}: downstream traffic volume of medium truck B_{13}: upstream traffic volume of large truck B_{14}: downstream traffic volume of large truck B_{15}: upstream passenger car and medium-sized bus volume B_{16}: downstream passenger car and medium-sized bus volume B_{17}: upstream bus volume B_{18}: downstream bus volume B_{19}: upstream oversize truck volume B_{20}: downstream oversize truck volume B_{21}: upstream tractor volume B_{22}: downstream tractor volume B_{23}: upstream container truck volume B_{24}: downstream container truck volume B_{25}: data collection site ID B_{26}: data collection time interval B_{27}: data collection time interval ID B_{28}: small truck traffic volume B_{29}: medium truck traffic volume B_{30}: large truck traffic volume B_{31}: passenger car and medium-sized bus volume B_{32}: bus volume B_{33}: trailer truck volume B_{34}: small trailer truck volume B_{35}: oversize truck volume B_{36}: container truck volume B_{37}: accident lane ID B_{38}: travel time B_{39}: toll station ID B_{40}: vehicle type B_{41}: gantry ID B_{42}: date B_{43}: time B_{44}: road section ID	B_{45}: road section length B_{46}: average upstream small truck speed B_{47}: average downstream small truck speed B_{48}: average upstream medium truck speed B_{49}: average downstream medium truck speed B_{50}: average upstream large truck speed B_{51}: average downstream large truck speed B_{52}: average upstream small and medium buses speed B_{53}: average downstream passenger car and medium-sized bus speed B_{54}: average upstream bus speed B_{55}: average downstream bus speed B_{56}: average upstream oversized truck speed B_{57}: average downstream oversized truck speed B_{58}: average upstream tractor speed B_{59}: average downstream tractor speed B_{60}: average upstream container truck speed B_{61}: average downstream container truck speed B_{62}: average small truck speed B_{63}: average medium truck speed B_{64}: average large truck speed B_{65}: average passenger car and medium-sized bus speed B_{66}: average bus speed B_{67}: average oversized truck speed B_{68}: average tractor speed B_{69}: average container truck speed B_{70}: accident ID B_{71}: accident start time B_{72}: accident end time B_{73}: accident start location B_{74}: accident end location B_{75}: number of lanes B_{76}: number of accident occupied lanes B_{77}: traffic direction B_{78}: accident classification B_{79}: number of involved vehicles in accident B_{80}: number of involved large car in accident B_{81}: number of involved passenger car in accident B_{82}: number of casualties B_{83}: weather B_{84}: emergency resource type B_{85}: emergency resource reserve B_{86}: emergency base ID B_{87}: emergency base location
Business data item	C_1: current time interval C_2: road section ID C_3: small truck volume C_4: medium truck volume C_5: large truck volume C_6: passenger car and medium-sized bus volume C_7: bus volume C_8: oversized truck volume C_9: tractor volume C_{10}: container truck volume C_{11}: vehicle conversion coefficient C_{12}: average small truck speed	C_{13}: average medium truck speed C_{14}: average large truck speed C_{15}: average passenger car and medium-sized bus speed C_{16}: average bus speed C_{17}: average oversized truck speed C_{18}: average tractor speed C_{19}: average container truck speed C_{20}: design volume C_{21}: design speed C_{22}: accident statistics C_{23}: emergency resources C_{24}: emergency base ID

Table 4 – Explanation of business data item and data source field

Items ID	Explanation
B_1	The time when the traffic detector records the data
B_2	The serial number of the current data recorded
B_3	The ID code of the device
B_4	The serial number of the road lane
B_5	A code that uniquely identifies a section of road
B_6	The name of the road section
B_7	The identification code of the administrative jurisdiction Where the highway network resides
B_8	The unique ID code for the detection site
B_9~B_{24}	The total number of different types of vehicles in two directions per time interval
B_{25}	The serial number of the road section where data collection is conducted
B_{26}	The time interval for data collection
B_{27}	The ID of the time interval for collected data
B_{28}~B_{36}	The total number of different types of vehicles in a time interval
B_{37}	The serial number of the lane where the accident occurred
B_{38}	The duration of vehicles passing the road section
B_{39}	The serial number of the toll station
B_{40}	The types of vehicles distinguished by characteristics, purpose, and function of the vehicle
B_{41}	The serial number of the gantry
B_{42}	A code that uniquely identifies the date
B_{43}	A code that uniquely identifies the time
B_{44}	The serial number of the road section
B_{45}	The length of road section
B_{46}~B_{61}	The average speed in two directions for different vehicle types
B_{62}~B_{69}	The average speed of different types of vehicles
B_{70}	The serial numbers of traffic accidents
B_{71}~B_{72}	The start and end time of the traffic accident
B_{73}~B_{74}	The geographical location where the traffic accident occurred and ended
B_{75}	The total number of lanes in one direction of a road
B_{76}	The total number of lanes occupied by traffic accidents
B_{77}	The direction of vehicular traffic
B_{78}	The specific classification of a traffic accident
B_{79}~B_{81}	The number of different types of vehicles involved in traffic accidents
B_{82}	The number of casualties caused by traffic accidents
B_{83}	The weather conditions at the time of the accident
B_{84}	The type of materials and equipment needed for emergency assistance
B_{85}	The reserves of materials and equipment required for emergency assistance
B_{86}	The serial number of the emergency base
B_{87}	The geographic location of emergency base
C_1	The time interval related to current time
C_2	The series number of the selected road section
C_3~C_{10}	The total number of different types of vehicles within a time interval
C_{11}	The conversion coefficient of different vehicle types to passenger car
C_{12}~C_{19}	The average speed of different types of vehicles observed within a time interval
C_{20}	The traffic volume used for designing the road section
C_{21}	The speed used for designing the road section
C_{22}	The basic statistics related to traffic accidents
C_{23}	The materials and equipment needed in response to emergencies
C_{24}	The series number of the emergency base

After obtaining the business model, business data item, data source, and data source fields, as listed in *Tables 1–3*, the business-data source relations can be constructed. For this purpose, this paper will set up three connections, including the business-to-business data item, the business data item to the data source field, and the data source field to the data source. For these three connections, it is straightforward to see that the connections between business model and business data item and between data source and data source field are the natural results of the breaking process mentioned previously, and therefore, the development of the business-data source relationship relies on essentially the connection between the business data item and data source field. In this study, the authors examined extensively all the business data items and data source fields and hence established the connections manually.

Using traffic volume measurement as a typical example, the relationship is shown in *Figure 4*. As can be seen in *Figure 4*, the first column to the left is the name of the business (D_1), and the next column is the business data items (C_1, ..., C_{10}) required for computing this business, i.e. measuring the traffic volume. For each business data item, the required data source fields are shown in the third column to the left, and the fourth column gives the data sources that will provide these data source fields. Note that each business data item will link to multiple data source fields.

It can be seen that this relationship is complex, and for each business listed in *Table 1*, such a diagram can be formulated. Together, this paper will integrate these diagrams into a complex network for supporting data source importance evaluation.

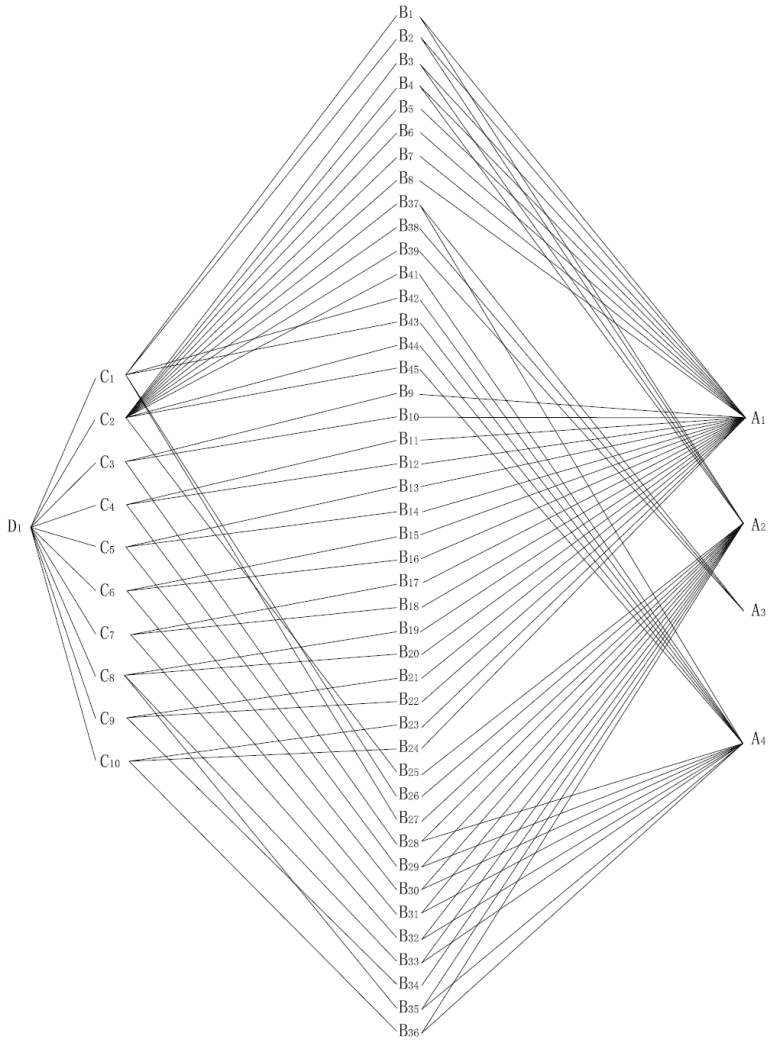

Figure 4 – Diagram for traffic volume and data source relation

4.3 Business-data source complex network construction

After generating all the business-data source relationship diagrams for the businesses listed in *Table 1*, this paper can construct the business-data source complex network as shown in *Figure 5* through integrating all the business-data source diagrams. Note that in the business-data source diagram, each edge or connection will connect two items such as business model, business data item, data source, or data source field, and essentially, during the integration process, this paper will merge all the edges in all the business-data source relationship diagrams together concerning the starting and ending items of each edge. Consequently, all the business models, business data items, data source fields, and data sources, i.e. nodes, are displayed in a single diagram, with the integrated edges connecting all these four types of nodes. In this way, a complex network can be formulated, based on which conventional instruments such as centrality measures can then be applied to evaluate the importance of the nodes. It can be seen clearly that in this complex network, some nodes show more complex connections to the rest of the nodes, implying potentially greater importance for these nodes in the network.

It is also worthwhile to mention that, based on such a complex network, it is possible to evaluate the importance of all types of nodes, while in this paper, only nodes representing the data sources are evaluated, i.e. nodes A_1 to A_6, as denoted by yellow.

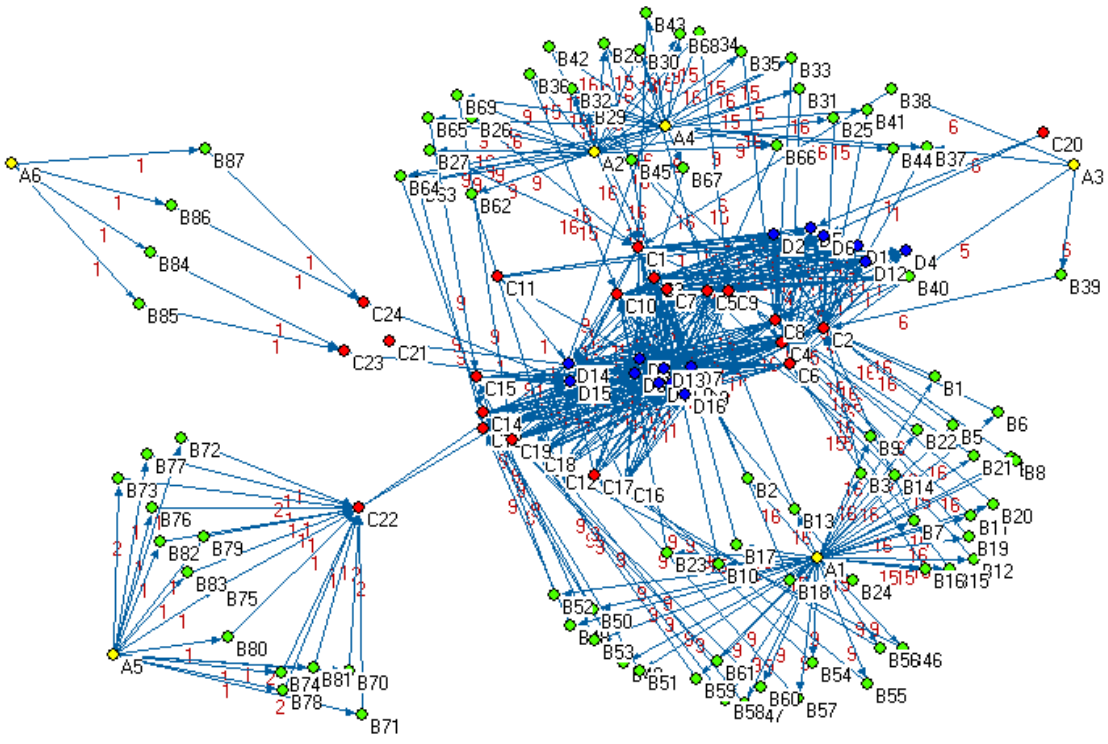

Figure 5 – Business-data source complex network topology

4.4 Centrality index computation

After establishing the complex network, the steps indicated in the flow chart of the proposed method can be followed, as shown in *Figure 1*. According to the flow chart, based on the complex network built in the previous section, in this section, the central indexes for the data source nodes can be computed and ranked as shown in *Table 5*. It can be seen that first, all three indexes indicate that continuous traffic survey station is the most important data source for the highway network. It is not surprising since this data source supplies a rich amount of data on the vehicle passing the detection zone. In fact, this type of data collection facility is the most important source of monitoring the operation of highway networks. In addition, there are also differences in the results for these different indexes on the importance of data source nodes, which indicates a need to use the entropy weight method for index integration.

Table 5 – Ranking of data source node importance under three evaluation indexes

Rank	DC		CC		SC	
	Node ID	Calculated value	Node ID	Calculated value	Node ID	Calculated value
1	A_1	0.3030	A_1	0.3782	A_1	0.0266
2	A_4	0.1742	A_4	0.3483	A_4	0.0458
3	A_2	0.1515	A_2	0.3411	A_2	0.0532
4	A_5	0.1061	A_3	0.2827	A_5	0.0796
5	A_3, A_6	0.0303	A_5	0.2264	A_6	0.2500
6	—	—	A_6	0.2112	A_3	0.2514

4.5 Indexes integration and importance evaluation

Data source importance is further investigated by integrating the three centrality index-based measures in this section. For doing so, the three data source node importance evaluation indexes are used to form the attribute set, i.e. $S = \{s_1, s_2, s_3\} = \{DC, CC, SC\}$. An initial decision matrix is constructed as *Equation 13* based on the results calculated for each index in *Table 5*, which is normalised to a standardised decision matrix. Then, the entropy weight method is used to calculate the weights of each index. According to the method discussed previously, the weights of DC, CC and SC are found to be 0.3419, 0.3269 and 0.3312, respectively. Finally, the importance ranking results of the data source nodes are obtained by integrating the three indexes, with results shown in *Table 6*.

$$X = \begin{bmatrix} 0.1515 & 0.3782 & 0.0266 \\ 0.0758 & 0.3411 & 0.0532 \\ 0.0152 & 0.2827 & 0.2514 \\ 0.0871 & 0.3483 & 0.0458 \\ 0.0530 & 0.2264 & 0.0796 \\ 0.0152 & 0.2112 & 0.2500 \end{bmatrix} \tag{13}$$

Table 6 – The importance ranking of data source nodes

Importance ranking	Node (Name)	Q Value
1	A_1 (continuous traffic survey station)	0.00
2	A_4 (automatic gantry based station)	0.37
3	A_2 (vehicle detector-based system)	0.44
4	A_5 (video-based vehicle detection system)	0.67
5	A_3 (manual toll collection system)	0.94
6	A_6 (basic information of emergency base)	1.00

From *Table 6*, it can be seen that the continuous traffic survey station is the most important data source in the Xuzhou highway network, followed by the automatic gantry-based station and vehicle detector-based system. This finding is consistent with the findings based on individual indexes in that continuous traffic survey stations are the most important data source in supporting highway network operations. In addition, in terms of the degree of importance, the Q values directly represent the proportional relationship of the importance of the data sources. To this end, it can be seen that the importance of a continuous traffic survey station is about 4 times that of an automatic gantry-based station and vehicle detector-based system, about 7 times that of a traffic video-based vehicle detection system, and about 10 times than that of manual toll collection system and basic information of emergency base. Based on the above observations, it is clear that the highway management agency is expected to invest more resources to maintain the continuous traffic survey stations in the highway network.

5. DISCUSSION AND CONCLUSION

Highway systems are important for supporting the social and economic development of the society. With the advancement of information technology, many data sources are formulated based on various vehicle detecting techniques or systems in highway networks, and a rich amount of operational data have been collected and applied in highway network management. Under this circumstance, there exists a need for allocating appropriately the resources for constructing and maintaining such facilities in terms of supporting the highway network management tasks, generating consequently an issue of identifying and evaluating the importance of such data sources. Targeting this issue, this paper proposes a complex network based evaluation method. In this approach, the businesses and data sources in a highway network are systematically analysed, after which these businesses and data sources are broken into business data items and data sources fields, respectively. Then the relationships between the business, business data item, data source, and data source field are formulated, based on which a business-data source complex network can be built through integrating such relationships. Data source importance evaluation is then conducted by using separately and collectively the centrality indexes obtained from the complex network.

The application of the proposed complex network based approach is illustrated using a highway network in Xuzhou City, Jiangsu Province, China. For this highway network, this paper analysed the relationship between highway network management businesses and data resources and built a business-data source complex network accordingly. The importance of data source nodes was then computed and ranked based on the complex network. The results show that among multiple data sources in the highway network, the continuous traffic survey station is the most important data source. The proportional importance is then shown for these data sources. Upon the results, it can be inferred that continuous traffic survey stations are expected to receive more attention when allocating construction and maintenance resources.

Discussions are provided concerning the limitations, applications, and future studies on the proposed approach as follows. First, as shown in this work, the construction of the business-data source relation is critical for building the complex network for importance analysis. Currently, this relationship is built based on manual investigations of the business models and data sources, which is time-consuming and requires a high degree of familiarity of the analyser in the highway network management field. This manual solution limits the application of the proposed approach to other highway networks.

Second, the application of the proposed approach can be multi-fold. As shown previously, the identified importance index for data sources can be directly applied to generate the plan for data collection facilities construction and maintenance. In addition, business tasks, in particular, business data items can be investigated to show their relative importance for highway network management. Moreover, the complex network developed in the proposed approach can be applied to optimise the design of data collection facilities in terms of supporting highway network management tasks.

Finally, multiple future work can be conducted. Artificial intelligence technology can be applied to mine and build relationships from the documents concerning highway network business tasks and data sources. This will help improve the efficiency of developing and applying such complex networks onto more complicated highway networks. In addition, more centrality indexes and other weighting techniques can be explored under this complex network framework. Moreover, the proposed approach can be applied and tested to allocate resources for highway network data collection facility construction and maintenance.

ACKNOWLEDGEMENTS

This work is supported by the Key Laboratory of Transport Industry of Comprehensive Transportation Theory (Nanjing Modern Multimodal Transportation Laboratory) through the open project No. MTF2023005.

REFERENCES

[1] Federal Highway Administration. *Traffic monitoring guide*. Washington D.C, USA: U.S. Department of Transportation Federal Highway Administration; 2022. https://www.fhwa.dot.gov/policyinformation/tmguide/ [Accessed 18th June 2023].

[2] Kamouch A, Chaoub A, Guennoun Z. Mobile big data in vehicular networks: The road to internet of vehicles. In: Skourletopoulos G, et al. (eds.) *Mobile big data. Lecture Notes on Data Engineering and Communications Technologies*. Cham, Switzerland: Springer; 2018. p. 129-143. DOI: 10.1007/978-3-319-67925-9_6.

[3] Chan Y. Telecommunications-and information technology-inspired analyses: Review of an intelligent transportation systems experience. *Transportation Research Record*. 2017;2658(1):44-55. DOI: 10.3141/2658-06.

[4] Huang Y, et al. Spatiotemporal approach for evaluating the vehicle restriction policy with multi-sensor data. *Transportation Research Record*. 2022;2676(8):724-736. DOI: 10.1177/03611981221085518.

[5] Levenberg E, et al. Live road condition assessment with internal vehicle sensors. *Transportation Research Record*. 2021;2675(10):1442-1452. DOI: 10.1177/03611981211016852.

[6] Seedah DPK, Sankaran B, O'Brien WJ. Approach to classifying freight data elements across multiple data sources. *Transportation Research Record*. 2015;2529(1):56-65. DOI: 10.3141/2529-06.

[7] Robichaud K, Gordon M. Assessment of data-collection techniques for highway agencies. *Transportation Research Record*. 2003;1855(1):129-135. DOI: 10.3141/1855-16.

[8] Hasnat MM, Bardaka E. Distribution of highway infrastructure cost responsibility and revenue contribution shares among highway users in North Carolina: Present conditions and future alternatives. *Transportation Research Record*. 2023;2677(2):1082-1102. DOI: 10.1177/03611981221112403.

[9] Chen T, Ma J, Zhu Z, Guo X. Evaluation method for node importance of urban rail network considering traffic characteristics. *Sustainability*. 2023;15(4):3582. DOI: 10.3390/su15043582.

[10] Liu S, Gao H. The structure entropy-based node importance ranking method for graph data. *Entropy*. 2023;25(6):941. DOI: 10.3390/e25060941.

[11] Zhang Y, Lu Y, Yang G, Hang Z. Multi-attribute decision making method for node importance metric in complex network. *Applied Sciences*. 2022;12(4):1944-1944. DOI:10.3390/APP12041944.

[12] Sotoodeh H, Falahrad M. Relative degree structural hole centrality, $C_{RD\text{-}SH}$: a new centrality measure in complex networks. *Journal of Systems Science & Complexity*. 2019;32(05):1306-1323. DOI: 10.1007/s11424-018-7331-5.

[13] Yu H, Cao X, Liu Z, Li Y. Identifying key nodes based on improved structural holes in complex networks. *Physica A: Statistical Mechanics and its Applications*. 2017;486(C):318-327. DOI: 10.1016/j.physa.2017.05.028.

[14] Çalık A, Erdebilli B, Özdemir YS. Novel integrated hybrid multi-criteria decision-making approach for logistics performance index. *Transportation Research Record*. 2023;2677(2):1392-1400. DOI: 10.1177/03611981221113314.

[15] Chen C, Zhang H. Evaluation of green development level of Mianyang agriculture, based on the entropy weight method. *Sustainability*. 2023;15(9):7589. DOI: 10.3390/SU15097589.

[16] Hanson-DeFusco J. What data counts in policymaking and programming evaluation-relevant data sources for triangulation according to main epistemologies and philosophies within social science. *Evaluation and Program Planning*. 2023;97(3):102238. DOI: 10.1016/j.evalprogplan.2023.102238.

[17] Park N, et al. Estimating node importance in knowledge graphs using graph neural networks. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 4-8 Aug. 2019, Anchorage, USA*. 2019. p. 596-606. DOI: 10.1145/3292500.3330855.

[18] Narayan VV, et al. Evaluation of data sources and approaches for estimation of influenza-associated mortality in India. *Influenza and Other Respiratory Viruses*. 2018;12(1):72-80. DOI: 10.1111/irv.12493.

[19] Price C, Burley RA. An evaluation of information sources for current awareness on occupational diseases. *Journal of Information Science*. 1986;12(5):247-255. DOI: 10.1177/016555158601200504.

[20] Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology*, 1996;25(2):435-442. DOI: 10.1093/ije/25.2.435.

[21] Hjørland B. Evaluation of an information source illustrated by a case study: effect of screening for breast cancer. *JASIST*. 2011;62(10):1892-1898. DOI: 10.1002/asi.21606.

[22] Hjørland B. Methods for evaluating information sources: an annotated catalogue. *Journal of Information Science*. 2012;38(3):258-268. DOI: 10.1177/0165551512439178.

[23] Kaufmann D, Kraay A, Mastruzzi M. The worldwide governance indicators: methodology and analytical issues. *Hague Journal on the Rule of Law*. 2010;3(2):220-246. DOI: 10.1017/S1876404511200046.

[24] Wood S, Regehr JD. Hierarchical methodology to evaluate the quality of disparate axle load data sources for pavement design. *Journal of Traffic and Transportation Engineering (English Edition)*. 2022;9(2):261-279. DOI: 10.1016/J.JTTE.2021.02.005.

[25] Broach J, et al. Evaluating the potential of crowdsourced data to estimate network-wide bicycle volumes. *Transportation Research Record*. 2024;2678(3):573-589. DOI: 10.1177/03611981231182388.

[26] Jiang R, et al. Predicting bus travel time with hybrid incomplete data: A deep learning approach. *Promet - Traffic & Transportation*. 2022;34(5):673-685. DOI:10.7307/PTT.V34I5.4052.

[27] Yang L, Maria ST, Breitfuss G. Data sources in data driven circular business models. *New Business Models Conference Proceedings 2023. 21-23 Jun 2023, Maastricht, Netherlands*. 2023.. DOI: 10.26481/mup.2302.21.

[28] Khorashadizadeh H, Tiwari S, Groppe S. A survey on covid-19 knowledge graphs and their data sources. In: Mohanty SN, Diaz VG, Kumar GS. (eds.) *Intelligent systems and machine learning. ICISML 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Cham, Switzerland: Springer; 2023. p. 142-152. DOI: 10.1007/978-3-031-35078-8_13.

[29] Wan H, et al. Research on ship relation graph analysis driven by multi-source data, *2021 6th International Conference on Transportation Information and Safety (ICTIS), 22-24 Oct. 2021, Wuhan, China*. 2021. p. 655-660. DOI: 10.1109/ICTIS54573.2021.9798661.

[30] Kam KA, et al. Finding and exploring use of commodity-specific data sources for commodity flow modeling. *Transportation Research Record*. 2017;2646(1):77-83. DOI: 10.3141/2646-09.

[31] Nguyen K, Cao J. Top-K data source selection for keyword queries over multiple XML data sources. *Journal of Information Science*. 2012;38(2):156-175. DOI: 10.1177/0165551511435875.

[32] Tok AYC, et al. Online data repository for statewide freight planning and analysis. *Transportation Research Record*. 2011;2246(1):121-129. DOI: 10.3141/2246-15.

[33] Tijssen R, Raan TV, Heiser W, Wachmann L. Integrating multiple sources of information in literature-based maps of science. *Journal of Information Science*. 1990;16(4):217-227. DOI: 10.1177/016555159001600402.

[34] Wang W, et al. Factors affecting unmanned aerial vehicles' unsafe behaviors and influence mechanism based on social network theory. *Transportation Research Record*. 2023;2677(5):1030-1045. DOI: 10.1177/03611981221138782.

[35] Batista NA, et al. Dealing with data from multiple web sources. *WebMedia '18: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web. 16-19 Oct. 2018, Salvador, Brazil*. 2018. p. 3-6. DOI: 10.1145/3243082.3264609.

[36] Krogstie J. Evaluating data quality for integration of data sources. In: Grabis J, Kirikova M, Zdravkovic J, Stirna J. (eds.) *The Practice of Enterprise Modeling. PoEM 2013. Lecture Notes in Business Information Processing*. Berlin, Germany: Springer; 2013. p. 39-53. DOI: 10.1007/978-3-642-41641-5_4.

黄煌钦，郭建华，史祥雨，申雷霄

公路网数据源的重要性评估：一种基于复杂网络的方法

摘要：

数据采集技术或数据源对普通国省道路网的管理至关重要。然而，为了合理分配政府有限的管理资源，有必要确定普通国省道路网中多个数据源的重要性。因此，本文提出一种基于复杂网络的方法来评估普通国省道路网中多个数据源的重要性。该方法主要包括三个步骤。首先，识别和建立普通国省道路网的业务-数据源关系；然后，由此关系构建业务-数据源复杂网络；最后，利用熵权法将该复杂网络的度中心性（DC）、接近中心性（CC）和结构洞（SC）三个指标相融合，以此对数据源节点的重要性进行计算和排序。本文以中国江苏省徐州市的普通国省道路网开展实例分析，结果发现，重要性最高的数据源是连续交通情况调查站，其次是ETC系统和车辆检测器系统。本文讨论了所提方法的局限性、拓展应用和未来研究。

关键词：普通国省道路网运行；数据源；复杂网络；中心性；熵权法