Demonstration of Applicability of Qualitative Data in Origin-Destination Matrix Estimation

Christopher Job MUKUYE[1], Nobuhiro UNO[2] and Jan-Dirk SCHMOECKER[3]

[1]Masters Student, Department of Urban Management, Kyoto University
(1-2-438 katsura Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: kris@trans.kuciv.kyoto-u.ac.jp
[2]Associate Professor, ITS laboratory, Department of Urban Management., Kyoto University
(C1-2-436 katsura Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: uno@trans.kuciv.kyoto-u.ac.jp
[3]Associate Professor, ITS laboratory, Department of Urban Management, Kyoto University
(C1-2-436 katsura Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: schmoecker@trans.kuciv.kyoto-u.ac.jp

This paper presents a proposed methodology of utilising network describing qualitative data collected from traffic police and other traffic authorities in the Classical gravity model method of Origin-Destination matrix estimation. More unobserved link flow volumes are postulated by the fuzzy inference approach and are used to supplement the observed traffic counts utilised in the calibaration of the gravity model. The methodology is tested on a hypothetical 3*3 network & an improvement in the accuracy of the estimated OD matrix is to tested

Key Words : *OD estimation, gravity model, qualitative data, traffic counts*

1. INTRODUCTION

OD matrices obtained through a large scale survey such as home or roadside interviews are deemed time consuming and very expensive to implement hence the use of low cost and easily available data seems very attractive. several researchers have developed OD estimation methods utilizing traffic counts that are relatively inexpensive to collect, hardly distract traffic and are beneficial to other studies as well. An OD matrix estimated from observed link counts reflects the current distribution of the travel patterns and hypothetically, the more the traffic count stations, the more accurate the estimated OD matrix[3)].
Despite several studies and classification systems of methods utilising traffic counts in OD matrix estimation by different authors, the basic concept common to all is that the most likely OD matrix is obtained by obtaining consistency between the estimated and observed traffic flows[4)].
In many developing countries whose main transport mode is by road, traffic officers are placed strategically in order to control traffic during congestion and these personel coupled with transportation authorities know a great deal of information about the network as a whole that could be of potential use in transport modeling. This study therefore seeks to make use of this, possibly qualitative, information by modeling it, quantifying it and postulating more link flows that could supplement the observed link counts utilized in the calibration of the classical gravity model for OD estimation.

2. LITERATURE REVIEW

Knowing the travel pattern of any city facilitates easy transport planning decision making, however, this comes with the burden of carrying out comprehensive studies about the transport network. In literature, OD estimation has proven undoubtedly very relevant in this aspect albeit the financial and resource implications that are more pronounced in developing countries whose budgets are grossly stretched. In this regard, more emphasis and continuous research is being carried out in pursue of cheaper alternative ways of attaining the same.

(2.1) Classical Gravity Model for OD Estimation

Observed link counts have for years been considered informative enough on the OD matrix and network characteristics of any network and more studies are currently under way to devise cheaper, easy to manipulate ways of utilizing traffic counts in

OD estimation. As much as OD estimation with traffic counts is considered a cheaper way of matrix estimation, the selection criteria for the form of model formulation most appropriate for different practical scenerios is unknown[4]. Because of the above flexibility in model formulation, this permits several models to be explored. The method considered in this study is the gravity model calibration method where trip patterns, which are functions of the travel impedance between zones as well production and attraction factors, are sought in a way that minimizes the difference between the modelled and observed link flows. Quite often, only few link flows are observed compared to the size and complexity of their respective networks. Therefore, devising a way to postulate more 'close to accurate' link flows to supplement the physically observed ones within the estimation process would ideally improve on the accuracy of the OD matrix estimate.

Depending on the available data that is zone specific, the gravity model could be expressed mathematically in its general form as;

$$T_{ij} = \alpha U_i R_j f(C_{ij}) \quad (1)$$

Where T_{ij} is the demand from origin *i* to destination *j*, U_i & R_j represent zonal factors of production and attraction respectively (e.g population of origin and destination zones or population and employment in origin and destination zones respectively), $f(C_{ij})$ represents the travel cost function from *i* to *j* and α represents the constant of proportionality that can also be split into origin specific and destination specific balancing factors.

With an aim to estimate the OD trip table replicating the existing flows, the above travel demand is assigned to the network to generate link flows. These are then compared to the observed link flows & the impedance parameter together with the constant of proportionality are calibrated to minimize the difference between the modelled & observed link flows. This procedure is iterative in nature & is repeated until a stable proportionality parameter is obtained. Upon achieving this, the final OD obtained in the last iteration is taken to be the estimated OD matrix. This procedure is illustrated as shown in the figure below as adjusted from Holm et al[2].

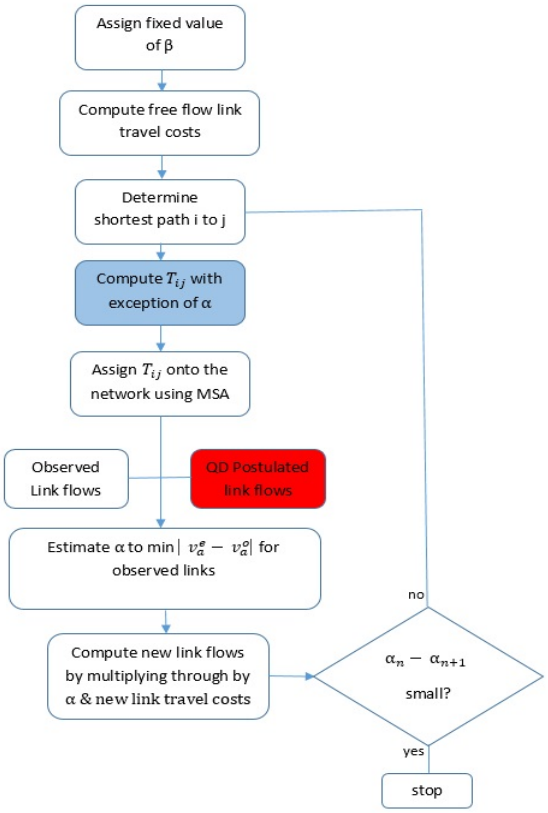

Figure 1 A flowchart showing methodology approach as adjusted from Holm et al

2.2 Qualitative Data in transport

The application of qualitative data in modelling transport is not new as various studies have somehow directly or indirectly incorporated this kind of data. The evolution of today's digital age has accelerated the creation of massive potential data sources that are mostly qualitative in nature through online media. In exploring the capacity of social media data for modelling travel demand, a discussion is presented how this data can be used to extract socio-demographic attributes and travel attributes like trip purpose, trip mode, activity duration as well as land use variables[8].

Zangiabadi et al acknowledges that consideration of qualitative information in any multi objective transportation problem is scarce in literature as in most transport studies, the existing objectives are considered by quantitative information even though there exists a variety of equally important qualitative information to consider such as public health, safety and comfort[5].

There are barriers to the adoption of qualitative data to shape transport policy which include generalisability of findings, issues to do with the perceived subjective nature of findings, the amount of data generated and the positioning of such findings to a policy addueince. Examples of transport studies utilizing qualitative data include data collected to look at the public acceptability of road pricing, to understand the reasons

and motivations behind attitudes held, examining older people's motivations and needs for travel and understanding the social nature of road user safety amongst the public[6].

Hellinga et al describes a prototype data fusion system that can integrate information from loop detectors, probe vehicles, and driver-based linguistic reports (e.g. cellular phone reports) to provide a time-varying estimate of link travel times for the entire traffic network. A sub-model was developed for each individual data source to transform the incoming data into an estimate of link travel time. These link travel time estimates are then combined to provide a composite estimate. Driver reports are linguistic and often qualitative, rather than numerical and quantitative. e.g. a driver reporting that the road on which she is travelling is severely congested. While this description contains useful information pertaining to the state of the traffic network at the specified location and time, there is some uncertainty associated with the precise meaning of the term "severely congested". This necessitates the linguistic report be interpreted and translated into some quantitative descriptor. The Driver Report sub-model is built on the assumption that driver linguistic reports are inherently imprecise in their description of traffic condition and following this assumption, a fuzzy logic rule-based model for estimating delay as a function of driver reports was formulated. Fuzzy membership functions were defined for the driver report information items (i.e. volume, speed, queue length, incident severity, incident time remaining) and the fuzzy logic system used was based on the Mamdani inference system, in which a set of if-then rules is defined[7]. In this study, we are looking at postulating more link flows or possible link flow ranges of unknown links based on traffic police reports as compared to link travel times in the former

3. QUALITATIVE DATA COLLECTION

Having obtained the observed link flow set, qualitative data is collected in form of questionnaires administered to several traffic police and transport authorities to be filled out during or at the end of each day, with specific questions targeting the assessment of traffic scenarios on different links per day. All this information collected is in form of linguistic variables e.g queue length with linguistic values like {short, medium, long}, traffic volume with linguistic values like {light, medium, heavy} and possibly many more . Several Links are assessed including those whose traffic volume is known as shown in Table 1 and 2. The respondent would just have to check the box applicable according to his judgement of all the roads.

Table 1 Example of questionnaire format

Linguistic Variable: TRAFFIC VOLUME			
	Linguistic Value		
Road A	light ☐	medium ☐	heavy ☐
Road B	light ☐	medium ☐	heavy ☐
Road C	light ☐	medium ☐	heavy ☐

Table 2 Example of questionnaire format

Compared to road A, please assess flow of vehicles per hour during morning rush hour on following roads?			
Road B	less ☐	similar ☐	more ☐
Road C	slow ☐	medium ☐	fast ☐

The collected information is preliminary checked to sieve out possible simple mistakes. In line with achieving unbiased data, the various questionnaires from various respondents are compared to check if they are all agreement. Thereafter, unknown link flows or their respective ranges are postulated from the observed link flows based on the comparative information in the table above, by utilizing the fuzzy inference theory with predetermined membership functions. This information could be collected over a period of time to investigate its consistency. Any other information, not catered for by the questionnaire, but provided by the respondents is as well collected. Other possible methods besides the fuzzy inference theory are also explored like the Delphi technique which aims at achieving a convergence of opinion. The postulated link flows or resulting link flow constraints are then added to the observed link flows and the OD is estimated using the classical gravity model where the calibration of the parameters takes into consideration the approximate link flows processed from the qualitative data. An accuracy comparison is also made between the OD's estimated with and without input from Qualitative information.

4. METHODOLOGY

The following section first discusses the estimation of link flows using qualitative data obtained through questionnaire data and then proceeds to OD estimation by calibration of the gravity model.

4.1 Link flow estimation from qualitative data

Based on the fuzzy logic inference approach, fuzzy membership functions are determined for the selected linguistic variables by considering the respective variables and their values for given links with observed traffic counts. these variables include link speed, length of queue, traffic volume and clearance time in case of congestion.

Consider A to be the set of network links with

$$A = A^o \, U \, A^u \, U \, A^f \tag{2}$$

Where A^o refers to the set of links with observed traffic volume, A^u refers to the set of links with unknown link volumes and A^f refers to the set of links with linkflows estimated from fuzzy inference approach. Based on a series of fuzzy if-then rules, each rule is evaluated according to the respondent's data and the results of each rule are aggregated. An estimate of the respective value or range is obtained upon computing the centroid of the aggregated membership shape or in comparison with the results of the observed links. There is an expected challenge in the definition of the appropriate membership functions however, a comparative assessment of the observed and unobserved links will aid in alleviating this. The figure below represents a triangular membership function for link travel time.

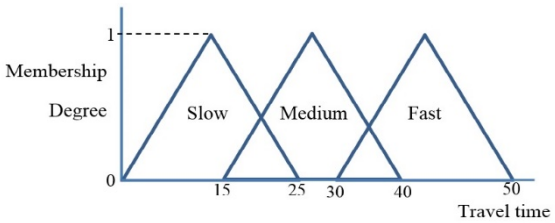

Figure 2 Example of triangular travel time membership function

4.2 Gravity Model Parameter Calibration

There are various techniques that can be employed in calibration however we adopt the method as described by Holm et al that utilizes the actual network traffic volumes for calibration.
Consider a gravity model of the type;

$$T_{ij} = \alpha U_i R_j (C_{ij})^{\beta} \tag{3}$$

Where; C_{ij} refers to the travel cost between *i* and *j* and β is an impedance parameter.
The above model is associated with a multi-path flow dependant traffic assignment algorithm considering link capacity. The minimum travel cost from *i* to *j* is known if the flow on each link is known. The average traffic volume at all iterations is given by;

$$v_a^n = \left(1 - {}^1\!/_n\right) v_a^{n-1} + {}^1\!/_n \, v_a \tag{4}$$

Where; v_a^n is the traffic volume on link a after n iterations and v_a is the traffic volume assigned on link a after the n'th iteration.

4.3 Parameter Estimation

In this section, we determine parameter α for a given trip cost exponent. For a given observed link *a*, the observed volume v_a^o is known and we use the model to predict linkflow on link *a* as, v_a^e for each value of α. The aim is therafter to minimize the difference between these two flows for all observed links.
The link volume for any link a predicted in the gravity model can be written as;

$$v_a = \alpha \sum_{ij} U_i R_j C_{ij}^{\beta} P_{ij}^{a} \tag{5}$$

Where P_{ij}^{a} refers to the proportion of traffic from *i* to *j* using link *a*. Considering x_a being an auxiliary variable where,

$$x_a = \sum_{ij} U_i R_j C_{ij}^{\beta} P_{ij}^{a} \tag{6}$$

This therefore means that $v_a = \alpha \, x_a$. With the statistical treatment of link flows, α is treated as the stochastic variable while x_a is the deterministic part which represents the assignment of traffic onto the network. This implies we retain the deterministic part during assignment iterations. Assuming the observed traffic flows are normally distributed as below;

$$v_a^o = v_a^n + \varepsilon_a \quad , \varepsilon_a \in N\left(0, (v_a^n)^p \sigma^2\right) \tag{7}$$

where p represents the weighting of the differences between the estimated and observed link flows. *P=0* means the traffic standard deviation is independent of of it's magnitude, *p=1* means the mean and variance are in proportion and *p=2* means the traffic standard deviation is proportional to the magnitude of the traffic. The parameters α and σ^2 are estimated by maximum likelihood method where the logarithm of the joint probability function is calculated and maximized by equating the partial derivatives with respect to α and σ^2 to zero. In this case, the value of α can be estimated as below.

$$\alpha = \frac{\sum_a (x_a^n)^{1-p}}{\sum_a (x_a^n)^{2-p}} \tag{8}$$

the weighted variance between the observed and predicted traffic can be estimated using the estimate $\acute{\alpha}$.

Steps to describe the algorithm are given below,
Step 1: Assign a fixed value of β
Step 2: calculate the link travel cost with no traffic on the network.

Step 3: Determine the shortest paths for all OD pairs and their respective path costs
Step 4: Using the above path costs, compute the number of trips between each OD pair using the gravity model without α and assign these trips onto the network to create estimated link flows.
Step 5: estimate the value of α so as to minimize the difference between the estimated and observed link-flows.
Step 6: Calculate the traffic volumes on all links by multiplyin the estimated flows in step 4 with the α value obtained in step 5.
Step 7: Calculate the new travel times on all links
Step 8: if the value of α stabilizes, stop otherwise go back to step 3 and repeat the procedure. Other values of β can as well be tested with the procedure.

This procedure is carried out with and without the inclusion of linkflows estimated from qualitative data obtained from the questionnaires for comparison.

5. SAMPLE TEST NETWORK

The methodology centred on in this study revolves around a hypothetical 3*3 directed road network as shown in the figure below with 'link 01,15,80' meaning link number from node 1 to node 2 is 01 and the reverse direction is link 15, with both links having a free flow travel cost of 80. The similar nomenclature applies to the rest of the links within the network and the overall algorithm used is written in matlab format.

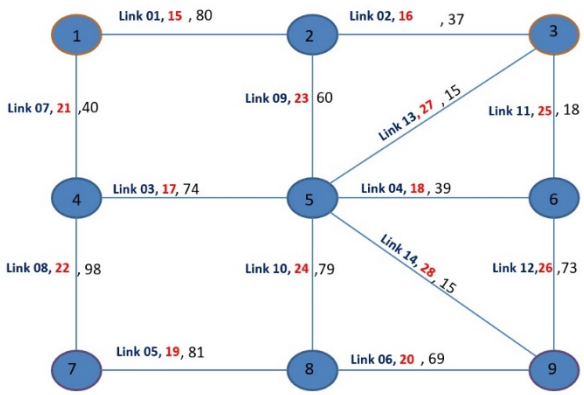

Figure 3 Hypothetical 3*3 network

The methodology followed within this study is divided into two stages.

Stage 1 (Scenario preparation). Assumed traffic demands from the hypothetical OD matrix are assigned onto the above network using the method of successive averages that considers link capacities and congestion effects. The final link flow volumes are obtained and recorded. A sample set of these link flow volumes will be used for comparison purposes in the following stage as known or observed link flows. The OD matrix assigned in this stage is T_0. In the application of this stage to Kampala data in the future, this stage will not be necessary.

Stage 2 (OD matrix estimation). Stage 2 involves the process of trying to retrieve the assumed demand in stage 1 and is as described below.
An OD matrix, T_1, is estimated following the procedure described in section 4.3 with a known set of linkflows. Secondly, with the inclusion of more linkflows resulting from the data collected through the questionnaire, a second OD matrix, T_2 is estimated. With T_0 taken to be the true matrix, relative mean square errors, E_1 and E_2 for matrices T_1 and T_2 are compared. The error E_2 is expected to be smaller than error E_1 and with the availability of more aggregated qualitative data describing more unknown network link characteristics, more link flow volumes can be postulated and supplemented onto the set of compared links. This therefore reflects the potential use of simple to collect qualitative information that improves on the accuracy of the estimated trip table.

The second portion of stage 2 of the described approach for this sample test network is shown in figure 1 with the highlighted box showing the resulting link flow estimates for some unobserved links. The highlighted red box shows this study's new contribution and the abbreviation "QD" within the figure means qualitative data postulated link flows.

This study will be limited to a case study of the hypothetical network and in further work, the general methodology described will be applied to Kampala, Uganda.

Upon this general idea, we propose further research into the feasibility and actual testing of the use of linguistic variables in estimation of unobserved linkflows and investigation into the selection criteria of these qualitative linguistic variables.

REFERENCES

1) Ennio Cascetta, Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator, *Transportation Research Board*, Vol. 18B, No. 4/5, pp. 289-299, 1984
2) J. Holm, T. Jensen, S.K. Nielsen: Calibrating traffic models on traffic census results only, *Traffic Engineering and control*, Vol. 17, No.4, pp 137-140, April 1976.
3) Sharminda Bera,. K. V. Krishna Rao,: Estimation of origin-destination matrix from traffic counts: the state of the art, *European Transport /Trasporti Europei*, no. 49, pp 3-23, 2011.
4) Josphat K. Z. Mwatelah, Methodological approach for

estimating O-D matrix and mode choice in developing countries with limited data, feasibility studies in Nairobi city 1994.
5) M. Zangiabadi, T. Rabie. : A new model for transportation problem with Qualitative Data, Iranian Journal of Operations Research, Vol. 3, No. 2, 2012, pp. 33-46
6) Dr. Charles Musselwhite, Qualitative methods in transport studies Senior Lecturer in Traffic and Transport Psychology, Centre for Transport & Society, University of the West of England, Bristol, UK.
7) Bruce Hellinga, Rajesh Gudapati,. Estimating link travel times from different data sources for use in atms and atis,. Department of Civil Engineering, University of Waterloo.
8) Exploring the capacity of social media data for modeling travel demand: a review of literature, Transport Reviews, A Transnational transdisciplinaru Journal,. Manuscript TTRV-2016-0004.