Australasian Transport Research Forum 2018 Proceedings
30 October – 1 November, Darwin, Australia
Publication website: http://www.atrf.info

"This is an abridged version of the paper presented at the conference. The full version is being submitted elsewhere. Details on the full paper can be obtained from the author(s)."

Levenshtein distance for the structural comparison of OD matrices

Krishna N. S. Behara[1], Ashish Bhaskar[1] Edward Chung[2]

[1]School of Civil Engineering and Built Environment, Queensland University of Technology, Brisbane, QLD 4000, Australia

[2]Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong.

Email for correspondence: krishnanikhilsumanth.behara@hdr.qut.edu.au

Abstract

The spatial distribution of Origin-Destination (OD) demands between different OD pairs reveals the structural information of OD matrices. Generally, OD pairs from geographical zones sharing similar activities, travel cost, and destination choices have a similar distribution of trips. Most of the traditional statistical measures are based on a cell by cell comparison and do not account for the additional structural knowledge in terms of similarity of trip distribution while comparing OD matrices. Thus, there is a need for new comparative measures to account for the structural information by computing statistics on the group of OD pairs. In this light, the paper adopts and extends an existing metric – Levenshtein distance for structural comparison of OD matrices. The proposed Mean Normalized Levenshtein distance for OD matrices comparison (MNLdOD) is an optimization-based metric and is computationally better than another popular metric – Wasserstein distance proposed by Ruiz de Villa et al. (2014).

Keywords: OD matrices comparison, Levenshtein distance, Wasserstein distance, OD matrix structure, statistical measures, Bluetooth OD.

1. Introduction

The distribution of travel demand between different origin-destination (OD) pairs is generally represented by an OD matrix. The “structure” of OD matrices is defined as the distribution of OD demands between different zonal pairs. The OD matrices have similar structure if groups of OD pairs share similar geographical zones, activities, travel cost and destination choices etc. resulting in similar travel patterns. Any changes in activity-travel patterns are reflected from the structural changes within OD matrix. The quality of OD matrices is generally assessed using traditional performance measures (RMSE, MAE etc.) that are generally based on cell by cell comparison of OD values and lack the potential to account for structure of OD matrices. By far in literature, very few statistical measures have focused on the structural knowledge of OD matrices. Bierlaire and Toint (1995 developed Matrix Estimation Using Structure Explicitly (MEUSE) approach to incorporate structural knowledge from parking surveys in traditional OD estimation. Kim et al. (2001) expressed the ratio of OD demand to origin flows as an important source of structural information that can be used as additional constraint in OD estimation process. Djukic et al (2013) applied Structural Similarity Index (SSIM) for comparing OD matrices and also proposed to use as an objective function in dynamic OD estimation. However, the SSIM approach is sensitive to the size of sliding window and is theoretical in nature. Ruiz de Villa et al. (2014) deployed Wasserstein Metric to compare OD matrices by taking into account network topology. To the best of authors' knowledge, this is the only method based on optimization formulation designed for structural comparison of OD matrices. Due to which it is computationally intensive for comparing OD matrices of large scale dimensions (further insights in section 5). In this light, the objectives of this study are to:

1. Develop a new measure –Mean Normalised Levenshtein distance for OD matrices structural comparison (*MNLdOD*); and
2. Compare the proposed *MNLdOD* with existing Wasserstein metric proposed by Ruiz de Villa et al. (2014).

2. Levenshtein Distance

The distribution of origin flows to different destinations provides insights into structural knowledge of travel patterns. For example, the preference of destinations could be different on different types of days. Say, the choice of destinations during Mondays are different as compared to that during a Sunday. This is due to different activities and their schedules during both days. Even if destination choices are same during both days, the number of trips could be different. This implies the structure of traffic flows are different if destination choices and number of trips are different between the same set of OD pairs.

Comparing OD matrices from this perspective requires a statistical metric that can exploit this additional structural information. For this purpose, we propose to extend traditional Levenshtein distance (details in section 4.1) into a new approach (presented in section 4.2) to suit its applicability for structural comparison of OD matrices.

2.1 Traditional Levenstein distance

Levenshtein distance, developed by Vladimir Levenshtein in (1966), is a measure of proximity between two strings, majorly applied to compare sequences in linguistics domain such as plagiarism detection and speech recognition, in molecular biology for comparing sequences of macro molecules, etc. It calculates least expensive set of *insertions*, *deletions* or *substitutions* that are required to transform one string into another. For example, if we have to compare two strings such as “MONDAY' and 'SATURDAY', one of the optimum ways is to insert the letters “S” and “A” and substitute “M”, “O” and “N” with “T”, “U” and “R” respectively leading towards a Generalized Levenshtein Distance (GLD) of 5.

Let's define S = $\{S_o, S_1, .. S_k \ldots S_s\}$ as the sequence of edit operations to transform string Y to X as represented by Y ⇒ X, and then the cost associated with each edit operation as $\{\beta_0, \beta_1 .. \beta_k .. \beta_s\}$. Generalized Levenshtein Distance (GLD) is the minimum total cost required to transform Y to X (*see* equation 1).

$$\text{GLD (X, Y)} = \min_{S}(\textstyle\sum_{k=0}^{k=s} \beta_k) \tag{1}$$

The Normalized Levenshtein distance (NLD) is the GLD normalized by the sum of the lengths of two strings (equation 2). This metric always lies between 0 and 1 (Yujian and Bo 2007).

$$\text{NLD (X, Y)} = \frac{\text{GLD (X,Y)}}{|X|+|Y|} \tag{2}$$

Refer, Heeringa (2004 for the pseudo code for computing GLD and NLD for two strings X and Y.

2.2 Levenshtein distance for OD matrices comparison

We extend the applicability of the technique to identify changes in preferences of destinations and number of trips made to different destinations from an origin. Here, for a given origin we can define a string that represents the order of destinations (in descending order of the demand to the destination from the origin). To compare two OD matrices, we compare the order of destination strings from each origins. However, contrary to the traditional application here different destinations have different demand values. We propose to include the demand in the estimation of Levenshtein distance. Hereon, the proposed approach for comparing origin row, n is termed as $LdOD_n$; the normalized comparison as $NLdOD_n$ and the mean comparison between OD matrices as *MLdOD* and its normalized version as *MNLdOD* respectively.

Here, the cost (in terms of trips) associated with each edit operation are $\beta_0, \beta_1 .. \beta_k .. \beta_s$ and the $LdOD_n$ is expressed as shown in equation 3. If the comparison is required between a scale of 0 and 1, one can use Normalized $LdOD_n$ ($NdLOD_n$) values as shown in equation 4. Here, $NdLOD_n$ is obtained by normalizing over the sum of origin flows from both matrices. The formulation for *MNLdOD* is shown in equation 5.

$$LdOD_n\ (\boldsymbol{R}_X^n, \boldsymbol{R}_Y^n) = \min_{S}(\textstyle\sum_{k=0}^{k=s} \beta_k) \qquad (3)$$

$$NdLOD_n\ (\boldsymbol{R}_X^n, \boldsymbol{R}_Y^n) = \frac{LdOD_n\ (\boldsymbol{R}_X^n, \boldsymbol{R}_Y^n)}{(\sum_{i=1}^{i=M} A_{x_i}^n + \sum_{i=1}^{i=M} A_{y_i}^n)} \qquad (4)$$

$$MNdLOD(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{n=1}^{n=N} NdLODn}{N} \qquad (5)$$

Where, for OD matrix **Y (**of size say, N x M**),** the sorted set of destination IDs and the corresponding demand from an origin n is expressed as $\boldsymbol{R}_Y^n = (\ \boldsymbol{D}_Y^n, \boldsymbol{A}_Y^n\ ) = \{(D_{y_1}^n, A_{y_1}^n), .. (D_{y_i}^n, A_{y_i}^n) \ldots (D_{y_M}^n, A_{y_M}^n)\}$ where D_{y_i} and A_{y_i} are the ith preferred destination and its corresponding demand value, respectively. Similarly, we express $\boldsymbol{R}_X^n = (\boldsymbol{D}_X^n, \boldsymbol{A}_X^n)$ for matrix **X**.

Unlike GLD, *LdOD* does not have any substitution operation because the destinations in the two OD matrices are same, however their order varies. Since, the *LdOD* formulation considers OD demand besides the sequence of destination IDs, we propose edit operation – "*absolute trips-difference*" that accounts for the demand variations over different destinations. This operation is in addition to the *insertion* and *deletion* operations.

3. Wasserstein vs Levenshtein distances

Levenshtein and Wasserstein metrics compare OD matrices through an optimization formulation to find minimal difference in the distribution of OD demand over the network. Despite this similarity, they are different from each other as discussed below.

Firstly, *MLdOD* computes the structural differences between OD matrices in terms of OD flows. On the other hand, Wasserstein metric expresses in terms of travel cost.

Secondly, Wasserstein metric is computationally very expensive as compared to *MNLdOD*. This is because solution search space for Wasserstein metric is spread over the entire OD matrix i.e. travel cost for all combinations of OD pairs need to be checked for an optimum distance. Whereas, the $LdOD_n$ is computed locally for each row due to which the solution search space is constrained to OD pairs originating from a specific origin only. To compare the computational strength of two metrics, Bluetooth OD matrices from Monday and Sunday are compared against each other. The OD matrices are constructed from Bluetooth observations for Brisbane City Council (BCC) region. The test is conducted on a Dell computer with Intel(R) Core(TM) i7-4770 CPU, 16GB RAM (3.40GHz) and time taken for computation is 0.33 seconds for *MNLdOD* and 1690 seconds for Wasserstein approach respectively.

4. Conclusion

The study focusses on the need for statistical measures that can perform holistic structural comparison of OD matrices. It opens with the review of indicators that are specifically meant for structural comparison and then discusses the development of a new approach- Mean Normalized Levenshtein distance for OD matrices (*MNLdOD*). This novel concept is borrowed from traditional Levenshtein distance that is popularly used for comparison of strings in linguistics field. The findings of this study are two-fold:

Firstly, *MNLdOD* is designed to capture the order of destination choices along with the distribution of trips while comparing OD matrices. The difference in the distribution among the group of OD pairs originating from the same origin zone are compared to identify the structural differences. For example, higher the similarity in the order of destinations and volume of OD demands, lower is the *MNLdOD* value between OD matrices.

Secondly, it is computationally more effective than Wasserstein metric for structural comparison of large scale OD matrices that are not sparse.

5. Acknowledgement

The Bluetooth data used in this study is provided by Brisbane City Council (BCC) and Queensland University of Technology. The conclusions of this paper reflect the understandings of the authors, who are responsible for the accuracy of the data.

6. References

Bierlaire, Michel and Ph L Toint. 1995. "Meuse: An origin-destination matrix estimator that exploits structure." *Transportation Research Part B: Methodological* 29 (1): 47-60.

Djukic, Tamara, Serge Hoogendoorn and Hans Van Lint. 2013. "Reliability assessment of dynamic OD estimation methods based on structural similarity index." In *Transportation Research Board 92nd Annual Meeting,* edited.

Heeringa, Wilbert Jan. 2004. "Measuring dialect pronunciation differences using Levenshtein distance, Citeseer.

Kim, Hyunmyung, Seungkirl Baek and Yongtaek Lim. 2001. "Origin-destination matrices estimated with a genetic algorithm from link traffic counts." *Transportation Research Record: Journal of the Transportation Research Board* (1771): 156-163.

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." In *Soviet physics doklady,* edited, 707-710.

Ruiz de Villa, Aleix, Jordi Casas and Martijn Breen. 2014. "OD matrix structural similarity: Wasserstein metric." In *Transportation Research Board 93rd Annual Meeting*, edited.

Yujian, Li and Liu Bo. 2007. "A normalized Levenshtein distance metric." *IEEE transactions on pattern analysis and machine intelligence* 29 (6): 1091-1095.