# Mobility Mining Using Nonnegative Tensor Factorization

**Conference Paper** · August 2017

DOI: 10.1007/978-3-319-65340-2_27



**3 authors:**

Hamid Eslami Nosratabadi
University of Porto
**18** PUBLICATIONS **65** CITATIONS
SEE PROFILE

Hadi Fanaee-T
University of Oslo
**20** PUBLICATIONS **119** CITATIONS
SEE PROFILE

João Gama
University of Porto
**360** PUBLICATIONS **6,131** CITATIONS
SEE PROFILE

# Mobility Mining using Nonnegative Tensor Factorization

Hamid Eslami Nosratabadi[1], Hadi Fanaee-T[2], Joao Gama[3]
[1] LIAAD-INESC TEC Rua Dr. Roberto Frias 4200 - 465 Porto, hamid.e.nosratabadi@fc.up.pt
[2] LIAAD-INESC TEC Rua Dr. Roberto Frias 4200 - 465 Porto, hadi.fanaee@fe.up.pt
[3] LIAAD-INESC TEC Rua Dr. Roberto Frias 4200 - 465 Porto, jgama@fep.up.pt

**Abstract.** Mobility mining has lots of applications in urban planning and transportation systems. In particular, extracting mobility patterns enables service providers to have a global insight about the mobility behaviors which consequently leads to providing better services to the citizens. In the recent years several data mining techniques have been presented to tackle this problem. These methods usually are either spatial extension of temporal methods or temporal extension of spatial methods. However, still a framework that can keep the natural structure of mobility data has not been considered. Non-negative tensor factorizations (NNTF) have shown great applications in topic modelling and pattern recognition. However, unfortunately their usefulness in mobility mining is less explored. In this paper we propose a new mobility pattern mining framework based on a recent non-negative tensor model called BetaNTF. We also present a new approach based on interpretability concept for determination of number of components in the tensor rank selection process. We later demonstrate some meaningful mobility patterns extracted with the proposed method from bike sharing network mobility data in Boston, USA.

**Keywords:** Mobility Mining, Nonnegative Tensor Factorization, BetaNTF.

## 1 Introduction

Extracting mobility patterns has been recently studied in the context of spatial data mining. It has lots of applications in urban planning, scheduling and public transportation. In the recent decade several data mining techniques have been exploited for addressing this problem. For instance, [5] used Markov models to tackle the problem of predicting next locations. In [20] the authors exploited association rules to extract patterns for tourist attraction problem. In [21] a heuristic method is proposed based on data mining which consider the trajectory of a focal tourist and the movements of past visitors. However, in neither of these works the spatiotemporal structure of traffic data is considered simultaneously.

Tensor decompositions are one of models that can naturally capture and model the spatiotemporal variance of traffic data. They are recently applied for solving many problems in relevant areas such as traffic flow prediction [2], data compression of urban traffic data [1], clustering and prediction of temporal evolution of global urban network [6], traffic speed data imputation [10], estimation of missing traffic volume [11, 14, 15, 16, 17, 18], and traffic volume data outlier recovery [19]. However, to the best of our

knowledge, Non-negative Tensor Factorization (NNTF) has never been applied to the mobility pattern extraction problem. This is while NNTF has been successfully applied to problems related to topic modelling [3] for extracting topic models from the text corpus. Our main objective in this work is to extend the application of NNTF from topic modelling to extract mobility patterns from dynamic traffic data.

Our initial empirical evaluation with traditional NNTF models such as CP-NLS [12, 13] against the recent method, BetaNTF [4] indicates the better performance of BetaNTF. The BetaNTF algorithm first time was developed in signal processing for blind source separation.

Our main objective in this work is to extract interesting, meaningful mobility patterns from bike sharing network data using BetaNTF model. The data being generated in bike sharing networks naturally has a tensor structure of "Origin x Destination x Time". That is why it is quite relevant to be analyzed with tensor decomposition models. However, one of the important problems in applying tensor decomposition models is how to determine the number of components. This becomes more difficult in pattern extraction since not only the model should be accurate but also it should be interpretable. To solve this problem, for the first time we introduce a new mechanism based on the interpretability of patterns for determining number of components. To summarize, our contributions include:

- For the first time we extend the ideas in topic modelling based on non-negative tensor factorization to the problem of mobility pattern mining.
- We apply BetaNTF a recent non-negative tensor decomposition algorithm (based on CP structure) for extracting patterns.
- We propose a new approach based on interpretability for determining of number of components in the BetaNTF model.
- We evaluate our proposal on a real-world bike sharing data set and provide realistic evidences regarding the validity of extracted patterns.

The rest of the paper is organized as follows. Details of the proposed method is presented in Section 2. Section 3 describes the experimental setup, data set and the empirical results. Section 4 gives the results. The last section concludes the exposition presenting the final remarks.

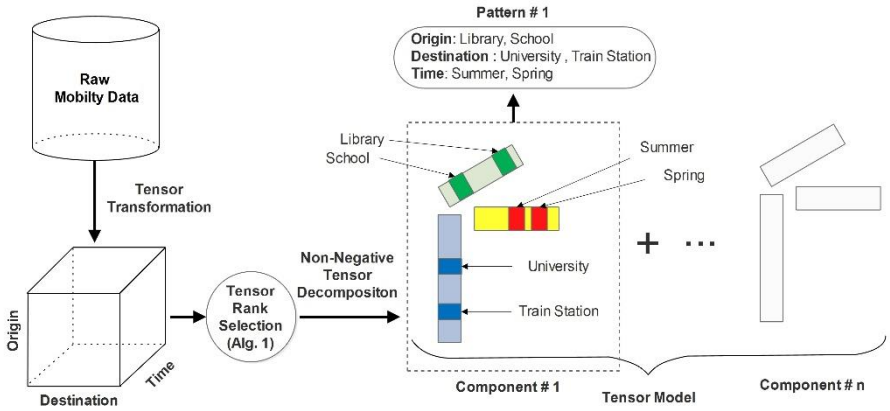


**Fig. 1.** Cartoon of the proposed method

## 2 Methodology

The overall picture of our methodology is illustrated in Figure 1. In the following section each of these components will be described in more details.

### 2.1 Tensor Transformation

The raw mobility data normally is presented in the format of transactional database. Each row usually contains information regarding the origin and destination of travel and also the timestamp when the travel is started and ended. A pre-processing step is required to transform this kind of databases to tensor format. A list of distinct stations and days is first retrieved and then we count number of performed travels between origin/destination stations during various day intervals. This can be carried out via group queries on the database. For instance, a query like "Count number of travels from station S#1 to station S#2 in day#3" constitutes a triplet of X (1,2,3)=15 in the third-order tensor. 15 in the example is one of the elements of the tensor and is the number of travels from station 1 to station 2 in day 3. Given the count for all possible triplets we can generate the full tensor of "Origin x Destination x Time" (ODT) which will be later used in analysis step.

### 2.2 Nonnegative Tensor Factorization

Non-negative tensor decompositions are considered more suitable than regular decomposition models for analysis of visual and count data (which is the case here). The reason is that in the non-negative models the elements in factor matrices have the non-negativity restriction which is more interpretable. In the case of mobility data for instance, negative values in the factor matrices cannot be justified with existing physical reality. Because, we cannot find any negative number of travels between the stations. In this step we apply BetaNTF decomposition on the ODT tensor and retrieve the factor

matrices. Note that BetaNTF is from the CP/PARAFAC family, therefore the decomposed space will include three factor matrices respectively for “origin”, “destination” and “time” dimensions. BetaNTF instead of using Alternating Least Squares technique [9] which is used in the majority of algorithms fits the tensor model by using beta-divergence as the cost-functions.

### 2.3 Rank Selection based on Interpretability

One of the difficult problem in tensor decomposition or in general in latent models is how to choose the number of components (or latent variables). This situation gets worse in the application of mobility pattern mining where the quality of extracted patterns is directly related to the chosen number of components. The common techniques for determining number of components are automatic methods such as triangle technique [8] (i.e. ranktest in Tensorlab) which are mostly used in the literature. But the problem is that this type of techniques only considers the trade-off between accuracy and simplicity. In the mobility pattern mining another factor gets importance which is interpretability. In order to solve this issue, we propose a new rank selection methodology which selects the tensor rank by considering the trade-off between the number of distinct extracted patterns and the model simplicity. Our proposed algorithm for rank selection is demonstrated in Algorithm 1. We first apply BetaNTF with R number of components varying from 1 to Max R on input X tensor. Next, we extract patterns using methodology described in section 2.4. Then we generate a table (See Table 1 for example) including a list of distinct number of discovered areas (c1) and number of patterns with different origin and destination zones (c2) and maximize c1 and c2 while minimize R. The logic behind this method that R is chosen as suitable number of components

when its corresponding model covers more various distinct patterns while keeps the model as simple as possible.

**Table 1.** Dispersion of the obtained pattern from R=1 to R=15

| R | The number Of encompassed area | The number of distinct area |
|---|---|---|
| **1** | 2 | 0 |
| **2** | 2 | 1 |
| **3** | 3 | 2 |
| **4** | 3 | 4 |
| **5** | 5 | 5 |
| **6** | 4 | 5 |
| **7** | 4 | 6 |
| **8** | 3 | 4 |
| **9** | 4 | 4 |
| **10** | 5 | 8 |
| **11** | 6 | 11 |
| **12** | 3 | 7 |
| **13** | 2 | 12 |
| **14** | 3 | 9 |

### 2.4 Pattern Extraction

There exists almost no work in the literature that provides a solution for automatic extraction of patterns (or topic models) from the decomposed tensor space. Usually the patterns are extracted by visual inspection of components. In this paper for the first time we propose an automatic mechanism for extraction of patterns from the factor matrices obtained from the decomposition model. Our proposed approach is as follows. Decomposition of ODT tensor (let's say with size of $N \times M \times K$) with R number of components gives us three factor matrices of size $N \times R$ (origin dimension), $M \times R$ (destination dimension) and $K \times R$ (time dimension). The set including the first column of $N \times R$ and $M \times R$ and $K \times R$ matrices constitutes the first pattern (see Fig. 1). Likewise, the second pattern can be built by the second columns of these matrices. Now we only need to find the elements with highest weights in these factor matrices. We tested three strategies for doing this, with z-score, with top N items and finally top N percentage. It seems that top N percentage gives a more interpretable results. Besides, choosing sigma threshold for z-score method was a bit difficult when there is a big difference between sizes of dimensions. So we select the top N% of items in the first column of factor matrices corresponding to each dimension and then generate a triplet of indices related to that weights. For example, suppose that the corresponding weight for *Central Station* in the origin dimension is selected as Top 1 and weight for *City park* is the highest weight in factor matrix of destination dimension. Also suppose that the weight for 2013-09-23 is the maximum weight among all in the first column of "time" factor matrix. A

triplet like {O: ”Central Station”, D:” City Park”, T: “2013-09-23”} would be outputted as the first extracted pattern.
We also relate the extracted temporal components to days of the week, holiday, month, season, and so forth to find the temporal dimension of patterns.

Algorithm (1) Tensor rank selection based on interpretability

```
Input: X (Origin × Destination × Time tensor), Max R
Output: determining of the best R rank
1 For R=1 to Max R
2 Apply BetaNTF on X given R
3 Extract patterns using the methodology described in
section 2.4
4 c1 ← number of distinct zones
5 c2 ← number of distinct areas with different origin
and destinations
6 End
7 Selected R ← Maximize c1,c2 and Minimize R
```

## 3 Experimental Evaluation

In this section we begin by describing the dataset and then explain the configuration used for experiments. Afterwards we demonstrate the obtained results.

### 3.1 Dataset

Boston bike-sharing data set has been extracted from hub-way data challenge 2013 [7]. It includes a historical usage log of all transactions in the network from 2011-07-28 to 2012-10-01, exclusive to the system's off-days in the winter, a total of 327 days. There are also 95 stations in total. After creating adjacency matrices for each day, the generated ODT traffic tensor will be in size 95 ×95 × 327.

### 3.2 Experimental Setting

These configurations are used in the experiments. Max R=15 is chosen in the Algorithm 1. The Selected R is chosen as 11 after generating Table 1 by taking into account the trade-off between simplicity of model and maximization of number of distinctive patterns and areas. Number of iterations in BetaNTF algorithm is set as 70. We also set Itakura-Saito cost function [22] in the BetaNTF algorithm. The N in the Top N% weight selection is set as 3 based on trial and error.

### 3.3 Results:

In this section we will demonstrate the mobility patterns extracted from the Boston bike sharing dataset (Fig 2 to Fig 12). In all figures, the λ value according to each component are shown. λ is obtained after normalization of decomposed tensor and has similar meaning as eigenvalue in matrix factorization. The component with higher lambda is more important. In our experiment the first and last λ are obtained respectively as 289 and 36.

Among the discovered patterns, Boston South Station and Boston North station are the main hotspots among than the others. Boston North station is surrounded by TD garden (multi-purpose arena) which seems to be related to sport and entertainment events. Boston south station also seems to be the transit hub as is surrounded by many transit stations. In terms of the time dimension, Tuesday and Wednesday in summer, especially in the month of August have appeared more frequently and probably play an important role in creating more diverse patterns.

In all figures, the red label marker specifies the point of destination, the green label marker displays the origin and the gray one demonstrates those points that origin and destination are overlapped. In the following we describe each extracted pattern in more details.

**Pattern#1**: In the first pattern (Fig.2) we observe two origin and two destination areas that demonstrate a mobility flow from two main stations in Aquarium and Arlington to Boston North Station and Boston South Station, one of two main transit hubs. This can be related to sport and entertainment events where people tend to use more bikes to transit from point of interests such as TD Garden and Boston Common. This temporal component reveals that this pattern is more frequent on Tuesdays and Wednesdays (working days) in the summer (months of July, August and September).

**Pattern#2:** In this pattern which is shown in Fig. 3, we can see a mobility link between North End Area which contains variety of tourist attractions to Boston South Station. This pattern also temporally occurs Saturdays and Sundays (weekend) in the seasons of spring and summer, in particular months of May, June, July and August. This pattern probably is related to weekend trips to point of interests and restaurants.

**Pattern#3:** Interesting mobility flow can be observed in this pattern (Fig.4) between two transit hubs of Boston South Station and Boston North station. The peak in the temporal component is related to month of August, so probably it uncovers a transit pattern of tourists who move between these two stations.

**Pattern#4:** This pattern (Fig. 5) probably demonstrates the mobility behavior of youth in Harvard medical school and Boston Sport club which also might be the central point for bikers. Temporally this pattern is more seen on Mondays and Wednesdays (working days) in the spring and summer, in particular months of July and August.

**Pattern#5:** The fifth pattern corresponds to a mobility flow originated in Downtown Crossing and Boston south Station approaching TD Garden and North End (Fig. 6) during Mondays and Wednesdays (working days) in the summer, especially on August and September.

**Pattern#6:** This pattern demonstrates the bi-directional mobility in the area close to Star Market and Portsmouth Playground (Fig.7) which temporally occurs on Wednesdays and Sundays in holidays in the working days in spring and summer, especially in April and July.

**Pattern#7:** The spatial and temporal component of this pattern seems to be related to the mobility of students of MIT campus during the working days, especially in period of school opening in September. It seems students arriving from the north and south rent/leave the bikes at three close stations nearby the MIT campus (Fig.8).

**Pattern#8:** This pattern (Fig. 9) seems to be related to shopping mobility between W Newton St. where there is a shopping center nearby and Dartmouth St which is more frequent on beginning of the week (Monday and Tuesday) in the Summer, especially in months of June and July.

**Pattern#9:** This pattern (Fig. 10) reveals a mobility flow from Downtown Crossing area to two origins at TD Garden (Boston North Station) and North End which is more frequent on Mondays and Thursdays (Working days) in the summer and autumn, especially in August and September.

**Pattern#10:** The stations appeared in this pattern are nearby stations to Boston University (Fig. 11). Apparently this reflects the mobility of Boston University's students. The temporal component also shows that this is a frequent pattern during the Tuesdays (working days) in the months of August and September.

**Pattern#11:** This pattern (Fig 12) includes two destinations close to Massachusetts college of Pharmacy and Boston children's Hospital and two origins close to Boston Public Librar. It seems there is a link between these points during Wednesdays and Thursdays (working days) in the spring and summer, especially in June and July.

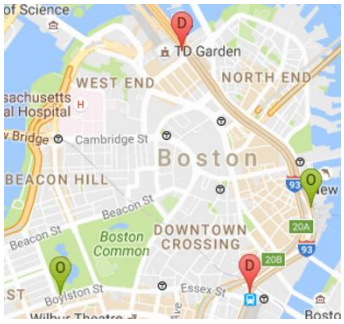

Fig2.Pattern#1, λ=289.68

Fig3. Pattern#2, λ =272.68

Fig4. Pattern#3, λ =186.95

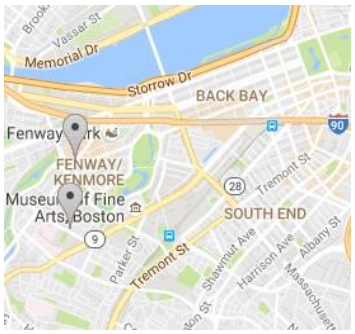


Fig5. Pattern#4, λ=180.64

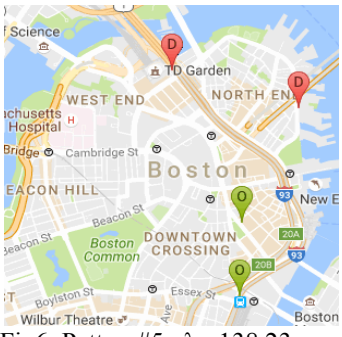


Fig6. Pattern#5, λ=138.23

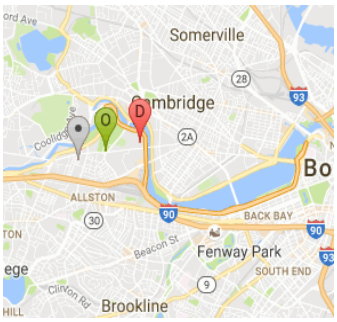


Fig7. Pattern#6, λ=136.42

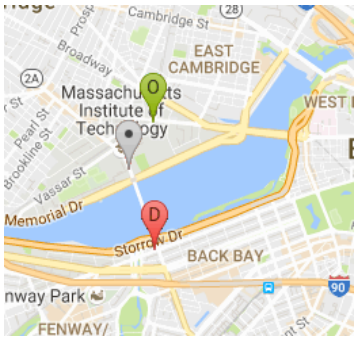


Fig8. Pattern#7, λ=129.10

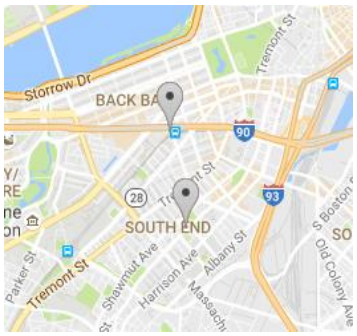


Fig9. Pattern#8, λ=125.24

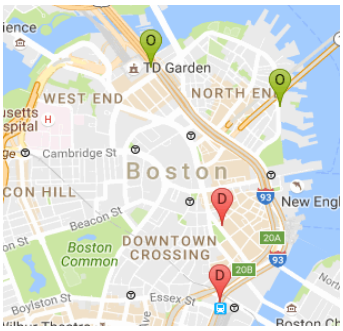


Fig10. Pattern#9, λ=95.93

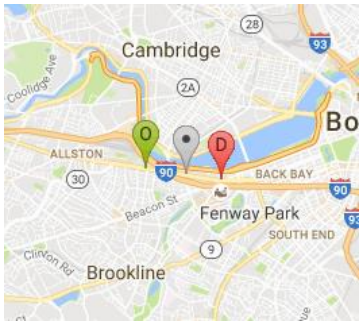


Fig11. Pattern#10, λ=75.33

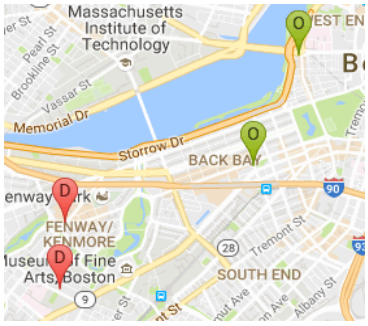


Fig12. Pattern#11, λ=36.02

## 4 Conclusion

We extend the application of Non-negative tensor factorization from topic modelling and pattern recognition to mobility pattern mining. In particular, we demonstrate that

the recent technique, BetaNTF which was originally developed for bling source separation has quite good potential for mobility pattern mining. We for the first time present a new technique for choosing number of components based on the provided interpretability. By applying our method on the real-world mobility data of Boston bike sharing network we provide some evidences of mobility behaviors that justify the usefulness of the proposed methodology. Some of the patterns such as mobility patterns of students close to the university campuses, or mobility close to shopping centers or tourist mobility makes sense and confirms the validity of patterns.

## Acknowledgements

This research was carried out in the framework of the project" TEC4Growth – RL SMILES –Smart, mobile, Intelligent and Large Scale Sensing and analytics NORTE-01-0145-FEDER-000020' which is financed by the north Portugal regional operational program (NORTE 2020), under the Portugal 2020 partnership agreement, and through the European regional development fund. The authors thank Antoine Liutkus for providing the code for BetaNTF and Huway Company for providing the dataset.

## References

1. Asif, M. et al.:Data compression techniques for urban traffic data.. Computational Intelligence in Vehicles and Transportation Systems (CIVTS), (2013) IEEE Symposium on.
2. Abadi, A., Tooraj, R., Petros A. Ioannou.: Traffic flow prediction for road transportation networks with limited traffic data. IEEE Transactions on Intelligent Transportation Systems 16.2 (2015): 653-662.
3. Bader, B.W., Michael, W. Berry, Murray, B.: Discussion tracking in Enron email using PARAFAC." Survey of Text Mining II. Springer London, (2008). 147-163.
4. Cichocki, A., et al.: Non-negative tensor factorization using alpha and beta divergences. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 3. IEEE, (2007).
5. Chen, M., Xiaohui, Y., Yang, L.: Mining moving patterns for predicting next location. Information Systems 54 (2015): 156-168.
6. Han, Y., Fabien M.: Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. International Journal of Intelligent Transportation Systems Research 14.1 (2016): 36-49.
7. http://hubwaydatachallenge.org/
8. J.L. Castellanos, S. Gomez, V. Guerra.: The triangle method for finding the corner of the L-curve, Applied Numerical Mathematics, Vol. 43, No. 4, (2002), pp. 359-373.
9. J. D. Carroll and J. J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition, Psychometrika,35 (1970), pp. 283–319.
10. Ran, Bin, et al.: Traffic speed data imputation method based on tensor completion. Computational intelligence and neuroscience 2015 (2015): 22.

11. Ran, Bin, et al.: Estimating missing traffic volume using low multilinear rank tensor completion. Journal of Intelligent Transportation Systems 20.2 (2016): 152-161.
12. Sorber, L., Marc, V B., Lieven, D, L.: Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(L_r,L_r,1) terms, and a new generalization. SIAM Journal on Optimization 23.2 (2013): 695-720.
13. Sorber, L., Marc V, B., Lieven D, L.: Unconstrained optimization of real functions in complex variables. SIAM Journal on Optimization 22.3 (2012): 879-898.
14. Tan, H., et al.: A tensor-based method for missing traffic data completion. Transportation Research Part C: Emerging Technologies 28 (2013): 15-27.
15. Tan, H., et al.: Low multilinear rank approximation of tensors and application in missing traffic data. Advances in Mechanical Engineering 6 (2014): 157597.
16. Tan, H., et al.: Traffic volume data outlier recovery via tensor model. Mathematical Problems in Engineering 2013 (2013).
17. Tan, H,.: Traffic Missing Data Completion With Spatial-temporal Correlations. Diss. Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 2014.
18. Tan, H., et al.: A new traffic prediction method based on dynamic tensor completion. Procedia-Social and Behavioral Sciences 96 (2013): 2431-2442.
19. Tan, H., et al.: Correlation analysis for tensor-based traffic data imputation method. Procedia-Social and Behavioral Sciences 96 (2013): 2611-2620.
20. Versichele, M., et al.: Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. Tourism Management 44 (2014): 67-81.
21. Zheng, W., Xiaoting, H., Yuan, Li,.: Understanding the tourist mobility using GPS: Where is the next place?. Tourism Management 59 (2017): 267-280.
22. Itakura, F., & Saito, S.: Analysis synthesis telephony based on the maximum likelihood method. In Proc. 6th of the International Congress on Acoustics (pp. C–17–C–20). Los Alamitos, (1968). IEEE.