Proceedings of the 9th International Conference on Applied Informatics
Eger, Hungary, January 29–February 1, 2014. Vol. 1. pp. 151–159
doi: 10.14794/ICAI.9.2014.1.151

Estimation of OD matrices with naive Bayes method

Árpád Nyilas, Zoltán Ács, Zoltán Vincellér

Eötvös Loránd University
anyilas@caesar.elte.hu
acszolta@inf.elte.hu
vzoli@inf.elte.hu

Abstract

The Origin/Destination (OD) matrix is an essential input for all traffic simulations. This matrix determines how many vehicles travelled from point X to point Y, where X and Y can be either a whole city or some kind of traffic zone. In practice, experts usually use estimations for getting the elements of these matrices, but the cost of these measurements can be very high. For mathematical estimations, we need a large amount of historical information and statistical data about the field of our interest. As example for our concrete case, it is nice to have information over the population, the number of cars or the number of trucks in the area. Sometimes the granularity of the sufficient data can go down to street level. So, the pure mathematical models are usually unusable, and the experts have to create the OD matrices manually with the use of their additional practical experiments. As demonstration of the size of this task, the national level OD matrix of Hungary contains more than 1 000 000 values.

In our proposed solution, we generate an OD matrix and we adjoin a probability to each value, which shows the estimated probability of accuracy for the actual value. If it is less than $P_{\min}$ percent, the experts can modify it. We accept this modification as reliable data. We add it to the training set with high weight. Therefore, in the next execution the probability attached to a value will be high in similar situations and the experts don't have to override result of our algorithm. So experts can calibrate the model by this feedback.

We use the additive naive Bayesian method to calculate values of OD matrix [2], the matrix contains also probabilities. The input of the algorithm is statistical data about demography, economy, vehicles and accidents; and the OD matrix of previous years.

Summing up, we have worked out a new method to generate an OD matrix that requires only minimal collaboration with experts. An expert has

opportunity to calibrate and override the model. Currently, the method is used as part of complex traffic simulation pilot software.

Keywords: traffic simulation, OD matrix, Bayes estimator

1. Introduction

1.1. Traffic simulation

Traffic simulation is very important part of road network analysis and planning. Simulation can show effect of a new road or new roundabout. For this it can useful if we want to design a new bridge. Experts can make a good decision about place of bridge. Other example of simulation usage is the traffic light setting. We can see simulation should be realistic, it have to model the driver behaviors or else the result will be useless.

First let's see a typical process of traffic simulation. First step is defining zones. Zone is base areas of simulation. Next step is to calculate the OD matrix. This matrix determines how many vehicles travelled from point X to point Y, where X and Y can be either a whole city or some kind of traffic zone. Of course, in practice experts use zones, because it's flexible. Zone can be part of the city or group of towns or region.

If we have OD matrix, we should generate traffic to road network by OD matrix. It means we should generate trip for all vehicle. So we assign the source and destination point and concrete route. When we assign the end point of trips, we should calculate capacity of roads. So, we shouldn't put all vehicle to center of zone, because there will traffic jam and traffic through zones will be unrealistic.

Next step is the execution of simulation. After it expert evaluate result of simulation. Typical, first time the result is unrealistic. So experts have to modify routing or trips. In practice OD matrix modification is rare, because it has too much cost. However sometimes it's necessary.

The simulation output depends on simulation software. Typically, they are vehicle tracking or road sectional data.

1.2. Input of simulation

Quality of simulation depends on quality of input data. In this section we show available inputs. We can use demographical, economical and traffic data. The first two is public data in statistical office webpage. This data is on town level. Of course, few towns have some detailed statistic, but they are rare in Hungary. Moreover we can get accident data from statistic office.

We can see, the available statistical data are incomplete for estimate direction and rout of traffic, but we can get a picture of driver behaviors. We can use vehicle fleet tracking data, like FCD. This information is useful to detect the preferred routs or to get information about avenge speed on a road. So, it informs us about traffic dynamic. More typical traffic measurement is the traffic counting. Contrary

to FCD it is cross-section information, so we do not get information about driver's behaviors, source and destination.

In simulation we can calculate some additional presumption. For example, we can simulate weather, or other external factors. We can configure the simulation models for this, but in this case we cannot separate data in different circumstances. Then we should use more complex data mining method.

In our article we concentrate to traffic source and destination. We will present some existing method for estimate of this. And we show our method, which mixed statistical and traffic silence tools for mine origin, destination information from above data.

2. Traffic simulation

2.1. Motivation

The traffic simulation is very popular tool in transportation science. It's powerful for examine effect of new road or roundabout. In practice experts use simulation to configure the traffic lights. It's also important in public transport design, for example in planning of time tables.

We can see the simulation has many application areas, but each has similar procedure. Therefore we can examine only one application of simulations problems.

2.2. General procedure

Many traffic simulation software exist, each can run traffic simulation. Each software use mathematical model. We distinguish micro and macro by type of background model. So, running simulation has well tested methodology. The real problematic areas are the generate input and interpret the result.

In this article we examine only the input generation. So the main problem we don't have enough data about people behaviors and population distribution. If we would like to do realistic simulation, we have to calculate the population distribution. For example we want to simulate the working time, and traveling between homes and working places. But we don't have information each people's addresses and working places. For this, we should estimate this and similar data.

In Figure 1, we can see the data flow schematic diagram of simulation. In this figure the first operation is zone definition. This is the basic area of simulation. It means, we define amount of traffic between this areas. But first of all, we should define zones. We use OD matrix for store the amount of traffic between zones. OD matrix usually includes daily information, but in practice we use few second step interval for simulation. Moreover OD matrix does not include the routs. So we should assign routs for each vehicles or vehicle groups (in case of macro simulation). In simulation step we apply one of mathematical simulation model for running simulation. Evaluate the result of simulation. After the simulation we

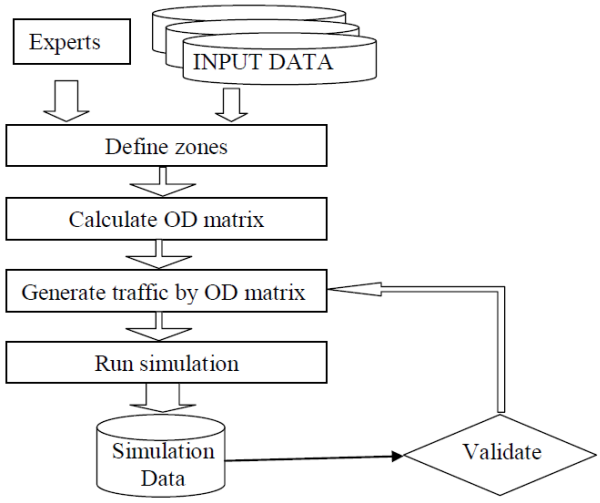

Figure 1: Simulation process

have lot of data, but we have to interpret this, because we want to model the real traffic, so we have to find the reason of traffic flow behaviors.

The routing and running simulation has supported by simulation software [5]. The zone definition is also difficult task, but in practice the natural zones usually are useable, for example towns, districts, housing estates. In this paper we focus to OD matrix estimation.

3. Exist procedures of OD matrix estimates

3.1. Full manually

The most scalable is the expert-oriented manual determination. It doesn't use integrated mathematical model. It doesn't use exact algorithm. The expert simply decides how he will calculate in each cases. Obviously, it is a slow and expensive procedure. We can see this independent on expert's knowledge. This is useful in small area, where we want to calculate by all factor.

3.2. Gravity model

The gravity model illustrates the macroscopic relationships between zones. In analogy with physics, we can say law of demography gravitation: The distance decay factor of 1/distance has been updated to a more complex function. In practice we need too much information to approximate function of generalized cost.

The gravity model is widely used model; it's based on physically gravity. The base equal is also similar, however there we use other constants, and it has many

parameter. The oldest application is the economical usage, we can't thing this science is similar to traffic engineering. Each area model the human behavior, but we can find significant differences, the sales and purchase is measurable in easy way. However traffic counting is very expensive procedure. Moreover we can't get enough precise information about workplaces, and homes.

Base equal of physically gravity model is the following:

$$F_{i,j} = G\frac{M_i M_j}{D_{i,j}^2}$$

where M_i is weight of i. zone, $D_{i,j}$ is the distance between i. and j. zone, G is the gravitation Coefficient and F is the force between zone i. and zone j. If we want to use this for demographical gravity we should use complex function instead of M. In scientific literature we can find some recommendations for this function.

Sum up with, the gravity model is powerful tool, if we have complete and precise information about zone's population and economy.

3.3. Observed base

If we would like to estimate OD matrix of current or previous period, we can observe the size of traffic flow. In road we can't do that, because the traffic counting don't inform us about source and destination vehicles. However we may find some area, where we can observe the traffic source and destination. The public transport is typical example of it, because people buy tickets. So number of sold tickets is good estimation of size of traffic flow.

In road we can use it if we have vehicle *tracking* data, such as FCD. Nowadays it's available only size. For this, in some area we can use observing for validation of on road OD matrixes.

4. Our hybrid estimation method

In our idea, we mix the Bayesian method and Gravity model. Our goal is fill the data gaps for gravity model. We get idea from computer networks, where Bayesian method is popular process for OD matrix estimation.

4.1. Bayesian method

The Bayesian method is classifier which is based on mathematically probability model. So we use this for classify zones. Each class includes zones with similar traffic data.

First let see what kind of input data available for this classifier? We can get demographical and economic data in town level from statistical office. So, first step is fitting this data to zones. So our assumption, demographical and economical attributes and amount of traffic is not independent. In country where people can choose home and workplaces is true.

We need trainer set to build probability model for Bayesian method. So, we need many zones where we know the size of input and output flow. It can be historical data, if we have same kind of historical demographical and economic data for those zones. Moreover the trainer set should include enough data to model building, and enough data for all classes. It is not trivial task because central zones for example environment of capital city usually have biggest traffic, and it is only few zones. So, sometimes we should analyze the extreme zones, and if need, add more historical data about this zones.

Main idea of Bayesian method is based on Bayes theorem. First we define the optional classification rule. Let Y_i, a one row (element) of input set is member of i-th class. Vector X is observable attributes of one element. Then we put element to j-th class, where $P(Y_i \mid X)$ is maximal. Now we can use Bayes theorem, which say:

$$P(Y_i \mid X) = \frac{P(X \mid Y_i)P(Y_i)}{P(X)}$$

$P(X)$ is constant for each class. So, to maximize $P(X \mid Y_i)P(Y_i)$ is enough. We can use maximum likelihood estimation for it.

The Bayesian method [4] has many versions, the base version works independent attributes, and however our statistical data set is not independent. So, we ought to use Bayesian Networks, it can handle correlations. On the other hand, in future we would like increment trainer set. In this case we can use additive Bayesian network. It has more complex algorithm, so we need more powerful hardware for execute implementation. Moreover statistical office usually gives very redundant data set. For example in Hungary we get 1024 attributes; witch is too many for store build correlation network in real hardware.

In our implementation we used principal component analysis with varmax rotation before Bayesian network, our analysis showed, we need only 5 factors.

Summing up, we known classes traffic metrics, we use this as estimation of size of input and output traffic flow for zones. So, in this way we hide the data gaps in input data, therefore we will be able to use gravity model with deficient information.

4.2. Increment trainer set by expert feedback

Simulation is useful, if it is realistic. In introductions we showed, for this we need realistic OD matrix. We cannot define realistic in mathematical way, so in this point we should add opportunity to traffic experts to modify the OD. However we would like to use this modification to train our model. Moreover, we want only few expert corrections, but we want correct complete OD-matrix.

In our solution, we take advantage of some property of Bayesian method. The Bayesian classifier do not only assigns zones to classes, but also show probability of result. So, we know which zone may be on incorrect class. And we can ask it from expert. After, we can add the answer to trainer set with high weight.

In this way training set will be grow, the bigger training set leads to more precise classifying. Therefore the system will learn from expert. It doesn't mean,

we shouldn't ask from expert after some years usage. But experts should correct only extreme zones.

The best advantage of this procedure is the user-friendly feedback opportunity. In this procedure experts should add only one number; they can understand the model or implementation. Moreover experts can accept the result of Bayes method. And after the simulation expert can go back to feedback step. The experts can't see the unrealistic behavior in numbers, but they can see in simulation.

In practice experts don't like modifying OD matrix after simulation, because is too big back step, but in our solution they can do in easy way.

4.3. Gravity model

In our idea, we choose M_i to count of outgoing vehicles from i. zone in gravity model [6]. Moreover OD matrix isn't symmetric, therefore M_j should be count of incoming vehicles from j. zone.

Gravity model has one mistake; we have to choose distance function. We can believe it's easy, Euclidian distance is good. However we work with road network, where people calculate by traveling time. For example, let's see tow zone different side of a river. If here are bridges people will travel between the zones. If here aren't bridges, then people will not move between this tow zones. About this problem, we find the following: in small area like city, the graph distance is better than Euclidian, however in country level the Euclidian better.

4.4. Our result

We implemented our method and we used public KSH data to execute simulation. In figure 2 we can see plot classified zones. Colors show the class of zones. You can see zone centroids in, the size represent demographical gravitational attraction. We can see it is may be true. Bigger city has bigger traffic. We have a reference matrix. The solution have surprised me, the extremity values approximated with high precision, and more than 90values error rate was under 1.5. In this salutation we used only historical OD matrixes and KSH data, we can say it is smaller data set than other country level data sets which are in used for traffic simulation [7].

5. Conclusion

In the previous sections we presented a new OD matrix estimation method. This method is less sensitive for quality and amount of input data than most of existing methods. We get idea from computer networks [3]. And we mixed it with road traffic science. In our solution we hide bad quality of input data by Bayesian classifying, each class member has similar size of traffic. Finally we use Demographical Gravitation model.

This procedure is enough flexible to correct the model by expert's feedback. It's user friendly model, so experts can't understand it. Here we take advantage of the

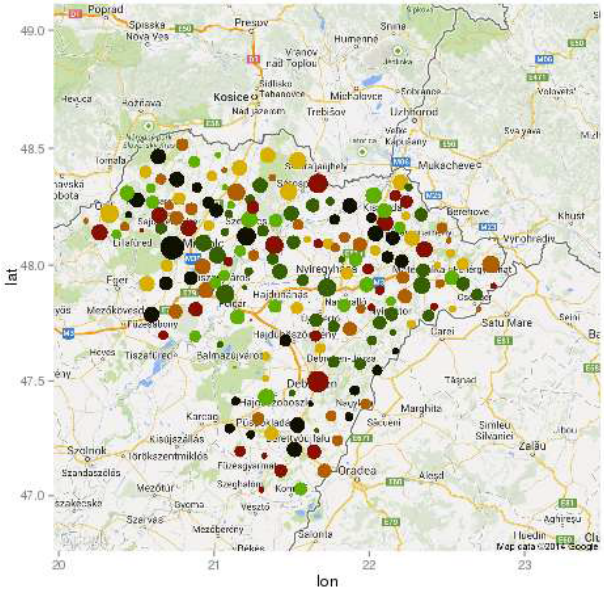

Figure 2: Centroids of zones

Bayesian classifier's probability model. Therefore we could find the low probability results, and we could ask it from experts.

6. Future goals

Each country has own traffic behavior, so we have to work out a global calibration process for our method. So we need process which calibrate the gravitation model and probability model of Bayesian method for any region. We note, both of them have existing calibration procedure, but we are looking for integrated calibration algorithm.

References

[1] AU, C., YUEN, M., Unified approach to NURBS curve shape modification, Computer-Aided Design Vol. 27 (1995), 85–93.

[2] FOWLER, B., BARTELS, R., Constraint-based curve manipulation, IEEE Comp. Graph. and Appl., Vol. 13 (1993), 43–49.

[3] Anders Gunnar, Mikael Johansson., Thomas Telkamp, Traffic Matrix Estimation on a Large IP Backbone– A Comparison on Real Data, Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, (2004), 149-160.

[4] Tom Auld, Andrew W. Moore, Bayesian Neural Networks for Internet Traffic Classification, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 18, NO. 1, JANUARY 2007

[5] G. Kotusevski, K.A. Hawick, A Review of Traffic Simulation Software, Res. Lett. Inf. Math. Sci., 2009, Vol. 13, pp. 35-54

[6] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, C. Diot, Traffic matrix estimation: existing techniques and new directions, SIGCOMM Comput. Commun. Rev., (2002), 161-174.

[7] Voellmy, A. Cetin, N., Vrtic, M., Nagel, K, Towards a Microscopic Traffic Simulation of All of Switzerland, Computational Science — ICCS 2002, (2002), 371-380.