Development of Origin-Destination Trip Matrices Using Mobile Phone Call Data

By

Md. Shahadat Iqbal

Department of Civil Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

January, 2013

Development of Origin-Destination Trip Matrices Using Mobile Phone Call Data

By

Md. Shahadat Iqbal

A thesis submitted to the Department of Civil Engineering,
Bangladesh University of Engineering and Technology, Dhaka,
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN CIVIL ENGINEERING
(Transportation)

Department of Civil Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

January, 2013

The thesis titled **"Development of Origin-Destination Trip Matrices Using Mobile Phone Call Data",** Submitted by **MD. Shahadat Iqbal, Roll No: 0411042439F, Session: April, 2011,** has been accepted as satisfactory in partial fulfillment of requirement for the degree of Master of Science in Civil Engineering (Transportation) on 15th January, 2013.

BOARD OF EXAMINERS

Dr Charisma F. Choudhury Assistant Professor, Department of Civil Engineering, BUET, Dhaka.	**Chairman** **(Supervisor)**
Dr Md. Mujibur Rahman Professor and Head, Department of Civil Engineering, BUET, Dhaka.	**Member** **(Ex-Officio)**
Dr Tanweer Hasan Professor, Department of Civil Engineering, BUET, Dhaka.	**Member**
Enam Ur Md. Rahman Khan Principal Engineer, Planning, Design and Dimensioning Tech. Grameenphone Ltd. Grameenphone headquarter, Basundhara, Dhaka	**Member** **(External)**

DECLARATION

I hereby declare that the research work presented in this thesis has been performed by me and this thesis or any part of it has not been submitted elsewhere for any other purposes except for publication.

January, 2013

MD. SHAHADAT IQBAL

ACKNOWLEDGEMENTS

All praise to Almighty Allah, the most Gracious and most Merciful.

The author would like to express his sincere appreciation and gratitude to his supervisor, Dr Charisma F. Choudhury, Assistant Professor, Department of Civil Engineering, Bangladesh University of Engineering and Technology (BUET), for her continuous guidance, invaluable suggestions and affectionate encouragement at all stages of this study. Without her valuable direction and cordial assistance, this research work could never be materialized. The author's debt to her is immense.

The author would like to express his deepest gratitude to the Department of Civil Engineering, BUET, the Head of the Department of Civil Engineering and all the members of BPGS committee for giving me such a great opportunity of doing my M.Sc. and this contemporary research work.

The author is indebted to Grameenphone Ltd. for the data without which the thesis could not be initiated. Special thanks to Enam Ur Md. Rahman Khan, Principal Engineer, Design and Dimensioning Tech, Grameenphone Ltd. for his immense support regarding the data transfer issues. Thanks are also due to Mr. Tanveer Mohammad, Chief Technology Officer and Mr. Matiur Rahman Siddiqui, Head of Assurance Technology, Grameenphone Ltd. for their support and help in this regard.

The author would like to thank Mr. Kaashif R. Choudhury, Director, Research and Development at Track Innovative Technologies Ltd., for managing server for database management and also for valuable advice in MySQL code writing. Special thanks to Mr Hafiz Imtiaz, Lecturer, Department of Electrical and Electronics Engineering, BUET for providing MATLAB Unix version software.

The author is also indebted to HEQEP (Higher Education Quality Enhancement project) for providing funding for the data collection and image processing software acquisition. Also, the author would like to thank Juwel Harun, Md. Abu bakkar Siddique, Mrinal Kranti Paul and Imran Hossain for their help in this thesis work.

Finally, the author would like to recall his family for their continuous support and motivation throughout the research.

ABSTRACT

Reliable Origin-Destination (OD) matrices are important for transportation planning and analyses. However, detailed field data collections by means of extensive questionnaire surveys or license plate matching are very expensive. On the other hand, mobile phone users leave footprints of their approximate locations whenever they make a call or send an SMS. As mobile phones have reached a high penetration rate in both developed and developing countries, the mobile phone call records can be used to extract the OD patterns of travelers. This can be scaled up to determine the actual OD matrix of a transport network.

In this thesis, two kinds of data are used for constructing the OD matrices: mobile phone Call Detail Records (CDR) and traffic counts extracted from video recordings. CDR is collected from Grameenphone (GP), the largest mobile operator of Bangladesh. The Dhaka city major road network has been divided into 29 nodes. Each BTS (Base Transceiver Station) is assigned to a specific node. So if someone makes a phone call, his/her location is represented in terms of a definite node. If he moves to another node and makes another call then its position is updated to that node and an entry is recorded in the corresponding OD. The database management software MySQL and MATLAB are used to process data from 1,397,118 users (one day) to determine the OD pattern (termed as seed OD in this thesis).

Video data is also collected for three selected days of the same month (eight hour each day) in thirteen important locations of Dhaka city road network. The traffic of Dhaka city is then simulated using a modified version of the microscopic traffic simulator MITSIMLab. The mobile phone CDR data are used as seed OD and an optimization based approach is deployed to determine the scaling factors that minimize the differences between the observed and simulated traffic counts.

The main outcome of this research is a methodology to calculate OD matrix for microscopic traffic simulator using mobile phone CDR data and limited traffic video data. This approach is not only economic but also suitable for easy periodic update of the OD matrix and also extendable in the context of dynamic traffic simulation. This could be a useful tool for local transportation agencies for generating their local OD matrices and subsequently using those in the microscopic simulation software for evaluation and impact assessment of alternative transport policies. It may be noted that although this research has taken Dhaka network as a test scenario, this methodology can be used for other city networks also.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF NOTATIONS AND ABBREVIATION

AUC	AUthentication Center
BSC	Base Station Controller
BSS	Base Station System
BTS	Base Transceiver Station
CDR	Call Details Records
EIR	Equipment Identity Register
GLS	Generalized Least Squares
GP	Grameenphone
GPS	Global Positioning System
GSM	Global System for Mobile communications
HLR	Home Location Register
NMC	Network Management Center
ML	Maximum Likelihood
MS	Mobile Station (or Mobile Phone)
MSC	Mobile services Switching Center
MVN	Multivariate Normal
OD	Origin-Destination
OMC	Operation and Maintenance Center
RAM	Random Access Memory
RNE	Road Network Editor
SMS	Short Message Service
SS	Switching System
TAZ	Traffic Analysis Zone
VLR	Visitor Location Register
WLAN	Wireless Local Area Network

CHAPTER ONE

INTRODUCTION

1.1 Background

Reliable data on usage pattern is an essential pre-requisite for transportation planning and design. The most commonly used form of data required for planning and design purposes is traffic flow rates in between the different zones of the design area or between different nodes of the transportation network. For example, to design a road connecting Point-A to Point-B, it is vital to know how much traffic (or how many people) are going to travel through that road. If the flow in between the point is low, then a single lane or dual lane carriageway may be sufficient. If the flow is greater, than the number of lanes can be increased according to the traffic needs. If the flow is not known than the design may be uneconomical (over-designed) or may be insufficient (under-designed). So for a perfect demand responsive design it is important to know the point-to-point (or zone-to zone flows in case of more aggregate level analyses). These zone-to-zone or point-to-point flows are presented by Origin-Destination (OD) matrices. OD matrix is not only important for transportation planning and design but also for transportation management purposes. For these reasons, a reliable and up-to-date OD matrix plays a fundamental role in successful transportation analysis.

Determination of the OD matrix is not a trivial task. Household survey data, roadside survey data, video data etc. are the main sources of data for generating OD matrix. But higher cost, longer time and lower reliability of those data collection has turned the transportation engineer to seek for different source of data for OD matrix generation. Modern trend is to use ubiquitous data sources like GPS, Mobile phone as a good source of data in transportation sector. Using these data, though still mostly limited to academic research, has already proven its superiority over the other data sources.

Now a day's mobile phone has become a part of parcel of everyday life. In Bangladesh the uses of mobile has been increasing. As mobile is carried everywhere a person travels so it can be a good source of travelling information for transportation engineers. But the use of mobile phone data has been mostly limited to some limited field. Phithakkitnukoon et al. (2010), Phithakkitnukoon et al. (2011) and Reades et al. (2009) work on visualization. Aggregate level analysis was done by González et al. (2008), Song et al. (2010), Simini et al. (2012), Candia et al. (2008) and Sevtsuk et al. (2010). Schlaich (2010) works on route choice modeling. Origin-destination flow estimation has also been done by Friedrich et al. (2010) and Calabrese et al. (2011) using mobile phone details data. But none of those used CDR data to generate OD matrix. This thesis has used Mobile Phone CDR data to generate a OD matrix which is a new addition in this field.

1.2 Research Objective

The main objective of this study is to generate an OD matrix (for the input of a microscopic traffic simulator) using the mobile phone CDR along with limited aggregate traffic counts (extracted from video data) collected from Dhaka city. The specific objectives of this thesis are as follows:

- Collecting mobile phone CDR and video data
- Create seed OD matrix using mobile phone CDR
- Determine the scaling factor by incorporating an optimization algorithm in conjunction with a microscopic traffic simulator
- Determine the actual OD matrix

The main outcome of this research will be development of the framework for calculating the OD matrix using mobile phone CDR and limited traffic count data and demonstrating it for the context of Dhaka city, Bangladesh. This approach is not only economic but also suitable for easy periodic update of the OD matrix and also extendable for dynamic traffic simulation. It may be noted that although this research will use Dhaka city as a test scenario but this methodology can be used to other networks also. This method would be very effective for generating complex OD matrix where land use pattern is heterogeneous and asymmetry in travelling pattern prevails.

1.3 Thesis Organization

Chapter 2 presents a review of available literature about various methods of OD estimation theories as well as of the new trends to use ubiquitous data in transportation engineering.

Chapter 3 describes the methodology employed in this study to achieve the stated objectives and the data collection techniques for this purpose.

Chapter 4 describes the details data collection procedures and description of the data for the case study.

Chapter 5 deals with different aspects of data analysis and development of models on a step by step basis.

Chapter 6 presents the conclusions of the entire study and suggests recommendations for future research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Many methodologies have been developed to estimate the Origin-Destination(OD) matrix. All those attempt to make the OD matrix more reliable with low cost and efforts. The traditional way of determining OD matrix is to use extensive surveys, license plate matching and/or link counts. The availability of ubiquitous data sources (like GPS data, mobile data etc.) has lead the transportation engineer to use those data in OD matrix estimation. In this chapter the details literature review of OD estimation is described. Some modern trends of mobile data uses in transportation are also discussed here.

2.2 Origin-Destination Matrix Estimation Models

There are many models for Origin Destination Matrix estimation. Lots of review has also been done on those models. Among those Abrahamsson (1998) gave a good review on existing literature on estimation of Origin-Destination Matrices using traffic count. The main part of it which is related to this thesis is presented below.

2.2.1 Problem specification

The region is divided into a number of zones represented by centroids. A set of directed links A connect a set of nodes N and the centroids makes up a subset of these nodes. For a subset, $\hat{A}$ of the links and for a specific time period (e.g. peak hour, average day), traffic count data are assumed to be available. The estimation problem is equivalent to finding a reasonable OD matrix g which, when assigned to the transportation network, reproduces the observed traffic count data. In a practical application the reproduction might not be exactly achieved for all traffic counts. This may be explained by internal inconsistencies in the traffic counts due to traffic flow collection at different times or the aggregated transportation network representation. The static OD matrix estimation problem treated here is referring to observed traffic counts for one single time period.

A crucial point in the estimation of an OD matrix using traffic counts is the assignment technique used: what route(s) in the transport network do trips from zone i to zone j take. The matrix P with elements p_{ij}^{a} represents of the proportion of trips between zone i and zone j that uses link a,

$$\mathbf{0 < p_{ij}^{a} < 1}$$

For a given link a, the sum of all $\boldsymbol{g_{ij}}$ (OD-flows from i to j) traversing this link is the link volume, $\boldsymbol{V_a}$. The car occupancy factor R (assumed to be 1:25) is included in this fundamental equation relating the link volumes and the OD-flows:

$$\boldsymbol{V_a = \frac{1}{R} \sum_{ij} p_{ij}^{a}\, g_{ij},\ a \in A} \quad \textbf{(2.1)}$$

Where, A is the set of links in the transportation network. The matrix with elements $\boldsymbol{p_{ij}^{a}}$ is often denoted the assignment matrix P. Depending on the treatment of congestion exogenous or endogenous determination of the assignment matrix is postulated:

I. **Proportional assignment:** In this case we assume independence between the traffic volumes and traffic proportions $\boldsymbol{p_{ij}^{a}}$. The link volumes (v_a) are proportional to the OD-flows (g_{ij}). The proportion of travelers choosing a route will not depend on congestion in the network but only on traveler and route characteristics. The values of $\boldsymbol{p_{ij}^{a}}$ can be determined before the estimation of the OD matrix is done and taken as exogenously given. The “all-or-nothing” assignment method can be used: all-or-nothing assignment of traffic is obtained when all traffic, for all OD pairs, is assigned to the cost minimizing route(s).

II. **Equilibrium assignment:** Wherever congestion effects are important, equilibrium assignment is a more realistic approach. The cost for travelling on a link depends on the flow volume, through a cost-flow relation. Equilibrium assignment techniques try to satisfy Wardrop's first equilibrium principle (Wardrop, 1952): the traffic system is in “equilibrium” when no traveler (user) can achieve a lower travel cost by switching to another route. The value of $\boldsymbol{p_{ij}^{a}}$ will depend on the volume on all links, i.e. $\boldsymbol{p_{ij}^{a}}$ = $p^{a}_{ij}(v)$, and cannot be determined independently of the trip matrix estimation process. The equations (2.1) become nonlinear in v.

A flow pattern with link flows v_a, $a \in A$, and a related assignment matrix P(v) that fulfills the equilibrium assignment requirements of ii) is recognized as being of the user-equilibrium type. The travelers are said to comply with a user-optimal behavior. Alternatively, a stochastic equilibrium concept can be employed that takes into account that travelers have different perceptions of travel costs, e.g. due to individual variations. The result is that multiple routes with unequal travel costs are used between any pair of zones. As noted in Cascetta and Nguyen (1988) the computational complexity of an OD matrix estimation model depends largely on the assignment technique used. Building on an equilibrium assignment technique and explicit treatment of congestion effects is more burdensome than relying on proportional assignment.

Given observed link volumes $\hat{\boldsymbol{v}}_a$ on a subset of the links $\hat{\boldsymbol{A}}$ and traffic proportions $\boldsymbol{p_{ij}^{a}}$ the OD matrix g is determined by solving the equation system (2.1). This equation system is normally highly underspecified: there are many more elements in the OD matrix g than links for which traffic counts are collected. Thus additional data (prior information) and/or assumptions about the travel behavior are needed in order to find a unique OD matrix.

A central part of the prior information is the target OD matrix. In statistical approaches the target OD matrix is typically assumed to come from a sample survey and is regarded as an observation of the “true” OD matrix to be estimated. The true OD matrix is assumed to belong to some statistical distribution and may be obtained by estimating the parameters of the statistical distribution. In traffic model based approaches, the target OD matrix is normally assumed to be an old OD matrix and one may ask for adjustments of the target OD matrix to satisfy the traffic counts. The distance between the estimated OD matrix and the target OD matrix is minimized subject to the flow constraints. In common for both approaches, the problem of finding the OD matrix g, given the target OD matrix $\hat{\boldsymbol{g}}$, is stated as minimizing a function $F_1(g; \hat{\boldsymbol{g}})$. In the notation below all prior information on the OD matrix is contained in the target matrix $\hat{\boldsymbol{g}}$.

Statistically, the observed set of traffic count data may also be assumed to be an observation of the "true" traffic count data to be estimated, related to and obtained as an assignment of the estimated OD matrix. Also for other reasons, touched upon above, deviations between estimated counts and observed counts may be accepted (this conception contrast with the alternative assumption of exact reproduction of the observed traffic volumes). Hence, an OD matrix is sought which produces "small" differences between the estimated link flows v and the observed flows $\hat{\boldsymbol{v}}$. This ambition can be expressed as a criterion F_2 (v; $\hat{\boldsymbol{v}}$) to be minimized subject to the assignment constraints.

Although the underlying motivations and assumptions of the OD estimation approaches differ the related optimization problems can be expressed in the following general form: Determine the demand for traffic between zones of the region, i.e. the OD matrix g solving the program:

$$\begin{aligned} \boldsymbol{min\, F(g, v) = \gamma_1 F_1(g, \hat{g}) + \gamma_2 F_2(v, \hat{v}) \quad v, g \geq 0} \\ \boldsymbol{s.t, \;\; v = assign(g)} \end{aligned} \tag{2.2}$$

Where $\hat{g}$ is the target OD matrix and $\hat{v}$ the observed traffic counts with F_1 and F_2 being some distance measures. The assignment of g to the transportation network is denoted assign(g), leading to a split of the OD-flows (g_{ij}) over available routes with path flows (h_{ijk}).

If the target OD matrix is very reliable and accurate γ_1 should be large compared to γ_2 which would result in a g close to $\hat{\boldsymbol{g}}$. Then larger deviations between v and $\hat{\boldsymbol{v}}$ would be accepted. If, on the other hand the observed traffic counts are reliable compared to the information in $\hat{\boldsymbol{g}}$ the magnitude of γ_2 should be large compared to γ_1. The second part of (2.2) would then guide the optimization, leading to estimated flows (v) that are close to the observed values ($\hat{\boldsymbol{v}}$) while larger deviations between the estimated OD matrix (g) and the prior information ($\hat{\boldsymbol{g}}$) would be accepted. The values of the weights (γ_i) are thus closely related to the conception of the modeling situation. General multi-criterion models in the case of proportional assignment were discussed in Brenninger-Gothe et al. (1989).

The distance measures F_1 are mainly of the minimum-information type and F_2 is often taken to be a Euclidean distance measure. In principle, any other combination of measures is conceivable.

2.2.2 Survey of modeling approaches

In the survey below, the first category of approaches is based on traffic modeling concepts. These approaches include the "minimum information" ("entropy maximizing") model and combined models for traffic planning. An estimate of the OD matrix is obtained by direct solution (F_1 = entropy measure) or by estimating the parameters of the combined model. The second category, statistical inference approaches, includes the Maximum Likelihood (ML), Generalized Least Squares (GLS) and Bayesian Inference approaches. Here the traffic volumes and the target OD matrix are assumed to be generated by some probability distributions. An estimate of the OD matrix is obtained by estimating the parameters of the probability distributions. The target OD matrix $\hat{\boldsymbol{g}}$ is in traffic modeling approaches normally obtained as an old (outdated) OD matrix, while statistical approaches rely on a target OD matrix obtained from sample surveys.

Gradient based solution techniques have been proposed, that solve optimization problems obtained from traffic modeling or statistical inference based approaches. This technique is important because an efficient solver is provided which can and has been applied to large-scale estimation problems based on equilibrium assignment.

Earlier reviews have been published by Cascetta and Nguyen (1988) and Nguyen (1984). Cascetta and Nguyen focus on statistical inference techniques while Nguyen concentrates on minimum information or entropy maximizing approaches. The survey of Cascetta and Nguyen does not promote any particular approach but provides a general framework for the problem of estimating an OD matrix from traffic counts. Nguyen (1984) only considers car traffic while Cascetta and Nguyen (1988) also consider transit traffic. In Fisk (1989) three entropy maximizing models are compared and under certain conditions shown to result in equivalent estimates. Some issues of implementation and computation are touched upon in this section. More comments related to these issues can be found in the annotated bibliography of section 4. This survey differs from the other reviews by its explicit intention to identify candidates for medium scale applications.

The categorization used is not the only relevant classification. Referring to the figure in the introduction another classification is related to the treatment of congestion. In the characterization used by Florian (1986) the models are classed into three groups: network equilibrium approaches, gravity-entropy models and combined distribution-assignment models. Most equilibrium based methods have not been applied to networks of sizes relevant in practical planning situations.

2.2.2.1 Traffic modeling based approaches

Because the information provided by the traffic counts on some links is insufficient to determine a unique OD matrix, it is possible to argue that one should choose a "minimum information" OD matrix. This is an OD matrix that adds as little information as possible to the information in the target OD matrix, while taking the equations relating the observed traffic counts with the estimated OD volumes into account. The estimated, minimum information matrix is obtained from minimizing the function I that correspond to F_1 in equation (2:2) of section 22.1 (see e.g.

Snickars and Weibull (1977) for a combinatorial derivation):

$$\boldsymbol{I} = \sum_{ij} \boldsymbol{g}_{ij} \ln\left(\frac{\boldsymbol{g}_{ij}}{\hat{\boldsymbol{g}}_{ij}}\right) \tag{2.4}$$

This is the minimum information or entropy maximizing ($\hat{\boldsymbol{g}}_{ij} = 1$) function. An OD matrix minimizing (2.4) while reproducing the traffic count constraints and taking the information contained in the target OD matrix $\hat{\boldsymbol{g}}$ into account may be derived as:

$$\boldsymbol{g}_{ij} = \hat{\boldsymbol{g}}_{ij} \boldsymbol{e}^{\lambda_1 p_{ij}^1 + \lambda_2 p_{ij}^2 + \cdots + \lambda_A p_{ij}^A} \tag{2.5}$$

Where each λ_i is a Lagrange multiplier associated with the constraint that relates the link flow with the travel matrix: equations (2.1). The expression (2.5) for g relies on the assumption of proportional assignment, i.e. constant p_{ij}^a .

Van Zuylen and Willumsen (1980) proposed two important models of this type. An OD matrix is found that when assigned to the network reproduces the observed traffic counts which are required to be consistent. Traffic modeling approaches, directly or indirectly, assume that the trip making behavior is represented by a certain trip distribution model. The models of Van Zuylen and Willumsen are based on minimum information and entropy maximizing principles leading to trip distribution models of the gravity type. In the theoretical part, a target matrix $\hat{\boldsymbol{g}}$ was assumed but in all applications presented, all origin-destination combinations were assumed equally probable, i.e. $\hat{\boldsymbol{g}}_{ij}$ = 1.

Fisk (1988) extended the entropy model of Van Zuylen and Willumsen to the congested case by introducing the user-equilibrium conditions as constraints. Smith (1979) has showed that the model of user-optimal behavior can be expressed as variational inequalities, as stated below. The proposed model has a bilevel structure that maximizes the entropy on the upper level and solves a user-equilibrium problem on the lower choice level:

$$\boldsymbol{min}\, g_{ij}, h_{ijk} \sum_{ij} g_{ij} \left(\ln(g_{ij}) - 1\right)$$

$$s.t.\begin{cases} C(h).(f-h) \geq 0 & \forall f \\ \sum_k h_{ijk} = g_{ij} & \forall i,j \\ \hat{v}_a = \sum_{ijk} \delta_{ijk}^a . h_{ijk} & \forall a \in \hat{A} \\ g_{ij}, h_{ijk} \geq 0 & \end{cases} \quad (\mathbf{2.6})$$

Where, C(h) is the cost of travel on paths given path flows h and f is any feasible path flow solution. If the observed flow pattern is a user-equilibrium flow pattern, the extended entropy model of Fisk will have the same solution as a combined trip distribution and assignment model. This is shown in Fisk (1989). For combined models efficient solution algorithms exist. Earlier work includes the combined trip distribution/assignment model by Erlander et al. (1979) and the combined model by Fisk and Boyce (1983). These combined models have the number of trips originating in/attracted to each zone given (O_i and D_j respectively) and this may be expressed as constraints. One important difference between these models and the one by Fisk (1988) above is seen in the assignment constraints. The observed traffic counts do not appear in the combined model but are used to determine the value of the parameter in an estimation phase. One of the so called "doubly constrained" models considered by Fisk and Boyce has the optimization formulation:

$$\boldsymbol{min}\, g_{ij}, h_{ijk}\, \mu \sum_a [\int_0^{v_a} S_a(v).dv] + \sum_{ij} g_{ij} . \ln(g_{ij}) \quad (\mathbf{2.7})$$

$$s.t.\begin{cases} \sum_{i \neq j} g_{ij} = O_i & \forall i\ (\alpha_i) \\ \sum_{i \neq j} g_{ij} = D_i & \forall i\ (\beta_i) \\ \sum_k h_{ijk} = g_{ij} & \forall i,j\ (X_{ij}) \\ v_a = \sum_{ijk} \delta_{ijk}^a . h_{ijk} & \forall a \\ g_{ij}, h_{ijk} \geq 0 & \end{cases} \quad (\mathbf{2.8})$$

Where, the Lagrange multipliers are indicated in the parenthesis in the right-hand column. The solution to the minimization problem (P_{FiBo}) can be obtained by forming the La-grangian of the problem and setting the first order derivatives of the Lagrangian with respect to h_{ijk} and g_{ij} equal to zero:

$$\sum_a \boldsymbol{S_a(v_a).\delta^a_{ijk} - X_{ij} = 0} \ \forall \boldsymbol{k};\ \boldsymbol{h_{ijk}} > 0 \qquad (\mathbf{2.9})$$

$$\sum_a \boldsymbol{S_a(v_a).\delta^a_{ijk} - X_{ij} \geq 0} \ \forall \boldsymbol{k};\ \boldsymbol{h_{ijk} = 0} \qquad (\mathbf{2.10})$$

$$\boldsymbol{g_{ij} = e^{-\alpha_i - \beta_j} e^{-\mu X_{ij}} = A_i O_i\, B_j\, D_j\, e^{-\mu \bar{C}_{ij}}} \qquad (\mathbf{2.11})$$

Equations (2.9) and (2.10) stipulate that the routes that are actually used, i.e. with h_{ijk} > 0, should have a minimal cost for travel between the relevant zone pair ij. The Lagrange multipliers X_{ij} are thus equal to the minimum costs for travel. From now we set $\boldsymbol{X_{ij}}$ to $\boldsymbol{\bar{C}_{ij}}$, the minimal costs in the road network. Equation(2.11) include the Lagrange multipliers $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_j}$ that are determined so that the trip production and attraction constraints are fulfilled. The balancing factors α_{ij} and β_{ij} are transformed Lagrange multipliers, A_i O_i = A_iO_i $= e^{\boldsymbol{\alpha_i}}$ and B_jD_j $= e^{\boldsymbol{\beta_j}}$ ^{j}. The value of the parameter $\boldsymbol{\mu}$ is undetermined in the problem formulation above and is calculated in an estimation phase.

The importance of this model stems from the relative ease with which a combined distribution and assignment model can be solved. Erlander et al. prove that the model may be rigorously estimated and that the corresponding optimization problem has a unique solution if the sum of the integrals of all link cost functions is known:

$$\hat{C} = \sum_{a\, \in A} \int_0^{\hat{f_a}} c_a(x)dx \qquad (\mathbf{2.12})$$

The observed traffic counts are reproduced by a combined model if observed counts for all links are available and if they are consistent with user-equilibrium. Other model refinements have been performed by for example Erlander, Jornsten and Lundgren for example in a technical report written in 1984. The problem of missing data in the base year, used in the estimation phase, is studied. The authors solve this problem by the prediction of values to be used in a prediction year.

If observed traffic counts only are available for a subset of the links in the network, Fisk and Boyce (1983) suggest how an estimate of $\hat{C}$ may be obtained, “weighting” average link costs by the importance of various link types in the network representation. The model of Fisk and Boyce will in general not reproduce the observed traffic counts. This may to some extent relate to the fact that the observed traffic counts might not constitute an equilibrium flow pattern. Different suggestions on procedures that remove internal inconsistencies have been proposed but no effective procedures for assuring the user-equilibrium property and internal consistency of the observed flow pattern seems to exist. See Yang (1994) for a recent treatment and review of these problems.

In Kawakami et al. (1992) the combined model of Fisk and Boyce is extended further to include two modes of travel, large size trucks and cars. An application to a medium sized network in the city of Nagoya, Japan, is reported. A target matrix appears to have been available but is not included in the model formulation.

In Tamin and Willumsen (1989) not only the gravity model but also the intervening opportunity model is considered, both models are of the doubly constrained type. Applications to a small test problem without congestion using a gravity model, an intervening opportunity model and a gravity-opportunity model are presented.

Nguyen (1977) presented one of the first formulations of the equilibrium based OD matrix estimation problem. The network was congested and Nguyen analyzed properties of the solution obtained thoroughly. The solution will reproduce the observed traffic counts. Though the solution is unique in link flow variables there normally exists many different OD matrices that correspond to these estimated link flows. It remains to choose a criterion for determining a unique OD matrix from all the different OD matrices that reproduce the observed counts.

Jornsten and Nguyen (1979) and later LeBlanc and Farhangian (1982) started from the assumption of equilibrium assignment and formulated models of the entropy maximizing and the "minimum least squares" type, respectively. The first approach does not require a target OD matrix while the model of LeBlanc and Farhangian(1982) also relies on a known target OD matrix. The model development starts from work suggested by LeBlanc in Gur et al. (1980). The motivation for least squares formulation is to obtain the OD matrix having a user-optimal behavior of travelers and being "closest" (in a GLS meaning) to the target OD matrix:

$$\boldsymbol{min} g_{ij} = \sum_{ij} \left(g_{ij} - \hat{g}_{ij}\right)^2$$

$$s.t. \begin{cases} \sum_{a} \int_{0}^{v_a} s_a(v)dv - \sum_{ij} \hat{C}_{ij} . g_{ij} = F \\ \sum_{k}^{a} h_{ijk} = g_{ij} \quad \forall i,j \; (\beta_i) \; (\hat{C}_{ij}) \\ v_a = \sum_{ijk} \delta_{ijk}^{a} . h_{ijk} \qquad \forall a \\ g_{ij}, h_{ijk} \quad \geq 0 \end{cases} \qquad \textbf{(2.13)}$$

Where, F is the objective function at optimum of Nguyen (1977)'s equilibrium based problem. This is a variable demand problem. The test problems presented in these two works are small.

Most formulations of the congested OD matrix estimation problem have a bilevel structure. The "upper" level problem estimates the OD matrix (assuming link flow volumes given) and the "lower" level problem is the equilibrium assignment problem that determines a link flow pattern of the equilibrium type. On the lower level the demand for travel, the OD matrix, is assumed given. Efficient solution algorithms for OD estimation problems having a bilevel structure are discussed in section 2.2.2.3.

2.2.2.2 Statistical Inference approaches

Maximum Likelihood:

The Maximum Likelihood (ML) approach maximizes the likelihood of observing the target OD matrix and the observed traffic counts conditional on the true (estimated) OD matrix. It is assumed that the elements of the target OD matrix $\hat{g}$ are obtained as observations of a set of random variables. The observed traffic counts $\hat{v}$ constitute another source of information about g, the OD matrix to be estimated, and $\hat{v}$ and $\hat{g}$ are usually considered to be statistically independent. The likelihood of observing $\hat{g}$ and $\hat{v}$ can be expressed as:

$$\boldsymbol{\mathcal{L}(\hat{g},\hat{v}|g) = \mathcal{L}(\hat{g}|g).\mathcal{L}(\hat{v}|g)} \qquad (\mathbf{2.14})$$

Due to the independence assumption on the observed traffic counts and the target OD matrix, the likelihood of observing both sets is equal to the product of the two likelihoods. Applying the ML principle for this problem amounts to seeking the OD matrix g that maximizes this likelihood. With the convention that 0 ln(0) = 0 we can as well maximize the logarithm of the product.

If the target OD matrix is obtained by simple random sampling in a region with a stable travel pattern, the target OD matrix may be assumed to follow a multinomial distribution. This is dependent on small sampling fractions α_i: if $\hat{g}_i$ trips are sampled out of a total of g_i trips at origin i then $\alpha_i = \hat{g}_i / g_i$. For the logarithm of the probability $\boldsymbol{\mathcal{L}(\hat{g}|g)}$ we have:

$$\boldsymbol{ln\,\mathcal{L}(\hat{g}|g) = \sum_{ij}(\hat{g}_{ij} ln(\alpha_i g_{ij})) + constant} \qquad (\mathbf{2.15})$$

This corresponds to F_1 in equation (2.2) of section 2. If the sampling fractions are sufficiently large, a Poisson probability distribution may be assumed for the target OD matrix and for the logarithm of $\boldsymbol{\mathcal{L}(\hat{g}|g)}$ we receive:

$$\boldsymbol{ln\,\mathcal{L}(\hat{g}|g) = \sum_{ij}(-\alpha_i g_{ij} + \hat{g}_{ij} ln(\alpha_i g_{ij})) + \ constant} \qquad (\mathbf{2.16})$$

If the observed traffic counts also are assumed to be generated by a Poisson probability distribution and independent of the target OD matrix, a similar expression for the probability $\boldsymbol{\mathcal{L}(\hat{v}|v(g))}$ is obtained, that is F2 in equation (2.2) of section 2:

$$\boldsymbol{ln\,\mathcal{L}(\hat{v}|v(g)) = \sum_{a\epsilon \hat{A}}(\hat{v}_a ln\big(v_a(g)\big) - v_a(g)) + \ constant} \qquad (\mathbf{2.17})$$

Where, $v_a(g)$ denotes the flow volume on link a resulting from an assignment of g. If a multivariate normal, MVN, distribution is assumed for the error terms of the observed traffic counts with zero mean and a variance-covariance matrix W, the expression for F_2 is:

$$\boldsymbol{ln\,\mathcal{L}(\hat{v}|v(g)) = -\frac{1}{2}\,\big(\hat{v} - \ v(g)\big)' W^{-1}(\hat{v} - \ v(g)) + \ constant} \qquad (\mathbf{2.18})$$

If (2.16), (2.17) and proportional assignments are valid assumptions, the OD matrix estimation problem can be formulated as:

$$\boldsymbol{max \sum_{ij}(-\alpha_i g_{ij} + \hat{g}_{ij} ln(\alpha_i g_{ij})) + \ \sum_{a\epsilon \hat{A}}(\hat{v}_a ln\big(v_a(g)\big) - v_a(g))}$$

$$\boldsymbol{s.t} \begin{cases} \boldsymbol{\sum_{ij} p_{ij}^a g_{ij}, \forall a\epsilon A} \\ \boldsymbol{g_{ij} \geq 0} \end{cases} \qquad (\mathbf{2.19})$$

This is one of the optimization problems considered by Spiess (1987). To solve problem (2.19) Spiess proposes an algorithm of the cyclic coordinate ascent type. The test examples of Spiess are small.

In the case of equilibrium assignment, problem (2.19) will have a bi-level structure with the assignment problem on the lower level and the estimation of the OD matrix on the upper level. See section 3.3 for solution methods of bi-level programs.

Generalized Least Squares:

The target OD matrix $\hat{\boldsymbol{g}}$ may be assumed to be obtained from the estimated, "true" OD matrix g with a probabilistic error term. In the same way the traffic counts may be viewed as obtained from a stochastic equation:

$$\hat{\boldsymbol{g}} = \boldsymbol{g} + \boldsymbol{\eta} \qquad (\mathbf{2.20})$$

$$\hat{\boldsymbol{v}} = \boldsymbol{v}(\boldsymbol{g}) + \boldsymbol{\epsilon} \qquad (\mathbf{2.21})$$

Where, is the probabilistic error that relates $\hat{\boldsymbol{g}}$ with g and $\boldsymbol{\epsilon}$ the error that relates the observed traffic counts $\hat{\boldsymbol{v}}$ with v(g). Often both $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are assumed to have zero means, $E(\boldsymbol{\eta}) = \mathbf{0}$ and $E(\boldsymbol{\epsilon}) = \mathbf{0}$.

Note that in the derivation of the generalized least squares (GLS) estimator below no distributional assumptions need to be made for the random parts and in (2.20 & 2.21). There is only a requirement on the existence of dispersion matrices. In the absence of accurate dispersion matrices often unity matrices (with diagonal elements equal to 1) have been used. There also exist a couple of reports on the sensitivity of the GLS model to using approximate dispersion matrices. Cascetta (1984) concludes that estimations or even "heavy approximations" of the dispersion matrix produced better results than a maximum entropy estimator. This independence of distributional assumptions is one important advantage of the GLS approach. One experience seems to be that the models are much more sensitive to variations and inaccuracies in the traffic count data and the target OD matrix than to values of the parameters, see e.g. Cascetta (1984) or Bierlaire and Toint (1994). The parameters include the dispersion matrices Z and W used in the problem formulation (2.22) below. As in the ML approach the target OD matrix and the observed traffic counts are assumed to be mutually independent. If the target OD matrix $\hat{\boldsymbol{g}}$ has an error with a variance-covariance matrix Z and the dispersion matrix of the traffic counts is W, the GLS estimator may be obtained by solving:

$$\boldsymbol{min}\ \frac{1}{2}(\hat{\boldsymbol{g}} - \boldsymbol{g})'\boldsymbol{Z}^{-1}(\hat{\boldsymbol{g}} - \boldsymbol{g}) + \frac{1}{2}\left(\hat{\boldsymbol{v}} - \boldsymbol{v}(\boldsymbol{g})\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{v}} - \boldsymbol{v}(\boldsymbol{g})\right)$$

$$\boldsymbol{s}, \boldsymbol{t}\ \boldsymbol{g}_{\boldsymbol{ij}} \geq \boldsymbol{0} \qquad (\boldsymbol{2.22})$$

The estimated OD matrix g should of course be constrained to be non-negative. Cascetta (1984) develops expressions for the mean and variance of the GLS estimator when non-negativity constraints are not active. An important property of the GLS approach is that the two sources of information, the observed traffic counts and target OD matrix, are readily combined. For example, if either dispersion matrix is close to zero, reflecting a great belief in this part of the information, the matrix inverse is very large. This means that the weights on the corresponding deviations are large and hence this part of the observed information is reproduced by the model when the minimum is attained.

The dispersion matrix Z can be approximated in different ways. If origin-based simple random sampling is adopted, an approximation that becomes sparse may be developed (see e.g. Cascetta and Nguyen (1988)). The dispersion matrix W is often considered to be diagonal and hence no covariance's between the different traffic counts are assumed.

Bell (1991) presents an algorithm that solves the problem (2.22). This algorithm explicitly considers the non-negativity constraint on the estimated OD matrix g. Both Bell (1991) and Cascetta (1984) assume proportional assignment. Bell derives the solution to the estimation problem (3.14) to be:

$$\boldsymbol{g} = \widehat{\boldsymbol{g}} + \boldsymbol{ZP}'(\boldsymbol{PZP}' + \boldsymbol{W})^{-1}(\widehat{\boldsymbol{v}} - \boldsymbol{P}\widehat{\boldsymbol{g}}) + (\boldsymbol{Z} - \boldsymbol{ZP}'(\boldsymbol{PZP}' + \boldsymbol{W})^{-1}\boldsymbol{PZ})\boldsymbol{\mu} \qquad (\mathbf{2.23})$$

Where, P is the assignment matrix and are the Lagrange multipliers associated with non-negativity constraints on the OD matrix g. From (2.23) one can see that the target OD matrix is “adjusted”, “changed” by two terms. The first term is related to the deviation of the observed traffic counts to those obtained by an assignment of the target OD matrix. The second term is related to active non-negativity constraints, i.e. constraints with non zero Lagrange multipliers corresponding to elements g_{ij} equal to zero.

If the traffic counts and the target OD matrix (their error terms $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$) are assumed to follow MVN distributions, the GLS estimator can be shown to coincide with a ML estimator. The same result is also produced using a Bayesian inference approach (Maher 1983) under the assumption of MVN distributions of the traffic volumes and the target OD matrix. In an earlier paper, Bell (1984) has shown that the minimum information approach of Van Zuylen and Willumsen (1980) is approximated by the GLS approach if the traffic counts are known to a high degree.

Yang et al. (1992) extend the GLS model by integrating equilibrium assignment of the OD matrix into the model. The OD matrix estimation problem is formulated as a bi-level program with the generalized least squares problem on the upper level and the equilibrium assignment problem on the lower level. The resulting problem becomes difficult to solve. A heuristic algorithm is suggested and applied to small problems (for other bi-level formulations and solution techniques, see section 2.2.2.3 below.)

If the observed traffic data is of the user-equilibrium type and available for all links an OD matrix that reproduces the observed link volumes can be obtained as a solution to an underspecified equation system. It is difficult to determine if the observed traffic data is of the user-equilibrium type. If the underspecified equation system is feasible, Yang et al. (1994) show that the traffic flow is of the user-equilibrium type. Adjusting observed traffic counts to be compatible with user-equilibrium is however difficult and no definite procedure has been suggested.

Bayesian Inference:

The Bayesian inference approach considers the target OD matrix as a prior probability function Pr(g) of the estimated OD matrix g. If the observed traffic counts are considered as another source of information about g with a probability $\boldsymbol{\mathcal{L}(\hat{v}|g)}$, then Bayes theorem provides a method for combining the two sources of information. For the posterior probability $\boldsymbol{f(g|\hat{v})}$ of observing g conditional on the observed traffic counts we then have:

$$\boldsymbol{f(g|\hat{v}) \approx \mathcal{L}(\hat{v}|g).Pr(g)} \qquad \boldsymbol{(2.24)}$$

The posterior probability function allows, in principle, for a determination of a confidence region for g but due to practical computational complications only point estimators may be obtained. This may take the form of the maximum value of the logarithm of the posterior distribution, the g that maximizes $ln\boldsymbol{f(g|\hat{v})}$. For the first term in (2.24), the observed traffic counts, a Poisson probability or a MVN distribution is usually assumed. The expressions for the logarithm of $\boldsymbol{\mathcal{L}(\hat{v}|g)}$ will then be (2.17) or (2.18). For the probability function Pr(g) a multinomial distribution may be assumed. Then, for the logarithm of Pr(g), using Stirling's approximation we have:

$$\boldsymbol{lnPr(g) = -\sum_{ij} g_{ij} ln(\frac{g_{ij}}{\hat{g}_{ij}}) + constant} \qquad \boldsymbol{(2.25)}$$

This is the minimum information function. A similar function is also obtained with a Poisson approximation of the multinomial distribution. If a multivariate normal distribution is assumed to hold for Pr(g), with mean q and dispersion matrix Z_q

Maher (1983) obtains:

$$\boldsymbol{ln\,Pr(g) \approx\ -\frac{1}{2}(g-q)'Z_q^{-1}(g-q)+constant} \qquad \boldsymbol{(2.26)}$$

Maher assumes that proportional assignment holds. For the traffic counts Maher makes the MVN assumption and shows that in this case the estimated OD matrix also becomes MVN distributed. The model of Maher is applied to a very small transportation network. The optimization problems of the Bayesian Inference approach contain, just as in the ML and GLS approaches, the sum of two parts. The first relates to the target OD matrix , (2.15) or (2.16), and the second concerns the observed traffic counts. The Bayesian Inference approach is a statistical inference technique with properties in common with the ML and the GLS approaches. But, as Cascetta and Nguyen note the roles assumed by g in the classical inference approaches (ML and GLS) and the Bayesian Inference approach differ. In the first case, the true g_{ij} are parameters of the likelihood function $\boldsymbol{\mathcal{L}(\hat{g},\hat{v}|g)}$ and in the second case the g_{ij} are random variables with given prior distributions.

2.2.2.3 Gradient based solution techniques

The most general optimization problems (with congested networks) of the previous sections 2.2.2.1 and 2.2.2.2 have a bi-level structure. The OD matrix estimation problem that determines g appears on the upper level and the equilibrium assignment problem on the lower level. Such problems were considered by Spiess (1990), Drissi-Katouni and Lundgren (1992), Yang et al. (1992), Florian and Chen (1993) and Chen (1994). The related solution algorithms have been applied to large city networks.

In gradient based solution techniques, the target OD matrix is taken as an initial solution to the OD matrix estimation problem. The target OD matrix is “adjusted” or “changed” to reproduce the traffic counts by iteratively calculating directions based on the gradient of the objective function. The link volumes are implicit functions of OD flows and obtained by the assignment procedure, v(g) = assign(g), of the user-equilibrium type. This means that the OD matrix estimation problem (using traffic counts) can be formulated in terms of g_{ij} variables only:

$$\min F(g) = \gamma_1 F_1(g, \hat{g}) + \gamma_2 F_2(v(g), \hat{v}) \qquad (2.27)$$
$$g \geq 0$$

where the F_i are appropriate distance measures. Except for the non-negativity constraints on g this problem is unconstrained. Drissi-Katouni and Lundgren propose general descent algorithms for solving the problem (3.19). Florian and Chen study a Gauss-Seidel type and an Augmented Lagrangian type method. Of these the former is more suitable for applications to problems of larger networks. Also Yang et al. have suggested a Gauss-Seidel type method. The algorithm of Spiess is more approximate and does in the practical applications presented performed well.
In theory all contributions except for the one by Spiess consider both F_1 and F_2 in the upper level problem. In applications to problems on networks having considerable size supplied by both Spiess and Drissi-Katouni and Lundgren only F_2 appears in the objective function. Chen (1994) and a comparative report by Denault (1994) contain results, where also the travel matrix variables, g_{ij}, are explicitly considered in the upper level objective function. Higher goodness-of- t seems to result from including the target OD matrix. As mentioned, an advantage of this approach is its computational tractability. Spiess presents applications to several large scale problems. The problems include an urban application of Bern, Switzerland, with about 2700 links and one interregional application to the road networks of Finland with about 12500 links. The method of Spiess is approximate since proportional assignment is assumed to hold locally and the method does not necessarily converge to a solution of the stated optimization problem. With this assumption the gradient of the objective function becomes easy to compute, attainable from the solution of two equilibrium assignment problems. The close relation to one of the methods suggested by Drissi-Katouni and Lundgren is shown by the latter ones. The descent directions employed in the Gauss-Seidel method of Florian and Chen may be interpreted as more elaborate and hence providing a less approximate solver to the estimation problem. As noted by Florian and Chen this interpretation also applies to the Gauss-

Seidel method of Yang et al..

The results obtained by Spiess are reasonable with a significant improvement of the goodness-of-fit. The method of Spiess is available within the "commercial" EMME/2 (1990) transportation planning system. The methods of Drissi-Katouni and Lundgren have been applied to problems from the city of Hull, Canada, with about 800 links. Drissi-Katouni and Lundgren investigate different gradient based descent directions possibly scaled by second order information. The results indicate that the computations involved are reasonable. The emphasis is on the quality of the search directions and not on attaining a computationally efficient OD matrix estimator. Chen applies a Gauss-Seidel type method to a medium size network of Winnipeg, Canada, with 2983 links, 154 zone cancroids' and with observed counts for about 2.3% of the links. The inclusion of OD matrix variables explicitly (i.e. F_1) estimates more precise results.

2.2.3 Annotated bibliography

2.2.3.1 Traffic Modeling Based approaches

Erlander, Nguyen and Stewart (1979):

On the calibration of the combined distribution-assignment model.

The models considered are of the combined distribution-assignment type similar to model (P_{FiBo}) as stated in section 2.2.2.1 The contribution of the paper is an investigation into sufficient conditions for calibrating the models. The observed values on the entropy or the total assignment cost for travel, eq. (2.12), are both shown to be sufficient to determine a unique value of the parameter of the model. On the other hand, by giving a counter example the total cost for travel is shown insufficient.

Erlander et al. show that if traffic counts are available for all links in the network the corresponding optimization problem has a unique solution, resulting in a unique OD matrix. For combined models the reader is referred to e.g. Boyce et al. (1983) or Abrahamsson and Lundqvist (1996). This formulation of the OD matrix estimation problem can handle large-scale applications.

Fisk and Boyce (1983):

A note on trip matrix estimation from link traffic count data.

Through the calibration of a combined distribution and assignment model an estimated OD matrix is obtained. Equilibrium assignment is assumed to hold and the observed traffic counts can be used to estimate the sum of the integrals of the link cost functions

Ĉ.

In practice, only a sample of traffic counts is available normally. The sample is generally not random and Fisk and Boyce proposes an unbiased procedure for estimating Ĉ. They stratify the sample into K groups with corresponding mean values of the cost function $\hat{C}_k$:

$$\hat{C_k} = \frac{1}{n_k} \sum_{a \epsilon A_k} \int_0^{\hat{f_a}} c_a(x) dx \qquad \mathbf{(2.28)}$$

Where, A_k is the set of links in group k and n_k the number of links in group k. The total cost may then be estimated as:

$$\hat{C} = \sum_{k=1}^{K} p_k \hat{C_k} \qquad \mathbf{(2.29)}$$

Where, p_k is the proportion of all network links in group k. With the estimate (2.29) of the total cost a doubly constrained combined distribution-assignment model is formulated. This problem has a unique solution and efficient solution algorithms exist, see e.g. Boyce et al. (1983) or Abrahamsson and Lundqvist (1996). The observed traffic count data are used in a very aggregate way and there is no requirement of a reproduction of the observed traffic counts. The relation to a bi-level formulation of the estimation problem as investigated by Fisk (1988 and 1989) should be mentioned.

No target matrix is required and no applications are reported. However, an extension of the model to include two travel modes has been developed and applied to the city of Nagoya, Japan, by Kawakami et al. in 1992.

<u>Fisk (1988)</u>

On combining maximum entropy trip matrix estimation with user optimal assignment.

The paper combines the entropy maximizing estimator of Van Zuylen and Willumsen with a user-equilibrium assignment procedure such as SATURN, Van Vliet (1982). In the latter work by Van Vliet, the user-equilibrium conditions are formulated as variational inequalities. By combining these two works, a mathematical programming problem with a bi-level structure is stated. No applications are reported. Prior information (target matrix) may be included.

See Fisk (1989) for linkages of this formulation to combined trip distribution and assignment models, e.g. Erlander et al.(1979), and outlines of solution methods.

<u>Fisk (1989):</u>

Trip matrix estimation from link traffic counts: the congested network case.

Three formulations for OD matrix estimation on congested networks are examined. When the observed traffic counts constitute a user-equilibrium pattern, the different formulations are shown to have equal solutions.
The first formulation is based on e.g. Jornsten and Nguyen (1979) that determines an OD matrix which reproduces the observed traffic counts and has maximal entropy.

The entropy maximizing model of Van Zuylen and Willumsen (1980), extended to the congested case by Fisk (1988), is the second formulation and nally formulations based on the combined distribution and assignment model, e.g. Erlander et al. (1979), are studied. The latter model does not have the bi-level structure present in the other two formulations. No applications and solution techniques of the models are explicitly discussed. For the combined model efficient solution algorithms exist. All models are based on equilibrium assignment.

The travel patterns that solve the models are shown to be equal if the observed flow pattern is of the user-equilibrium type. From a computational perspective Fisk favors the combined trip distribution and assignment model approach. Very few methods for assessing the consistency of observed traffic counts with user-equilibrium and for removing potential inconsistencies have been presented, see e.g. Yang et al. (1994).

Jornsten and Nguyen (1979)

On the estimation of a trip matrix from network data.

The paper extends the model of Nguyen (1977) relying on equilibrium assignment by seeking the OD matrix with maximal entropy. Also exogenous information on the OD matrix, such as trip production and attraction and the total number of trips or an old (target) trip matrix, may be considered. The model by Nguyen (1977) fails, as Nguyen has observed, to find a unique trip matrix due to the under specification problem. This problem is solved by requiring that the estimated matrix has maximal entropy. The resulting problem has a bilevel structure with equilibrium assignment on the lower level and a maximum entropy problem on the upper level. The observed traffic count data are internally consistent and are assumed to be of the user-equilibrium type and also available for all links.

An algorithm based on generalized Benders decomposition is developed. Three small numerical examples with up to 12 transportation links and a Sioux Falls, South Dakota, network with 76 links are reported.

Jornsten and Nguyen (1983) :

Estimation of an OD trip matrix from network data: dual approaches.

Following Jornsten and Nguyen (1979), methods based on user-equilibrium route choice and the minimum information principle are presented. A sequence of combined distribution and assignment problems are handled as part of solving the OD matrix estimation problem with a bi-level structure. Different formulations of the problem and the corresponding dual are derived and investigated. The computational requirements of the algorithm of Jornsten and Nguyen (1979) are extensive and the dual approaches presented here are claimed to be computationally superior in large scale applications.

Only the first combined problem need a complete cold start optimization. Later problems can use the previous solution as an initial solution to the present combined distribution and assignment problem. The observed traffic count data are constrained to be reproduced by the model. Applications to small problems including a network with 7 links and a Sioux-Falls, South Dakota, networks are referred to. The solution for the latter problem is better, with the same computational e ort, as compared to the results of the approach in Jornsten and Nguyen (1979).

Kawakami, Lu and Hirobata (1992)

Estimation of origin-destination matrices from link traffic counts considering the interaction of the traffic modes.

A combined trip distribution, modal split and assignment model is proposed. The model by Fisk and Boyce (1983) is extended to include also an entropy constraint with respect to mode choice. The trip distribution and modal split model is of the singly constrained nested combined type. If p_{ijmr} is the proportion of traffic from zone i to zone j by mode m and route r the entropy constraint with respect to mode choice reads:

$$-\sum_{ijm}(\sum_{r} Pijmr)\ln(\sum_{r} Pijmr) \leq E_M \quad \mathbf{(2.30)}$$

Where, E_M is the observed entropy value with respect to mode choice. An algorithm for solving the combined model is proposed and an application to a simpli ed road network in Nagoya, Japan, is presented. The transportation network has about 500 transportation links with two modes for travel (large sized trucks and car). No target OD matrices are assumed in the model but observed OD matrices are available and reasonably well reproduced. The model is based on equilibrium assignment but the degree of congestion in the application is not obvious from the results presented. The OD matrix estimation application is concerned with all day (24 hour) traffic, which suggests that a detailed analysis of congestion is not aimed for.

LeBlanc and Farhangian (1982):

Selection of a trip table which reproduces observed link flows.

The paper extends the model of Nguyen (1977) relying on equilibrium assignment by seeking the OD matrix closest to a target OD matrix. In the bilevel formulation, a GLS problem is stated on the upper level and the problem on the lower level is of the user-equilibrium type. The relation to the method of Jornsten and Nguyen (1979) is noted where the upper level problem is an entropy maximization problem. The lower level user-equilibrium problems are solved by the Frank-Wolfe method developed to use an evolving bounds technique.

Computational results from an application with 76 links “a Sioux Falls, South Dakota, network” are presented. The reported results are termed `attractive' and the observed traffic count data are reproduced.

Sherali, Sivanandan and Hobeika (1994)

A linear programming approach for synthesizing origin-destination trip tables from link traffic volumes.

A linear programming model consistent with user-optimal behaviour of travellers is developed. The model recognizes that the observed traffic counts often do not comply with user-equilibrium and is capable of handling a prior target matrix.

Most notable is the small computational demands compared to a maximum entropy model and a network equilibrium approach (relying on an algorithm proposed by Gur et al. (1980)) on small networks. The model is also applied to a portion of the real network of Northern Virginia being of a modest size. The algorithmic tuning in terms of convergence criteria and other 'model refinements for such real-world applications' are on the research agenda of the authors. The model may have computational limitations when applied to problems of larger networks.

Tamin and Willumsen (1989)

Transport demand model estimation from traffic counts.

The three transport demand models considered by Tamin and Willumsen are the gravity, the intervening opportunity and a combined gravity/intervening opportunity model. All models are of the doubly constrained type. Three different estimation methods are investigated corresponding to each one of the models. Proportional assignment of traffic is assumed and both an all-or-nothing and a stochastic assignment technique are used. No target OD matrix is assumed in the models though the results are compared to a base year, target matrix.

A small application to real data from the town of Ripon, Great Britain, is presented. The transportation network has 188 transportation links. The gravity-opportunity model is found to produce the best t in Ripon, i.e. the best matching of observed and estimated link volumes. If the estimated and observed OD matrices are compared the gravity model is however found to produce the best fit. The results indicate a dependence on the assignment technique. The amount of congestion in the Ripon application is negligible.

<u>Van Zuylen and Willumsen (1980)</u>

The most likely trip matrix estimated from traffic counts.

The proposed models estimate OD matrices that reproduce the traffic counts. Proportional assignment is assumed. The algorithms iteratively adjust an initially given target matrix. The estimated OD matrix is constrained to reproduce the observed traffic counts. The models are based on minimum information and entropy maximizing principles. The problem of inconsistent observed traffic counts is discussed and a maximum likelihood method to remove inconsistencies and produce a better estimation of the observed link flows is described.

In applications to a small artificial network having 72 links the models are found to perform well for all but one case. In all test cases no target OD matrix was assumed known and the models were of the maximum entropy type. In the case with unsatisfactory results the authors note that these could certainly be improved with better prior information (a target OD matrix).

<u>Willumsen (1984)</u>

Estimating time-dependent trip matrices from traffic counts.

The models proposed extend the entropy maximizing model by incorporating a target OD matrix. The objective function considered is:

$$-\sum_{ij} g_{ij} (\log\left(\frac{g_{ij}}{\hat{g}_{ij}}\right) - 1 \qquad \mathbf{(2.31)}$$

Congestion effects in urban areas are here treated by letting the assignment matrix P vary over time. The number of variables in the related problem is noted to be large. A heuristic solution approach is developed and some results are reported. The main focus of the paper is on the estimation of matrices that vary over time. This reference is included because of the minimum-information type of objective function.

2.2.3.2 Statistical Inference approaches

<u>Maximum Likelihood, Spiess (1987):</u>

A maximum-likelihood model for estimating origin-destination matrices.

In Spiess (1987) it is assumed that the elements of the target trip matrix g are obtained by sampling, for all OD pairs, Poisson variables. In a first model, Spiess estimates the means of the random variables by the maximum-likelihood technique:

$$min \sum_{i,j} \alpha_i \, g_{ij} - \hat{g}_{ij} \, ln(g_{ij} \, \alpha_i))$$

$$s.t. \begin{cases} \hat{v}_a = \displaystyle\sum_{ij} P_{ij}^{a} g_{ij} \,, a \in \hat{A} \\ g_{ij} \ \geq 0 \end{cases} \tag{2.32}$$

This model reproduces the observed traffic count data. If the feasible set defined by the constraints is non empty, there exists an optimal solution (this requires the constraints to be internally consistent). Proportional assignment is assumed. A convergent algorithm of the 'cyclic coordinate descent' type is developed. In a section on statistical inference, the validity of the model is displayed using tests derived from the asymptotic behavior of the distribution.

Also a doubly constrained model and a model where the observed traffic counts are not required to equal the traffic volumes obtained by an assignment of the estimated OD matrix are studied. Solution algorithms to both these extensions are supplied.

One important practical problem is mentioned: the derivation of the extended model (when the observed traffic counts are not required to equal the estimated traffic volumes) relies on the assumption of mutually independent traffic counts. This means that the observation of the traffic counts must take place at different times because, otherwise, a certain traveler might be counted on more than one link and the model may only be an approximate ML model.
An advantage of the model is the feasibility with respect to values of g^. The maximum likelihood method is always feasible while an entropy maximizing method might become infeasible because of zero valued elements of g. The test problems are small, i.e. networks with no more than 20 transportation links.

Generalized Least Squares, Bell (1991)

The estimation of origin-destination matrices by constrained generalized least squares.

The approach of Bell explicitly considers the non-negativity constraints on the OD matrix g. It is shown that this can improve the accuracy of the estimated OD matrix.

The route choice is assumed to follow a proportional assignment procedure. The optimization program of Bell is that of section 2.2.2.2 above with the solution stated in equation (2.23). It can be seen that active non-negativity constraints affect the estimated OD matrix g through the last term in (2.23) (also strictly positive matrix elements g_{ij} are affected). Bell further supplies a solution algorithm to the optimization problem with non-negativity constraints. The convergence is proved and expressions for variances and covariance's of the estimated OD matrix are derived.

A small numerical example with 5 transportation links is reported.

<u>Generalized Least squares, Bierlaire and Toint (1995):</u>

MEUSE: An origin-destination matrix estimator that exploits structure.

The model is a developed GLS estimator, (Bell (1991) or Cascetta (1984)), that also contains terms derived from parking survey data. A type of augmented Lagranian algorithm is used to solve the problem. The determination of weights corresponding to the different term should ideally reflect the relative confidence one has in the associated terms' and the values of different parameters are determined beforehand. The model assumes proportional assignment.

In applications to both an artificial example and a real case study of Namur, Belgium, the performance is displayed. Comparisons with a GLS and an equilibrium based estimator (relying on the SATURN, Van Vliet (1982), software package) are supplied and the "MEUSE-estimator" is found to estimate a reasonable OD matrix. The examination of the model sensitivity is noted to be important, in particular with respect to parameter variation. The results are found more sensitive to variations in the observed traffic counts than to changes in the parameters of the model.

<u>Generalized Least Squares, Cascetta (1984)</u>

Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator.

The route choice is assumed to follow a proportional assignment procedure. First Cascetta derives the estimator when the observed traffic counts are not necessarily reproduced by the estimated traffic volumes. If the observed traffic counts are known with a high degree of certainty the corresponding dispersion matrix of the traffic counts (W in equation 2.22) is very small and the related estimator is also studied. Statistical characteristics of the two GLS estimators are derived that include the expected value and the variance-covariance matrix of the estimators. The derivations rely on the assumption of a strictly positive OD matrix g, i.e. the non-negativity constraints on the travel matrix do not become active and no OD flows equal to zero are allowed.

In an application to a small artificial network with 5 zones, the effects of using approximate dispersion matrices and a comparison with an entropy maximizing estimator reproducing the traffic counts are studied. Approximate dispersion matrices only affect the characteristics of the estimator in a minor way and the GLS estimator has a mean square error superior to that of the entropy maximizing estimator.

Cascetta discusses the possibility of using an equilibrium assignment model and among others refers to models following e.g. Jornsten and Nguyen (1979). In the equilibrium assignment case it is claimed to be difficult, if ever possible, to obtain expressions for the statistical characteristics.

Generalized Least Squares, Yang et al. (1992)

Estimation of origin-destination matrices from link traffic counts on congested networks.

The OD matrix estimation problem is formulated as a bi-level program. Generalized least squares terms constitute the objective function of the upper level. On the lower level equilibrium assignment is assumed. Similar bi-level formulations have also been suggested by LeBlanc and Farhangian (1982) and Jornsten and Nguyen (1979) with a generalized least squares and entropy maximum objective function on the upper level respectively. The latter entropy maximizing model is discussed as an alternative formulation. The models of

Yang et al. do not require the observed traffic counts to be internally consistent nor to be of the equilibrium type. This is further discussed in comparisons with other equilibrium based models such as the one by Nguyen (1977) and Fisk (1988) that are interpreted as special cases obtained for certain parameter values. Fisk has discussed a version of the proposed solution algorithm. The important difference is that the observed traffic counts are viewed as observations of random variables and appear as a part of the upper level objective function and not as constraints to be fulfilled.

The difficulty with solving the bi-level program is noted and a heuristic solution procedure is proposed. The upper level and the lower level (user-equilibrium) problems are solved iteratively using proportional assignment from the lower level solution in the next upper level problem. The algorithm was shown to converge satisfactory in an application to a small test problem with a transportation network that has 24 transportation links. The method can be interpretated as a gradient based approach. Some statistical measures of error for the estimated OD matrix are calculated and show improvement. High quality traffic counts, with small internal variation, give more significant improvement.

Generalized Least Squares, Yang et al. (1994)

The equilibrium-based Origin-Destination matrix estimation problem

The observed traffic data are assumed to be of the user-equilibrium type and available for all links in the network. Then, it is shown that an OD matrix reproducing the traffic counts can be obtained by solving an underspecified system of linear equations. The problem of identifying if the observed traffic counts are of the user-equilibrium type can be solved by finding out if these equations have a feasible solution. Among the OD matrices that are compatible with the equation system, the OD matrix solving a generalized least squares or an entropy maximizing problem is taken to be the estimated OD matrix. The resulting model does not have the bi-level structure present in other models solving the congested OD matrix estimation problem.

Before the actual estimation problem is solved the observed traffic counts must be expanded to all links. An approach that determines the volumes on all links from an equilibrium assignment of the target matrix resulting in volumes of the missing links is suggested. These link volumes are scaled so that they become compatible with the total magnitude of the observed volumes. A small OD matrix estimation problem is supplied. Further work to develop estimation methods that also consider uncertainties in and non user-optimality of observed traffic volumes remains, compare Yang et al. (1992).

Bayesian Inference, Maher (1983)

Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach.

Maher assumes that the target (prior) OD matrix and the observed traffic counts follows multivariate normal (MVN) distributions and that proportional assignment of route choice holds. Traffic count data for all links are assumed available. The estimated OD matrix becomes MVN distributed.

An advantage, as to traffic modeling approaches, is the allowance of different beliefs in the target OD matrix and the observed traffic counts. The relative beliefs are expressed in the variance-covariance matrices related to the target OD matrix and observed traffic counts. An iterative solution method that updates the target OD matrix and the related dispersion matrix is suggested. At a general level, with dense dispersion matrices the computational demands are considerable and Maher studies some special cases with related updating equations. These cases include independent observations, observations without error and least-informative prior.

A small test example, taken from Bell (1983), with 6 OD-pairs and 4 links is presented. The minimum information (entropy maximizing) approach is seen to be one extreme case of a whole range of possibilities as to beliefs in the target OD matrix and the observed traffic counts. The minimum information approach represents a minimal belief in the target matrix.

2.2.3.3 Gradient based solution techniques

Chen (1994) :

Bi-level programming problems: Analysis, algorithms and applications.

The PhD thesis studies the bi-level programming problem thoroughly. Both historical and theoretical developments are reviewed. This includes results on the existence of solutions and the complexity of algorithms solving the problem. Our interest focuses on real world applications of the bi-level OD matrix estimation problem.

Two methods are suggested, an Augmented Lagranian method and a Gauss-Seidel type method. They were implemented by using macros of the "commercial" EMME/2 (1990) software. The first method shows satisfactory results on a small network taken from Nguyen and Dupuis (1984). Though only local convergence is guaranteed the estimated OD matrixes differ only marginally for different initial (target) OD matrices.

Chen argues that for large-scale problems algorithms that work directly in the link flow space (as opposed to the route flow space) and that do not require repeated objective function evaluations are needed. A heuristic Gauss-Seidel method is developed and applied to problems from Winnipeg, Canada, with a network containing 2982 road links, 154 zone centroids and about 2.5% of the links with observed traffic counts. It is concluded that the estimated results show higher quality if a target OD matrix is included in the objective function.

Denault (1994:)

Two OD matrix adjustment methods applied to problems of congested networks are com-pared. The first is an Augmented Lagranian approach suggested by Chen (1994) and secondly a gradient based method proposed by the author is studied. The latter relies on sensitivity analysis developed by Tobin and Friesz (1988), compare the gradient based method of Drissi-Katouni and Lundgren (1992). The performance of the methods were studied on a small network and a medium sized network of Winnipeg, Canada. The first method performed well on the small network but problems related to the large number of parameters were experienced in the Winnipeg application. The gradient method resulted in excellent results and was preferred because of its relative ease of use.

<u>Drissi-Katouni and Lundgren (1992):</u>

Bi-level origin-destination matrix estimation using a descent approach.

Following the formulation in equation (3.19) a general descent algorithm consistent with equilibrium assignment is proposed. The gradient of the objective function F is:

$$\nabla F(g) = \gamma_1 \nabla F_1(g) + \gamma_2 \nabla F_2(v(g)) \qquad (\mathbf{2.33})$$

The difficulty with computing this gradient is related to the Jacobian $J = \nabla_g v(g)$ Where, $v(g)$ are the link volumes obtained by the assignment procedure. Following Tobin and Friesz (1988), who proved the Jacobian, J, to be unique, the computation of these derivatives will in theory require two matrix inversions. By solving a set of quadratic problems the computations become reasonable and with some minor additional calculations $\nabla F(g)$ is attained. It is noted that using a 'projected gradient' determined with a minimal computational effort, is equivalent to the approach of Spiess (1990). The method of Spiess is in essence equivalent to making simplifying assumptions in the quadratic problems. The search directions based on the gradient can be improved by using second order information obtained from the Hessian of F, the matrix of second order derivates. Computing the diagonal of this Hessian can be performed through calculations similar to those for obtaining the Jacobian J. Scaling the gradient with the inverse of this diagonal matrix then result in a method using second order information.

In the algorithm, equilibrium assignment sub-problems are repeatedly solved and since path-flow variables are used each assignment problem can be solved by reoptimizing the previous one. The traffic assignment procedure uses a code by Larsson and Patriksson (1992). However, the requirements for memory space are considerable. In theory the descent algorithm can be shown to converge but no guarantee can be given that the limit point is a global or even a local minimum. Since the algorithm is a descent method Drissi-Katouni and Lundgren however notes that the objective function is always improved from the starting point (the starting matrix), which leads to some sufficiently improved OD matrix.

The approach is implemented and applied to test problems from the city of Hull, Canada with 798 links and 146 OD pairs with non zero demand for traffic. The upper level objective function only includes link flow variables and a least squares function is employed:

$$Z(g) = \frac{1}{2}\sum_{a\in A}(v_a(g) - \hat{v_a})^2 \qquad \mathbf{(2.34)}$$

of the methods but rather on the quality of the search directions in the solution process. In the applications traffic counts were assumed to be available for 5 % of the total number of links. The results of methods using various directions are all reasonably accurate and no large differences were visible in the final OD matrix estimated. The more elaborate search directions (using second order information) do render a faster objective function improvement than the gradient based search directions.

<u>Florian and Chen (1993)</u>

A coordinate descent method for the bilevel OD matrix adjustment problem,

The OD matrix estimation problem is formulated as a bilevel program with user-equilibrium assignment on the lower level. A practical algorithm to solve the OD matrix estimation problem when the flows in the network are distributed according to the user optimal principle is developed.

It is argued that using path-flow variables in large networks is impractical. A Gauss-Seidel type method is proposed that does not use path information explicitly and iterates between the upper and lower level problems. The path dependent information is used implicitly

and it is devised how the algorithm may be implemented in the EMME/2 (1990) software package. In each iteration at most three equilibrium assignment problems are solved. The algorithm is applied to two test problems. The first is a small network taken from Nguyen and Dupuis (1983) with 20 links and 4 OD pairs. The results from an application to a larger Winnipeg network is termed 'very encouraging' but it is stressed that the method indeed is heuristic and very dependent on the correct structure of the initial (target) OD matrix relative to the observed traffic counts.

Spiess (1990)

A gradient approach for the OD Matrix adjustment problem,

The problem of Spiess has a bilevel structure and uses a steepest descent method for the solution of the upper level problem. On the lower level, equilibrium assignment is assumed to hold. The optimization problem is the following:

$$\min Z(g) = \frac{1}{2}\sum_{a\in\hat{A}}(v_a(g) - \hat{v}_a)^2, \qquad (\mathbf{2.35})$$

$$s.t. v = assign\ (g)$$

This is the same problem as considered by Drissi-Katouni and Lundgren (1992). The approach of Spiess uses a gradient based on the relative change in the demand, and can be written as:

$$g_i^{l+1} = \begin{cases} \hat{g}_i & l = 0 \\ g_i^l(1 - \lambda^l \left[\dfrac{\delta Z(g)}{\delta g_i}\right]_{g_i^l} & l = 1,2,3,\dots \end{cases} \qquad \mathbf{(2.36)}$$

An expression for the gradient $\frac{\delta Z(g)}{\delta g_i}$ is easily determined if proportional assignment is assumed to hold locally within an iteration. An explicit expression for the optimal step length can also be determined.

It is shown how this gradient method can be implemented using the standard version of the EMME/2 (1990) transportation planning software. Each iteration of the gradient method only corresponds to one or two equilibrium assignments plus some minor additional calculations. Even though this represents a substantial e ort, it is claimed to be computationally reasonable for the largest networks that can be handled within the EMME/2 system. The method is available as a macro DEMADJ.

The method has successfully been applied to several large scale problems in Switzerland, Sweden and Finland. The networks of Sweden and Finland are both national networks while the two problems of Switzerland are urban applications of Bern and Basel. The largest network has 469 traffic zones and 12 476 transportation links. The method has been found to perform satisfactorily.

2.3 Use of Ubiquitous Data in Transportation

Ubiquitous sources (e.g. GPS trajectories, WLAN access records, mobile phone call records, etc.) has now become a good source of raw data in transportation application. These data sources are increasingly being used in travel behavior modeling because of several advantages. Firstly, compared to other data sources, it is relatively easy to get panel data over long durations using these sources which makes it possible to model travel behavior of individuals at relatively lower costs. Gonzalez et al. (2008) and Song et al. (2010) shows such thing in their paper. This is particularly useful since there have been evidences that there are often little repetition of tours from one day to the next and therefore models estimated with cross-sectional data can lead to serious errors in prediction (Stopher et al., 2011). Secondly, unlike traditional travel diaries, ubiquitous sources do not have reporting errors and fatigue effects and have smaller non-response bias e.g Schonfelder et al. (2002), Wolf et al. (2003), Forrest et al. (2005), Stopher et al. (2011), Bricka et al. (2009), Giaimo et al. (2010). However, extraction of trajectories from these ubiquitous data sources is extremely challenging because of many issues: the coarse spatial precisions, missing data points, absence of place labels (e.g. home, workplace, shops) for identifying trip purposes, privacy issues and potential of sampling bias to name a few.

Some of the major uses of ubiquitous data are:

- Human travel pattern visualization e.g. Phithakkitnukoon et al. (2010), Phithakkitnukoon et al. (2011) and Reades et al. (2009).
- Aggregate level analyses e.g. Gonzalez et al. (2008), Song et al. (2010), Simini et al. (2012), Candia et al. (2008) and Sevtsuk et al. (2010)
- Route choice modeling e.g. Wolf et al. (1999), Jan et al. (2000), Li et al. (2005), Frignani et al. (2010), Spissu el el. (2011) Schlaich et al. (2010) and Yalamanchili et al. (1999)
- Origin-destination flows estimation using mobile phone details data e.g. Friedrich et al. (2010) and Calabrese et al. (2011)
- Traffic model calibration e.g. Barth et al. (1996), Gurusinghe et al. (2002), Wang et al. (2005), Wang et al. (2006) and Di et al. (2010)

2.4 Summary

Though lots of works has been done regarding using ubiquitous data in transportation application. Mobile phone continuous data has been also used to determine the OD matrix. But using CDR data is a new attempt to determine the OD matrix.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The traditional methodology to estimate OD matrix is to collect questionnaire survey, license plate reader data and/or traffic counts and combine them. But the reliability of questionnaire data, time needed for survey and the cost of survey has led transportation engineers to think about alternate data sources for the OD matrix estimation. In this thesis mobile phone call details records (CDR) has been used as the primary source of data for the OD matrix estimation. A proportional OD matrix (termed as ‘Seed OD’ in this thesis) is generated using the CDR data. Video data count data is then used to determine the ‘scaling factor’ and determine the actual OD matrix. The methodology used for OD estimation is described in the following sections.

3.2 Framework

The mobile call data contains Latitude and Longitude of Base Transceiver Station (BTS) to which the call is generated. The overall OD estimation work is done in three steps. These are:

1. Associating BTS of GSM network with key nodes of the transportation network
2. Determining the flows among the nodes based on the CDR
3. Finding the scaling factor and determining the actual OD matrix

The first two steps are for determining the Seed OD matrix and the third one is to estimate the actual OD matrix.

3.2.1. Assigning BTS to nodes

To use mobile phone data for identifying mobile user's movements, we need to assign each BTS to a definitive node of the road network. The assigning of the BTS to a node is done depending on the following assumptions:

1. **Defining tower coverage area:**

   Firstly, the tower coverage area (A) is identified. A BTS tower coverage area differs depending on:

   a. Adjacent tower location
   b. Population in the coverage area
   c. Geography (Obstacles)

   Normally in urban areas the population density is so high that the tower locations are very close to each other (around 0.5km) compared to rural area where BTS service area can be as large as 10-12 km. The tower coverage area is quite like combination of three hyperbolas shown in Figure 3.1.

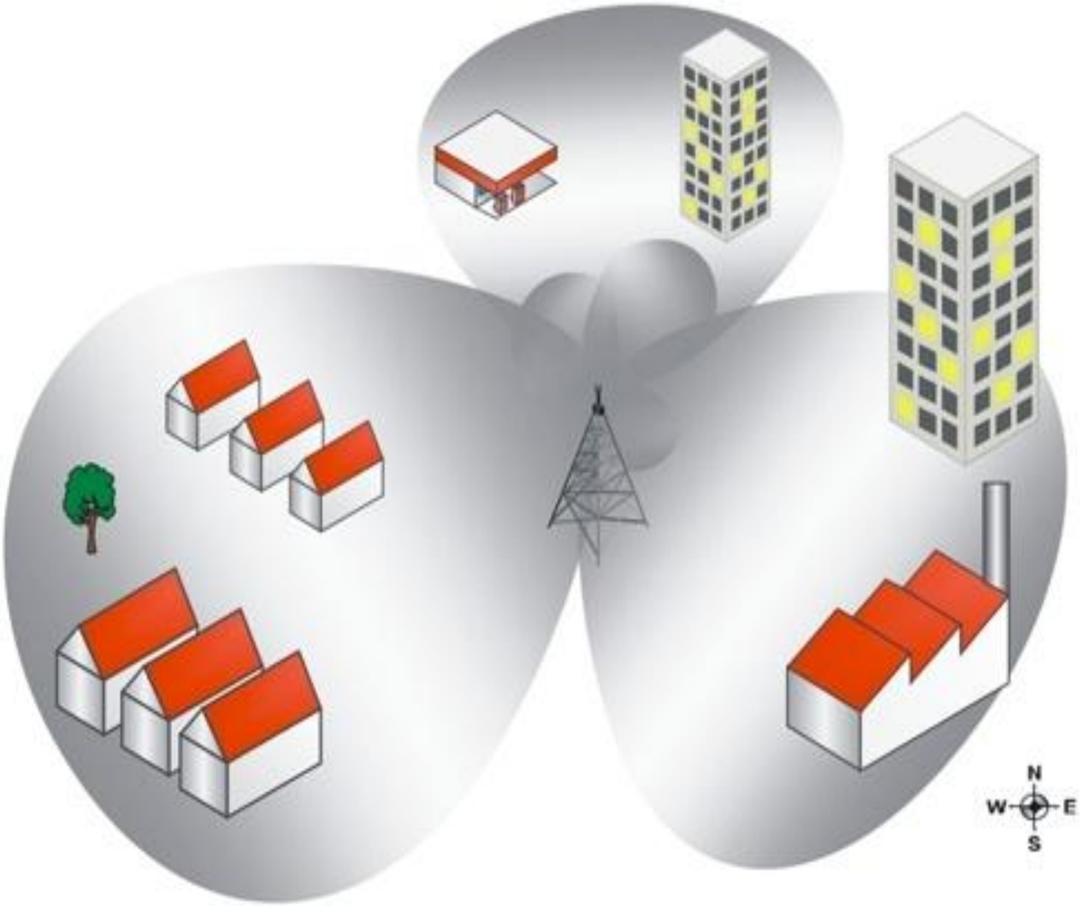


Figure 3.1: BTS coverage area (http://www.truteq.co.za/tips_gsm/)

As a mobile always connects to its nearest node and Dhaka is a densely populated urban area, it can be assumed that the area between the BTSs is equally split. That is, if the distance to the nearest BTS in a particular direction is x km, area within 0.5x km radius in that direction falls under each BTS (in the Dhaka network, x is 1km on average).

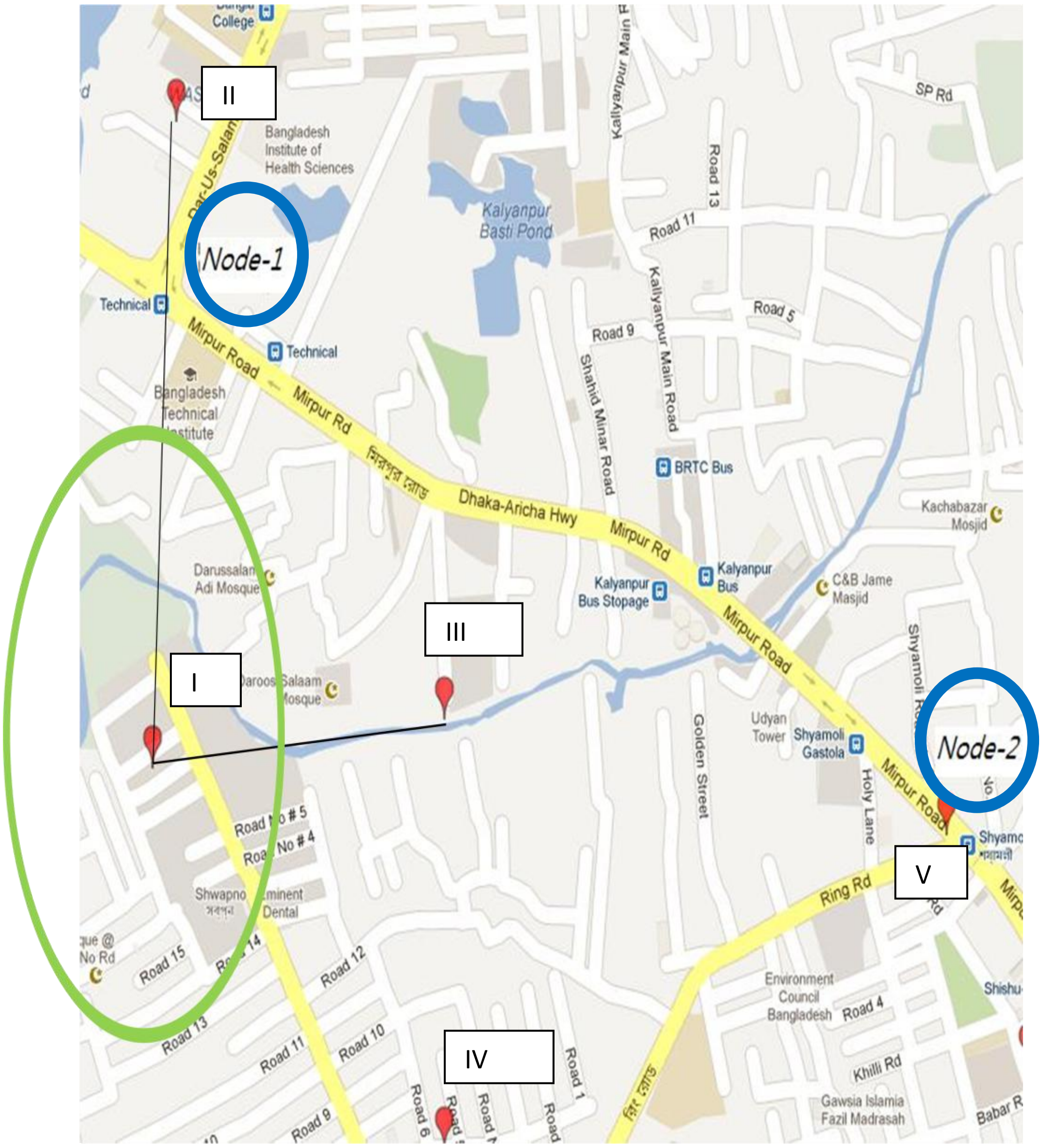


Figure 3.2: Finding the nearest road

Figure 3.2 is a representation of nodes and towers. The red points are tower locations and the nearest nodes (Node-1 and Node-2) are indicated by blue circles. For tower I (left most tower), according to step one, the coverage area of the tower is the half of the area to its nearest nodes (Node-1 and Node-2). Here it is indicated by green line.

2. **Identifying the nearest node:**

   After defining the area (A) of a tower, the nearest node under which the area serves is identified. It depends on the feeder roads and connectivity to this node.

   In Figure 3.2, in step 2 if we examine the area under the green line than it is seen that the most of the area is connected by feeder road to node 2 and some upper portion (separated by a canal) is connected to node-1.

3. **Sequence of nodes for multiple node connection:**

   If the area (A) associated with a BTS has two nodes closely located (border towers), then it is assigned to different nodes with a definite sequence based on proportion of A that feeds to each node. The sequence depends on the following assumptions:

   a. First node would be the node which serves greatest portion of A
   b. Similarly the second node is the one which serves the second highest portion of A
   c. The third node is the one which serves the third highest portion of A
   d. Three nodes will be selected in this manner

   In Figure 3.2, the selected tower is assigned to two nodes. As the major portion is connected to node 2 and the rest portion is connected to 1 so the first node is 2 and the second node is 1.

4. **Finding a user maximum call positions:**

   After assigning each tower to different nodes we will get a refined data which will give information that refer each call is generated from a node or multiple (two/three) nodes. To eliminate the multiple node entries of a call with a single node the user wise call data is filtered out. Then for a user (U) the call records which have single node assigned are separated. After that the total number of calls per node is calculated. Finally the nodes are ranked in decreasing order according to the number of calls made in each node.

Table 3.1: CDR data sample

UNIQ_ID of User	Call No	Tower ID	Node
AAH03JABeAANU1fAiP	1	1	1
AAH03JABeAANU1fAiP	2	2	2
AAH03JABeAANU1fAiP	3	3	2 Or 3
AAH03JABeAANU1fAiP	4	3	2 Or 3
AAH03JABeAANU1fAiP	5	2	2
AAH03JABeAANU1fAiP	6	4	2
AAH03JABeAANU1fAiP	7	1	1
AAH03JABeAANU1fAiP	8	2	2

If Table 3.1 represents the CDR data of a user then it is seen that it has six single node entries. If we ranked it with the number of calls made than we will get that maximum call is made from node 2 and the 2nd maximum is node 1.

5. **Defining the node for multiple node entry:**
   For the calls of user U which has several nodes identification (1st, 2nd, 3rd nodes) take the nodes which nodes has higher rank. If the nodes do not fall in any rank then the 1st node is taken.
   In Table 3.1 two entries (call no 3 and 4) have multiple entries having entries 2 or 3. If we look at the node ranking of that user than we will get that the maximum call is made from node 2. So node 2 can be selected from the multiple entries 2 or 3.

A flow chart describing the details of assigning towers to nodes is presented in Figure 3.3

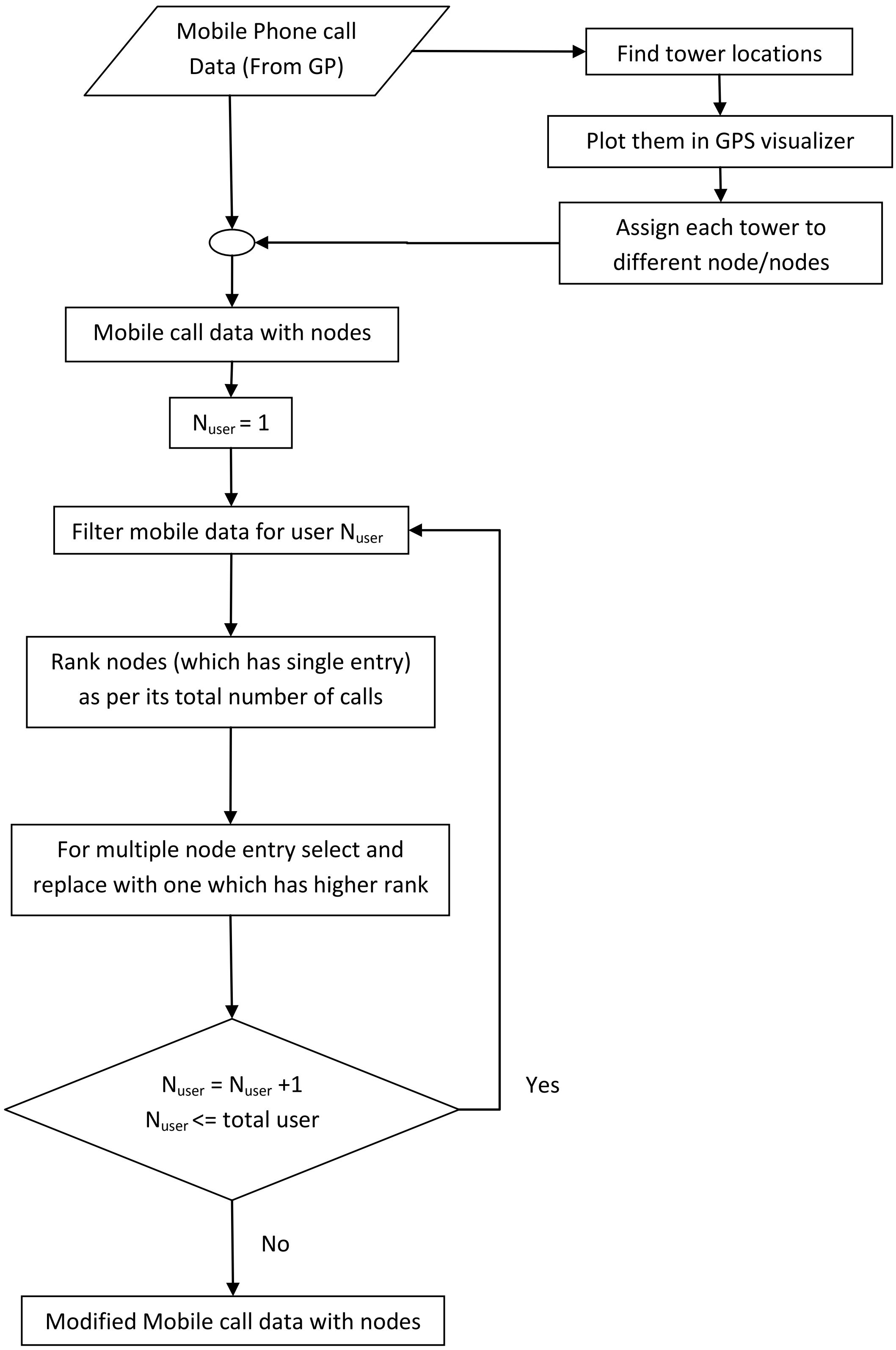


Figure 3.3 Frameworks for Assigning Towers to Nodes

3.2.2 Finding flow between nodes

From the previous section it is determined that each call is generated from a definite node. So when a user makes a call his position is determined by the corresponding node number. If the next call of the same user refers to a different node then it can be said that the user had made a trip from the previous node to this node. Thus a origin-destination pair is noted. If the same user made another moves in the remaining time of the day those OD pairs are also stored. This process is repeated for all users. If a user call from same node all aver the day then it is assumed that no movement is made by that user. They step wise procedure is described below:

1. For user U and day D find its first node (N_0) of the day where the first call was happened.
2. Then search the different node (N_1) of the user in the later time of the day. Save ($N_0 – N_1$)
3. Follow the previous steps until the end of the day
4. Save ($N_0 – N_1$, N_1 - N_2, N_2 -N_3 …….etc)

Following the above procedure a proportional OD matrix is generated (termed as a Seed OD matrix). A flow chart describing the details of finding the flow between nodes is presented in Figure 3.4

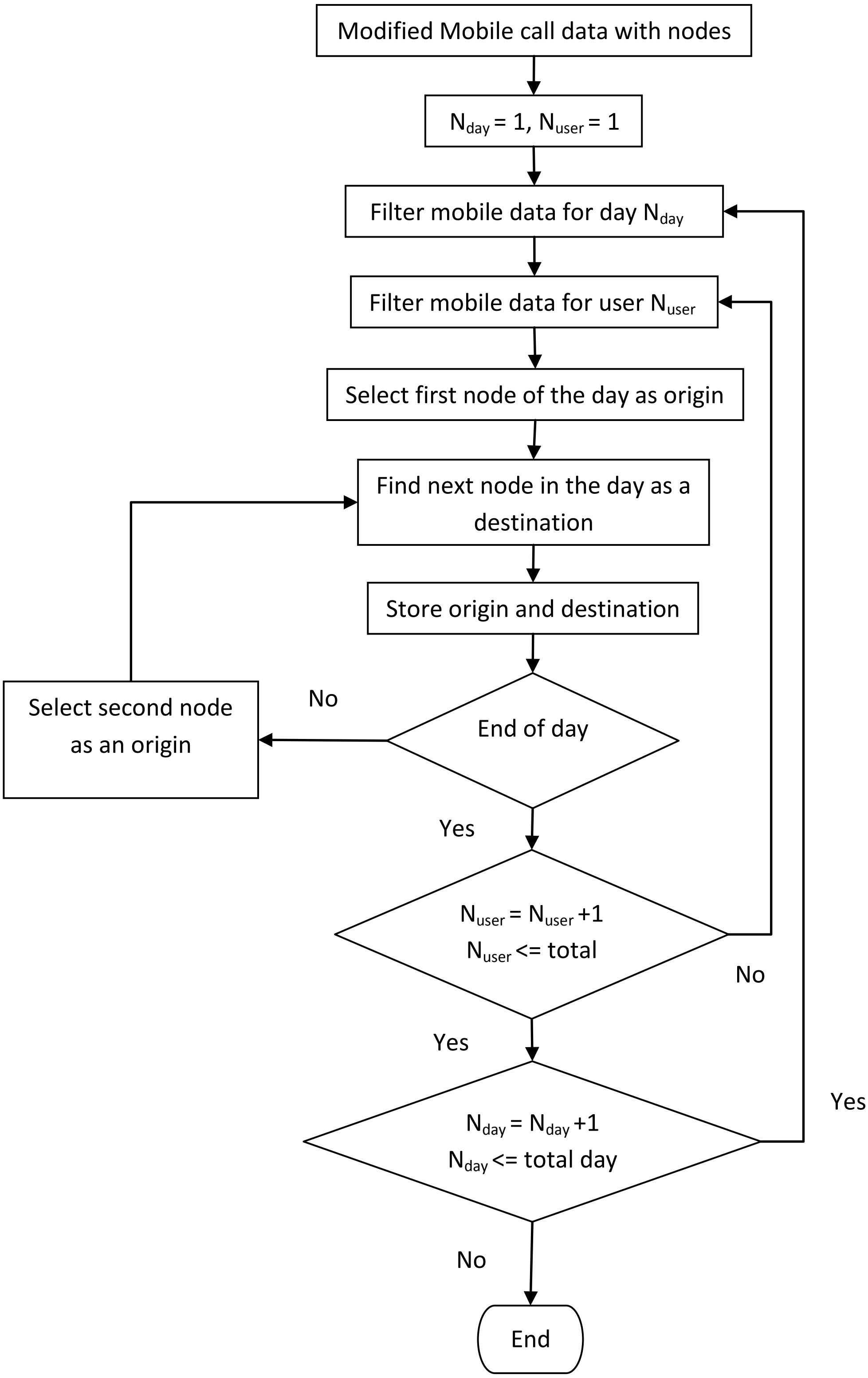


Figure 3.4 Flowchart for finding flow between nodes

3.2.3 Finding actual OD matrix

Actual OD matrix is generated from the Seed OD matrix. The video data is collected from selected locations of the road network are used to count the traffic of that road. The Seed OD matrix is used as an input to a traffic simulator to simulate the traffic scenario of the same road network.

Let,

Traffic count of link i of the road network from video data is N^{i}_{actual}

Traffic count of link i of the road network from simulation is $N^{i}_{simulated}$

Now a scaling factor (β) is determined such a way that the difference between those two observation can be minimized. So the objective function is

$$minimize, Z = \sum_{i=1}^{n}(N^{i}_{actua\ l} - N^{i}_{simulated})^2 \qquad \text{Eqn. 3.1}$$

Such that, $OD_{simulator} = \sum_{j=1}^{m} \beta_j * OD_{seed,j}$

Where,

n	=	Total number of links which count data are available
m	=	Total number of scaling factors
OD_{seed}	=	OD matrix found from mobile data
$OD_{simulator}$	=	OD matrix used in the simulation

The optimization algorithm used in this thesis is BOX algorithm.

3.3 Summary

The main advantage of this methodology is that it makes the data collection process easier and more reliable. As mobile phone is being used by most of the people in a city that we can easily get the CDR data from the mobile company. Video data collection in some foot over-bridges of the network is also convenient. Now-a-days there are software which can analyze video traffic data to give the traffic count. So this methodology would be an effective way for OD estimation.

CHAPTER FOUR

DATA COLLECTION

4.1 Introduction

The main data of this thesis is mobile phone call records (CDR) data. Supplementary data is collected from thirteen different location of Dhaka city road network. The overall data collection processes are described in the following sections.

4.2 Mobile Phone Call data Collection

Mobile phone call records data is collected from the leading mobile phone company of Bangladesh, Grameenphone (GP) Ltd. (Grameenphone, 2012), through a memorandum of understanding between Grameenphone Ltd. and Bangladesh University of Engineering and Technology (BUET). According to this memorandum, anonymized CDR data of cellular phones (without personal data) of Dhaka city for the duration on one month is provided. Call Detail Record (CDR) data include:

- An unique ID of each individual,
- Location of the base transceiver station (BTS) where the activity (mobile call) occurred (Latitude and Longitude of BTS), and
- Time of the activity occurrence (Date and time)
- Duration of the activity

As per Memorandum, GP gives CDR data for all of its users within the Dhaka city over the duration of one month starting from 19 June, 2012 to 18 July, 2012.

GrameenPhone uses GSM (Global System for Mobile Communications) Network system. To understand the CDR data collection process we need to know about some parts of GSM network. The relevant parts are described below.

GSM Networks Components:

The GSM network is divided into two systems (GSM, 2012). They are:

1. Switching System (SS)
2. Base Station System (BSS)

Each of these systems are comprised of a number of functional units which are individual components of the mobile network. In addition, as with all telecommunications networks, GSM networks are operated, maintained and managed from computerized centers. The main parts of GSM network is shown in Figure 4.1

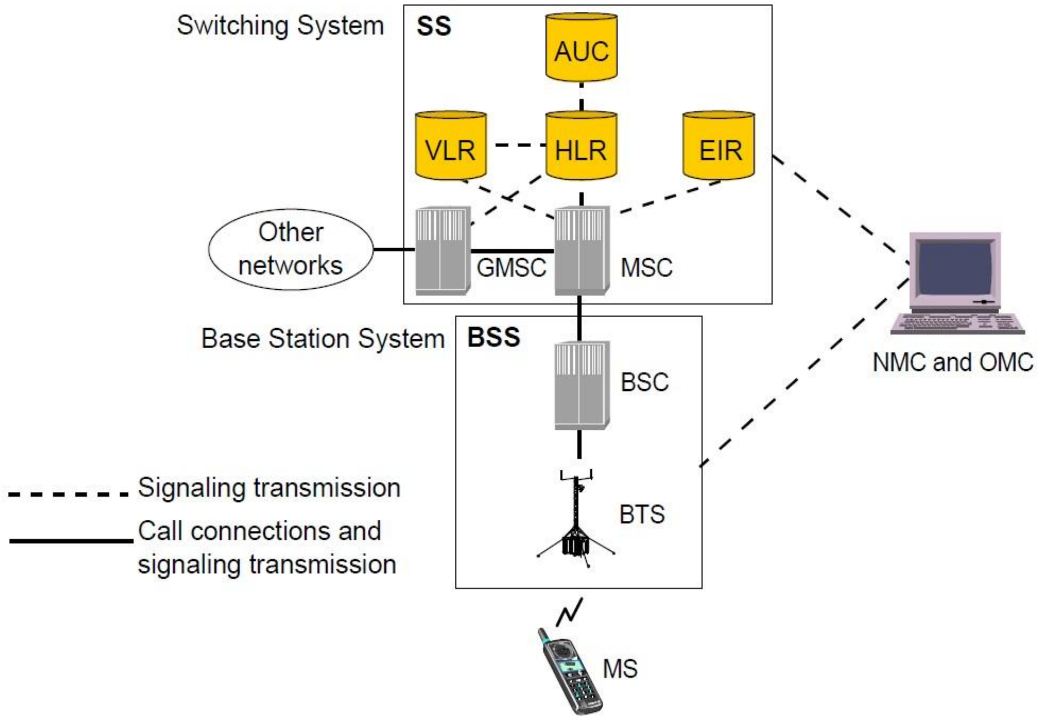


Figure 4.1 GSM network systems

The SS is responsible for performing call processing and subscriber related functions. It includes the following functional units:

1. Mobile services Switching Center (MSC)
2. Home Location Register (HLR)
3. Visitor Location Register (VLR)

4. AUthentication Center (AUC)
5. Equipment Identity Register (EIR)

The BSS performs all the radio-related functions. The BSS is comprised of the following functional units:

1. Base Station Controller (BSC)
2. Base Transceiver Station (BTS)

The OMC performs all the operation and maintenance tasks for the network such as monitoring network traffic and network alarms. The OMC has access to both the SS and the BSS. MSs do not belong to any of these systems.

For this thesis work it is not needed to learn about all the terms. So in the following sections the related term and their responsibilities for collecting CDR data are described.

MS = Mobile Station (or Mobile Phone)

The mobile phone we use to communicate with others is termed as a mobile station. It has two components:

1. A mobile terminal (Mobile Set)
2. A Subscriber Identity Module (SIM)

When a call is made from a MS then it connects to its nearest BTS.

Base Transceiver Station (BTS)

The BTS controls the radio interface to the MS. The BTS comprises the radio equipment such as transceivers and antennas which are needed to serve each cell in the network. A group of BTSs are controlled by a BSC. CDR data contains Latitude and Longitude of those BTS under which the call is made.

Base Station Controller (BSC)

The BSC manages all the radio-related functions of a GSM network. It is a high capacity switch that provides functions such as MS handover, radio channel assignment and the collection of cell configuration data. A number of BSCs may be controlled by each MSC.

Mobile services Switching Center (MSC)

The MSC performs the telephony switching functions for the mobile network. It controls calls to and from other telephony and data systems, such as the Public Switched Telephone Network (PSTN), Integrated Services Digital Network (ISDN), public data networks, private networks and other mobile networks. The important thing is that the CDR are stored in the MSC and leter the data is extracted by revenue management team for billing purposes.

A sample of CDR data of a originating call is shown in Appendix A with describing (in green colours) the entities which are needed in this thesis.

From this type of CDR data the required data is extracted for one month duration. The data contains anonymous mobile number, date of call, time of call, call duration, latitude and longitude of the BTS under which the call was served.

4.3 Video Data Collection

Video data is the supplementary source of data in this thesis. Video data is collected in thirteen different location of Dhaka city network. The locations are selected on the following considerations:

- Covering major roads (links) of Dhaka city
- All locations are spatially selected so that all major OD flows are covered
- Availability of foot over bridge for locating the video camera
- Avoiding excessive jam area or intersection

Figure 4.1 shows the data collection points in Dhaka city road network.

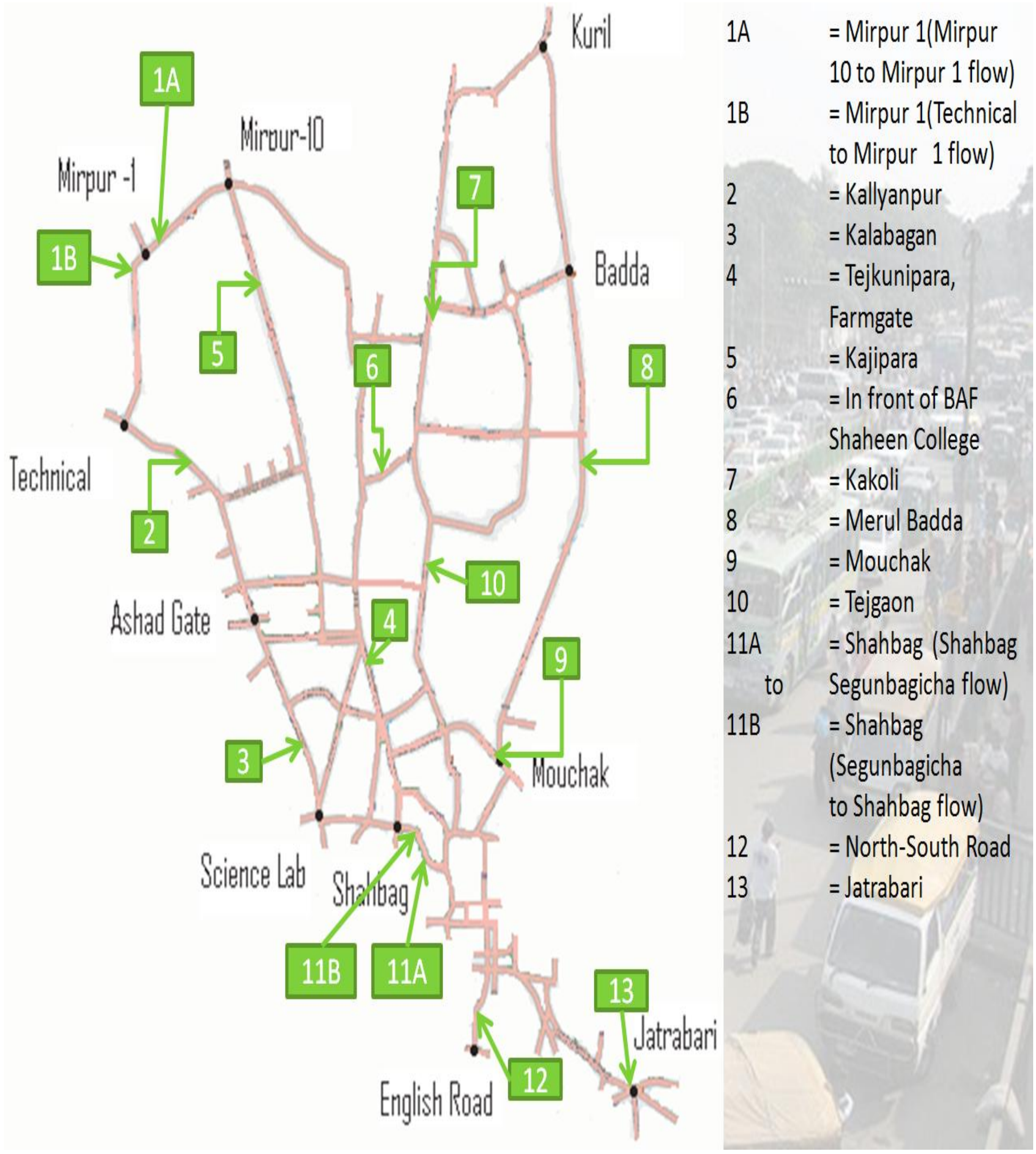


Figure 4.2 Video data collection locations

At those locations video data is collected for three days having eight hour each day. The time of data collection for each day was:

- 8.00 am to 12.00pm
- 3.00 pm to 7.00pm

This time was taken such that the video can cover the pick hour traffic flow. As there are two peak hour in the Dhaka (one is the home to working place trip in morning and another is working place to home trip in the evening) so the time is selected four hour in the morning and four hour in the evening.

A brief description of each point along with picture is described below.

4.3.1 Mirpur 1 (Mirpur 10 to Mirpur 1 Flow) [1A]

In this location the foot over-bridge located at the junction is used for data collection. This location covers traffic coming towards mirpur1 from mirpur10.

Figure 4.3 Mirpur 1 (Mirpur 10 to Mirpur 1 Flow)

4.3.2 Mirpur 1 (Technical to Mirpur 1 Flow) [1B]

In this location also the foot over-bridge located at the junction is used for data collection. This location covers traffic coming towards mirpur1 from Technical.

Figure 4.4 1B Mirpur 1 (Technical to Mirpur 1 Flow)

4.3.3 Kallyanpur [2]

In kallyanpur the video data collection location is just in-front of the bus stand. Here video data is collected in both directions i.e. one from shamoli and one from technical.

Figure 4.5 Kallyanpur

4.3.4 Kalabagan [3]

In Kalabagan the video data is collected in front of Kalabagan field. Two directional traffic flows one from Sukrabad and other from Sciencelab are captured here.

Figure 4.6 Kalabagan

3.3.5 Tejkunipara, Farmgate [4]

This site is located in in Tejkunipara near farmgate. Two video cameras are located in two directions. One is to capture flow from Farmgarte and another is from Kaoranbazar.

Figure 4.7 Tejkunipara, Farmgate

3.3.6 Kajipara [5]

Flow from mirpur-10 and flow from Taltola are recorded using two video cameras from foot over-bridge near Kajipara.

Figure 4.8 Kajipara

3.3.7 In front of BAF Shaheen College, Dhaka [6]

In this location traffic flow from Mohakhali and flow from Bijoysarani is captured.

Figure 4.9 In front of BAF Shaheen College, Dhaka

3.3.8 Kakoli [7]

Here two directional traffic flows are captured. One is from Mohakhali and another is from Airport.

Figure 4.10 Kakoli

3.3.9 Merul Badda [8]

Traffic flow from Badda is captured by one video camera and the flow from Rampura is captured by another video camera.

Figure 4.11 Merul Badda

3.3.10 Mouchak [9]

The two directional flows are one from Malibagh and one is from Mogbazar.

Figure 4.12 Mouchak

3.3.11 Tejgaon [10]

The data collection point of Tejgaon is in front of Dhaka Polytechnic Institute. Here the two directions are: one from Sat Rastar More and another from Mohakhali

Figure 4.13 Tejgaon

3.3.12 Shahbag (Shahbag to Segunbagicha flow) [11A]

This location is situated in front of IEB. Only the flow from Shahbaag to Segunbagicha is captured here.

Figure 4.14 In In front of IEB (Shahbag to Segunbagicha flow)

3.3.13 Shahabag (Segunbagicha to Shahbag flow) [11B]

To capture Segunbagicha to Shahbag flow a video camera is placed on foot over-bridge near Dhaka Shishu-Park.

Figure 4.15 In front of shishu park(Segunbagicha to Shahbag flow)

3.3.14 North South Road [12]

To capture flow from Gulisthan and flow from English Road two video cameras are placed in North South road. This loacation is near Suritola High School.

Figure 4.16 North South Road

3.3.15 Jatrabari [13]

One of the data collection locations is in Jatrabari. Two video cameras are set here to capture two way flows. One is to capture flow from Shanir Akhra and another is from Saidabad.

Figure 4.17 Jatrabari

4.4 Summary

The main data collected from Mobile Phone Company and the supplementary data from video recording is used to find the OD matrix of Dhaka city. Due to rain the video data collection was hampered on the last day of data collection. Overall the data collection process is quite simple and easy compared to survey method.

CHAPTER FIVE

CASE STUDY

5.1 Introduction

To demonstrate the methodology proposed in chapter three, a case study has been executed. The major road network of Dhaka has been selected for this case study. The selected road network and the Seed OD matrix have been used as inputs to the simulator to find out the actual OD matrix.

5.2 Network description

As the mobile phone CDR data contains call records of Dhaka city so the Dhaka city network is selected to estimate the OD matrix.

The mobile phone data contains total 1360 BTSs around the Dhaka city. So each call will generate from those towers. To estimate OD matrix, the towers need to be assigned to nearest nodes. So depending of the towers location and coverage area total 29 nodes has been selected in the network.

Figure 5.1 shows some of the BTS in Dhaka city. As this OD matrix will be used as a Seed matrix to simulator as node to node OD matrix (as opposed to zone to zone OD matrix). So the calls has been assigned to 29 nodes where the nodes have been defined according to following assumptions:

- Has a distinct BTS under it
- Traffic enters into the city road network through the point
- Spatially separated and covers major ODs of Dhaka city

The selected 29 nodes locations are shown in Figure 5.2.

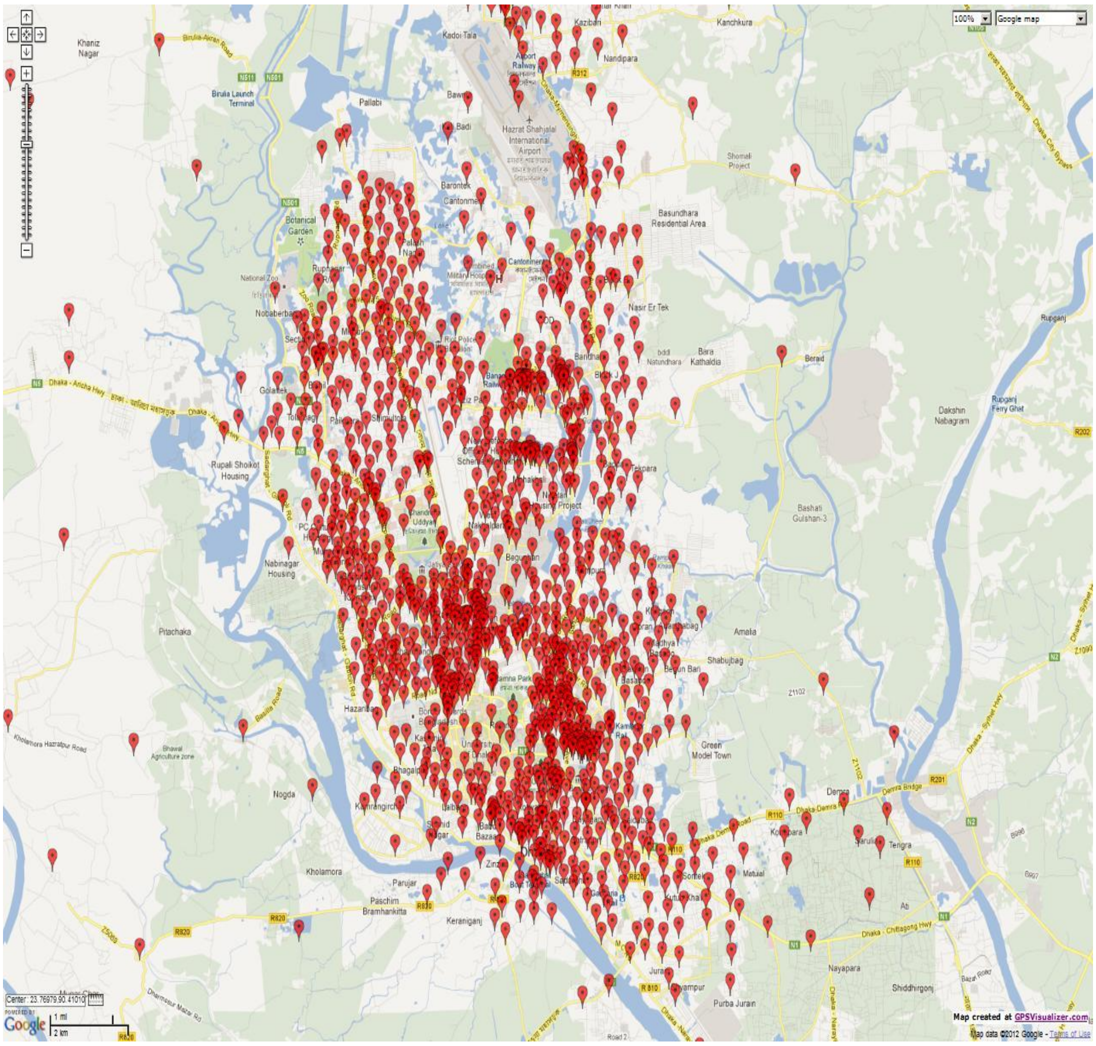


Figure 5.1 BTS locations in Dhaka city[1]

[1] This excludes some of the BTSs located in the outer periphery of Dhaka city

Figure 5.2 Dhaka road network and selected traffic nodes

The twenty nine node locations are presented in Table 5.1.

Table 5.1: Node locations

Serial No.	Name of the node	Node ID
1	English Road	1
2	Technical junction	2
3	Mirpur-1	3
4	Mirpur-10	4
5	Agargaon	6
6	Shyamoli	7
7	Asad Gate	8
8	Dhanmondi 32	9
9	Science Lab	10
10	Shahbag	11
11	KawranBazar	12
12	Farmgate	13
13	Mahakhali	14
14	Mirpur-14	15
15	Kakoli	16
16	Kuril Bishwa road	18
17	Natun Bazaar	20
18	Gulshan-2	21
19	Madhya Badda	22
20	Gulshan-1	23
21	Rampura Bridge	24
22	Mouchak	25
23	Square of seven Roads	27
24	Jatrabari	28
25	Azimpur	29
26	Motijheel	30
27	ZeroPoint	31
28	Chawk Bazar	32
29	Bashabo	33

Each node serves a definite area and contains some BTS which are under this node. There are some BTS which are termed as "Border BTS" (i.e. has the probability of being assigned to multiple nodes). A sample nodes catchment area is shown in Figure 5.3. The catchment areas of the other nodes are presented in the Appendix B.

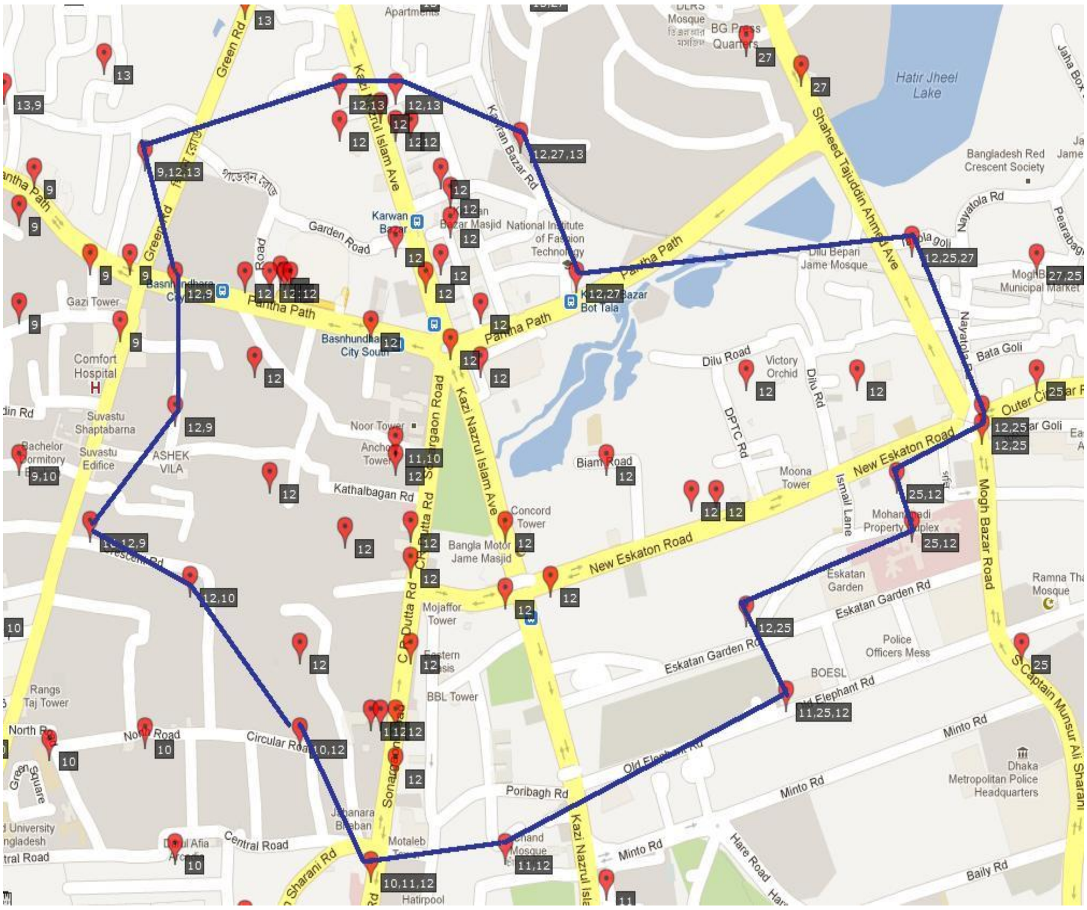


Figure 5.3 KawranBazar Node [Node ID: 12]

5.3 Analysis Process

Mobile phone data and Video data are first analyzed separately. Then they are combined to estimate the OD matrix. So the analysis process is divided into three subsections:

1. Analysis of mobile phone data
2. Analysis of video data
3. Combined analysis

5.3.1 Analysis of Mobile Phone Data

Mobile phone CDR data has been used as a key source of data in this thesis. Some features of the data are:

- Time period of data collection is from 19^{th} June, 2012 to 18^{th} July,2012.
- Total call entry is about one billion (972,558,877)
- Total number of BTS is 1359
- Total size 11.9 GB (zipped)

The node wise tower number is presented in Figure 5.4.

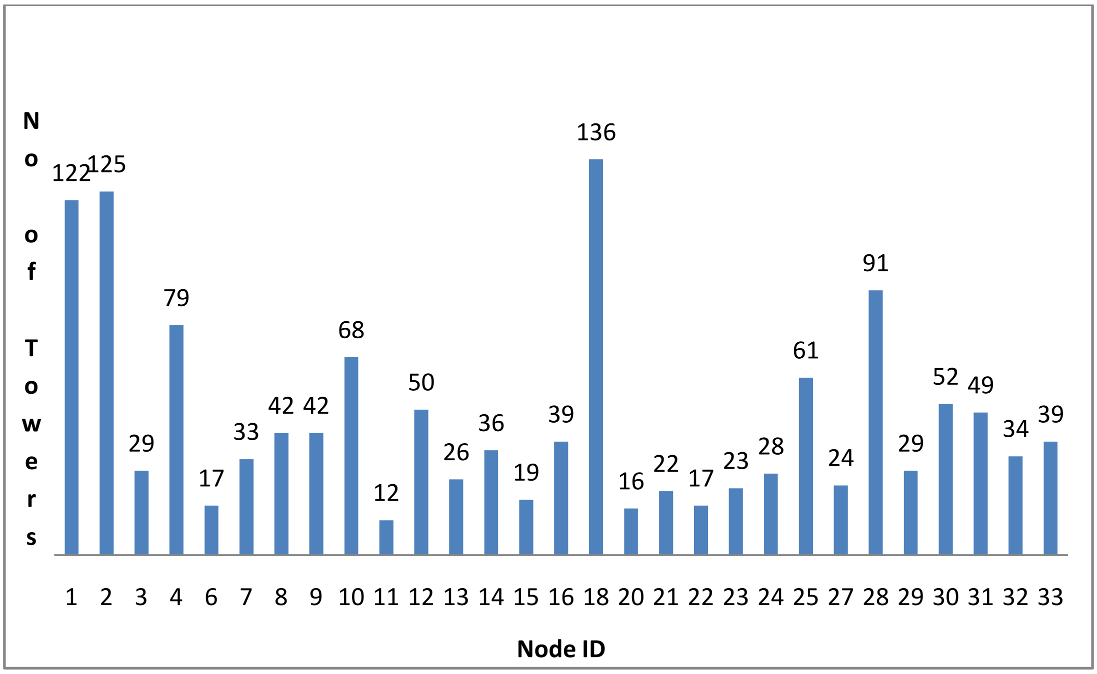


Figure 5.4: Tower numbers in each node

One of the main challenges of the thesis work was the handling of huge data. Finally the data is manipulated using MySQL database. The details code written in MySQL is presented in the Appendix C1.

Due to time constrain and data handling issues, among the 30 days mobile phone data the one day (15.7.12) data (around 34 million calls from around 1.4 million users) is filtered out from the main data table. The one day data is then sorted according to user ID.

The sorted and ordered data is then analyzed using MATLAB to determine the Seed OD matrix. The MATLAB code gives an output of 29x29 OD matrixes which is used a Seed matrix in the later analysis. The MATLAB code details are also given in Appendix C2.

The call traffic to and from each node (origins and destinations of each pair of calls) are then examined carefully to get an idea about the range of candidate scaling factors. Example of flow from and to a random node (Node 6, Shyamoli) is shown in Figures 5.5 and 5.6 respectively.

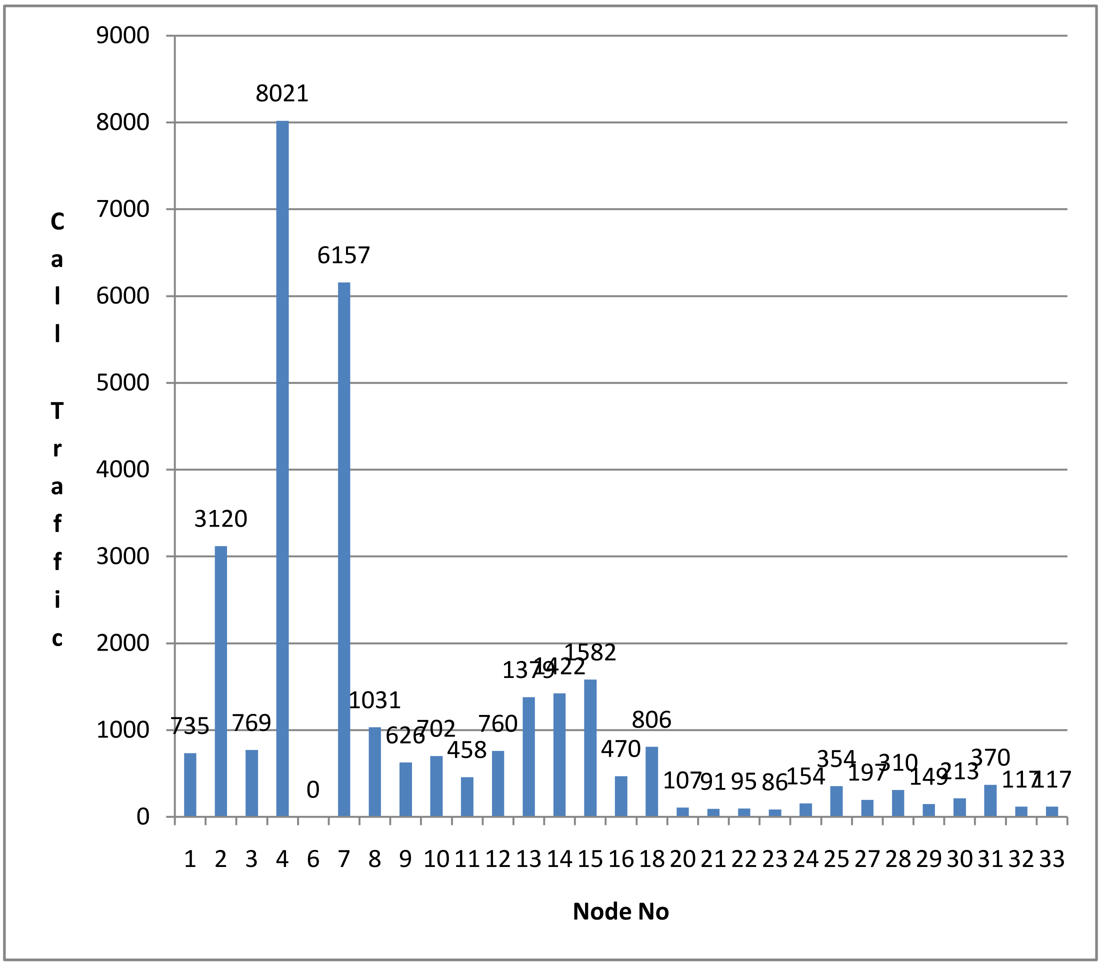


Figure 5.5 Call traffic from Node 6 (Shyamoli Node)

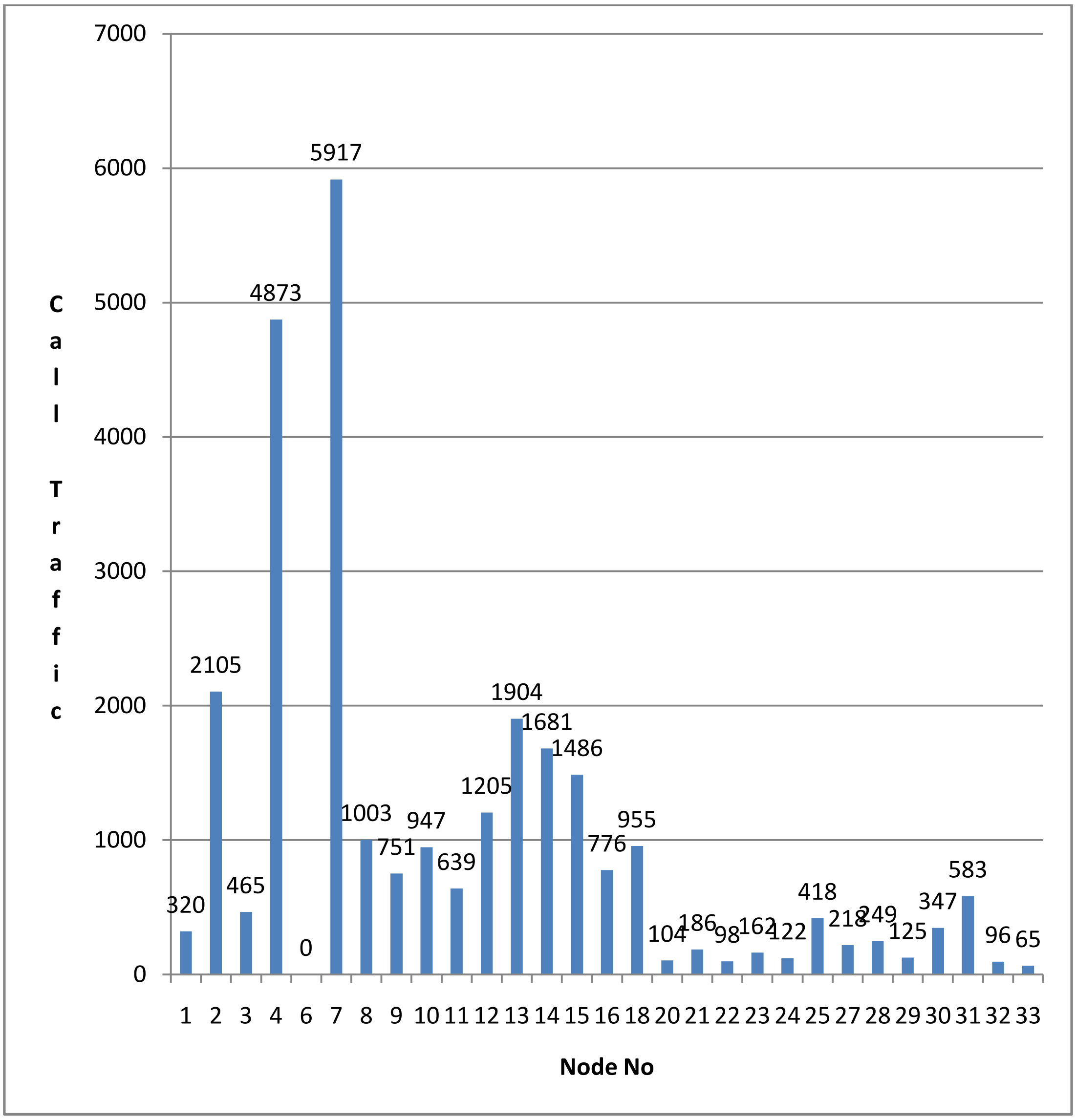


Figure 5.6 Call traffic to Node 6 (Shyamoli Node)

As observed in Figure 5.5, some flow (to node 4 and 6) is exceptionally high compared to the other nodes. This is due to the mobile phone tower switching problem of the border nodes. Tower switching problem may occur when someone calls using a tower which is assigning to node 6 but in his/her next call from the same place he/she uses a tower which is assigned to node 7. To overcome this problem two distinct scaling factors are considered in the MITSIMLab. One is for adjacent nodes and other is for other nodes.

5.3.2 Analysis of Video Data

Secondary data (Video data) is used to get the traffic count of the selected data collection points.

Software named TRAZER (Traffic AnalyZer and EnumeratoR) is used to count traffic from video recording data. TRAZER (Kritikal Solution, 2012) is a technology that helps in collecting useful traffic data catering to heterogeneous and homogeneous traffic conditions of the developed as well as developing nations.

TRAZER can give classified counts for four major type of vehicles:

1. LMV (Low motorized vehicle) i.e. car, microbus, jeep, etc.
2. HMV (High motorized vehicle) i.e. bus, truck, etc.
3. 3W (Three wheeler) i.e. CNG, auto-rickshaw, etc.
4. 2W (Two wheeler) i.e. motorcycle, cycle, rickshaw, etc.

A snapshot of TRAZER data analysis is shown in Figure 5.6.

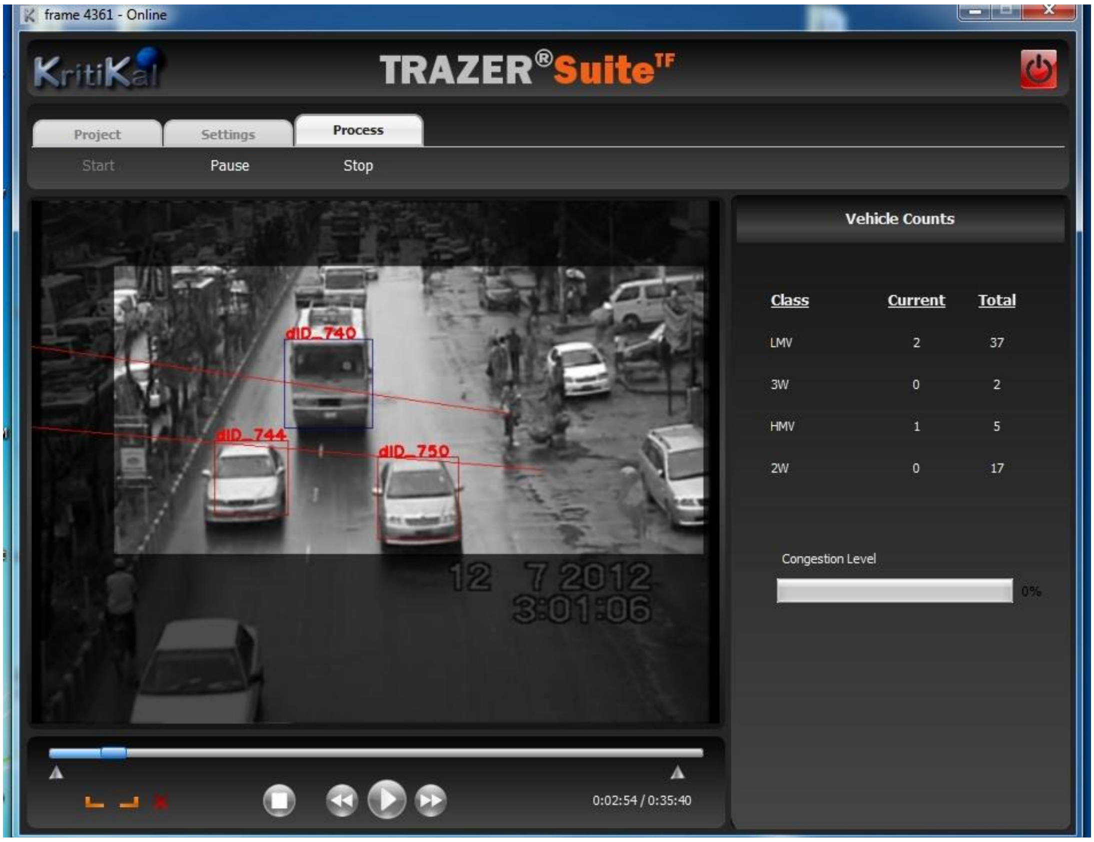


Figure 5.7 Traffic count by TRAZER software

5.3.3 Combined Analysis

After finishing the mobile phone and video data analysis, two data files have been generated:

1. A Seed OD matrix of Dhaka city
2. Vehicle count data in some selected locations (13 locations)

Taking Longitude and Latitude of different places from the Google map, a road network is coded on Road Network Editor (RNE). This is used in the simulator to mimic the Dhaka road network.

To determine the actual OD matrix combined analysis is done. This analysis is run on MITSIMLab (Yang et al., 1997), micro-simulation software. Actually the Seed OD matrix contains a portion of the actual OD matrix. So the main thing is that, a scaling factor need to be determined using optimization based algorithm.

Due to tower switching problem, there are some ‘inflated’ flows in adjacent nodes. So the flows are greater than actual in the adjacent nodes. To solve the problem, two scaling factors are considered: one is for adjacent nodes and another is for others.

So the optimization problem is to adjust the two scaling factors such a way so that the summation of the difference between the actual count and the sensor count of the MITSIMLab is minimized. A MATLAB code is written to find out the scaling factor. The code is presented in Appendix D.

A lot of factors is included the scaling factor. Some of the factors included in the scaling factor are presented below:

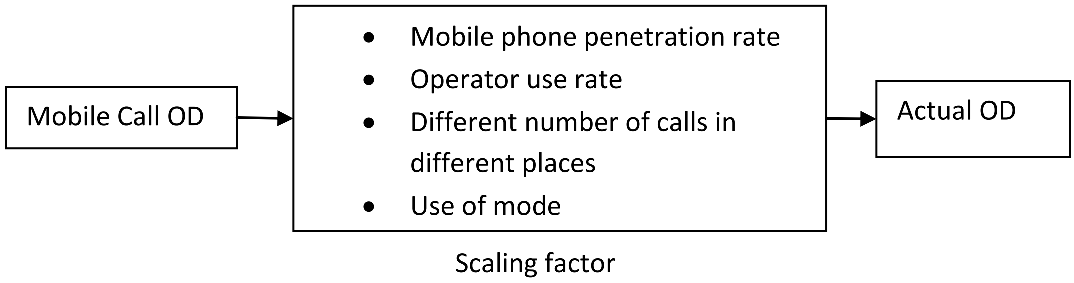


Figure 5.8 Considered factors in Scaling factor

5.4 Result

The total 29 nodes generate total 812 possible node-to-node flows in the network. Figure 5.6 and 5.7 represents the full OD matrix graphically.

Due to usage of two distinct scaling factors, the 'inflation' of flows observed in the Seed OD extracted from the CDR has been attenuated. A comparison of Seed OD and scaled OD for Node 6 has been presented in Figure 5.9 and Figure 5.10.

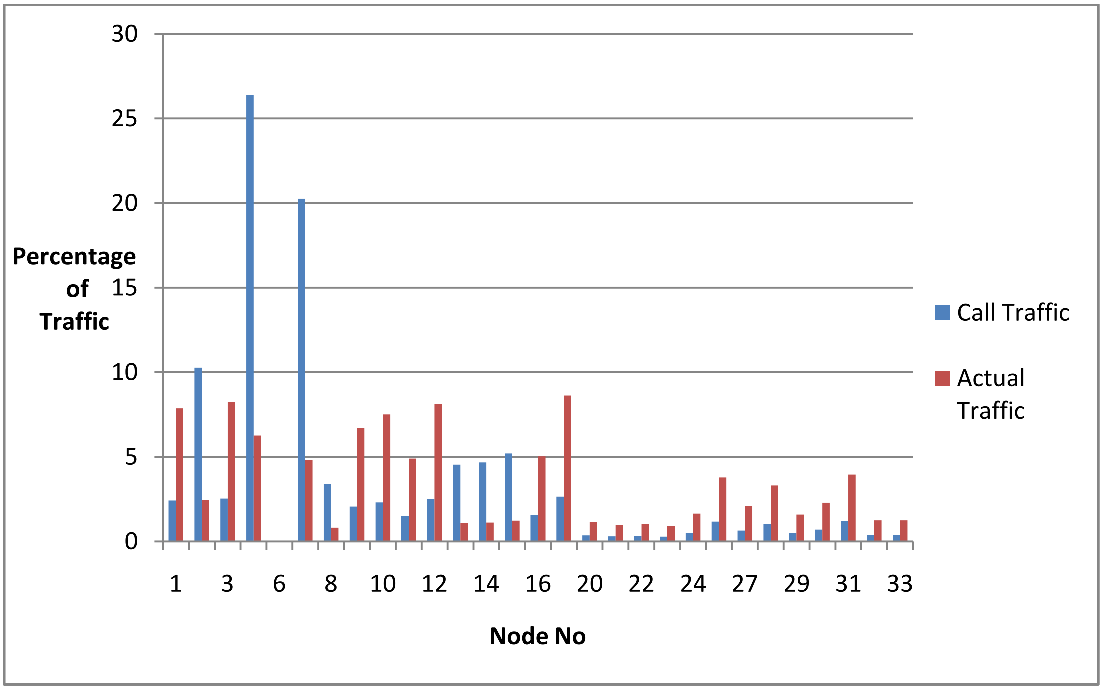


Figure 5.9 Percentage of traffic to each node going from Node 6 (Shyamoli Node)

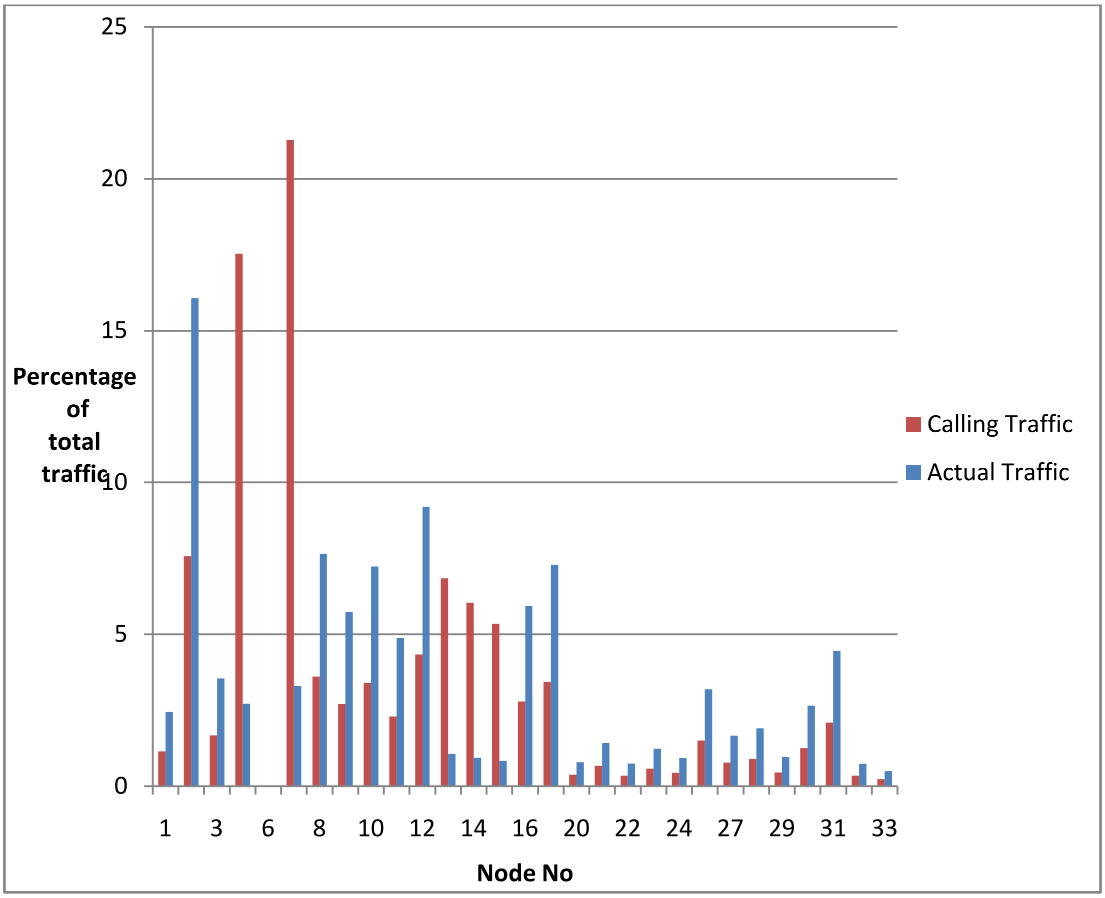


Figure 5.10 Percentage of traffic from each node coming to Node 6 (Shyamoli Node)

Figure 5.11 describes the OD flows in bar chart. The x-axis presents the origin node and the y-axis represents the percent of flows to different destinations.

Figure 5.12 represents the OD flows in the network drawn in RNE.

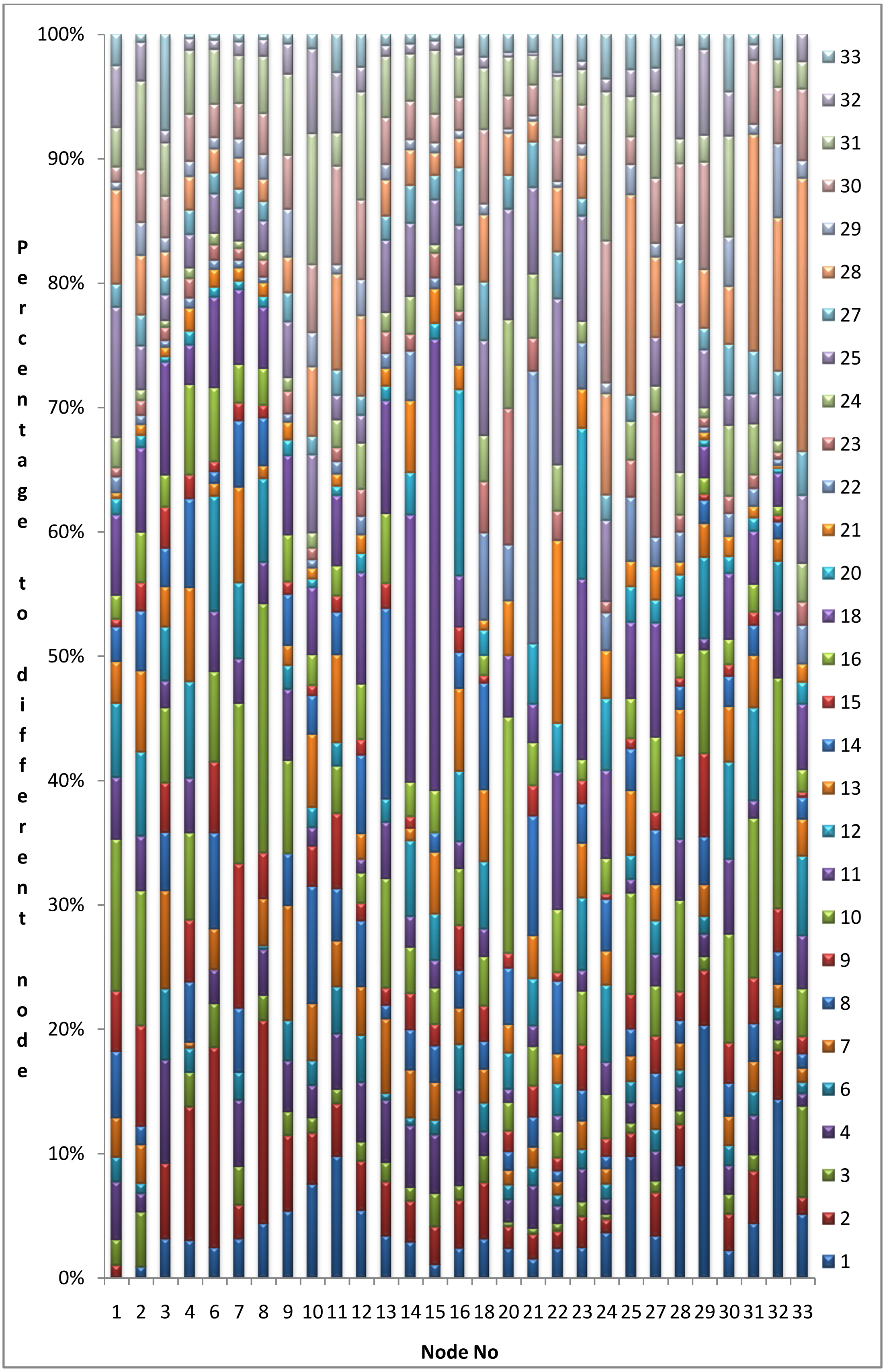


Figure 5.11 OD flow in bar charts

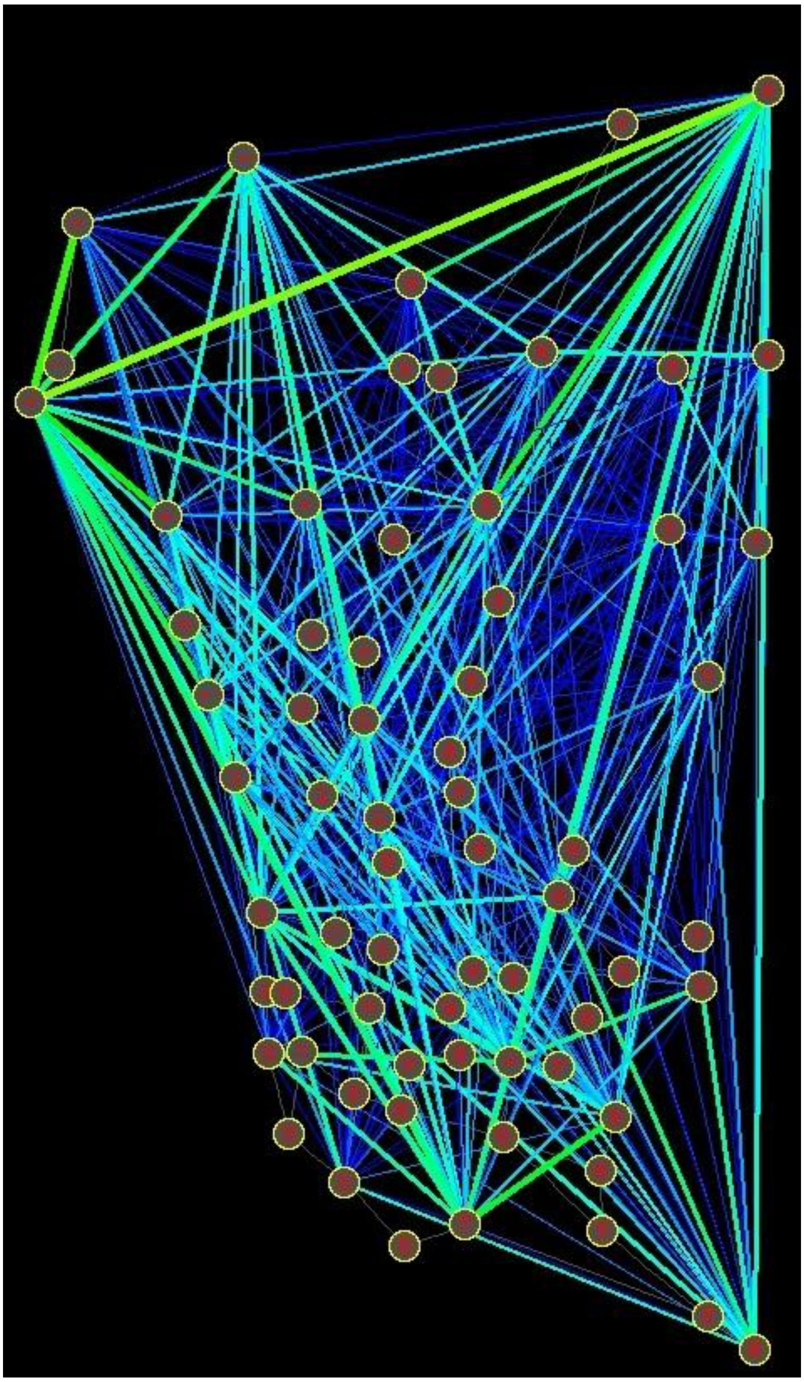


Figure 5.12 OD flows of Dhaka Network

It is tough to understand all the flows in a same graph so the node wise OD flows is described in the following sections.

5.4.1 Flow from English Road Node

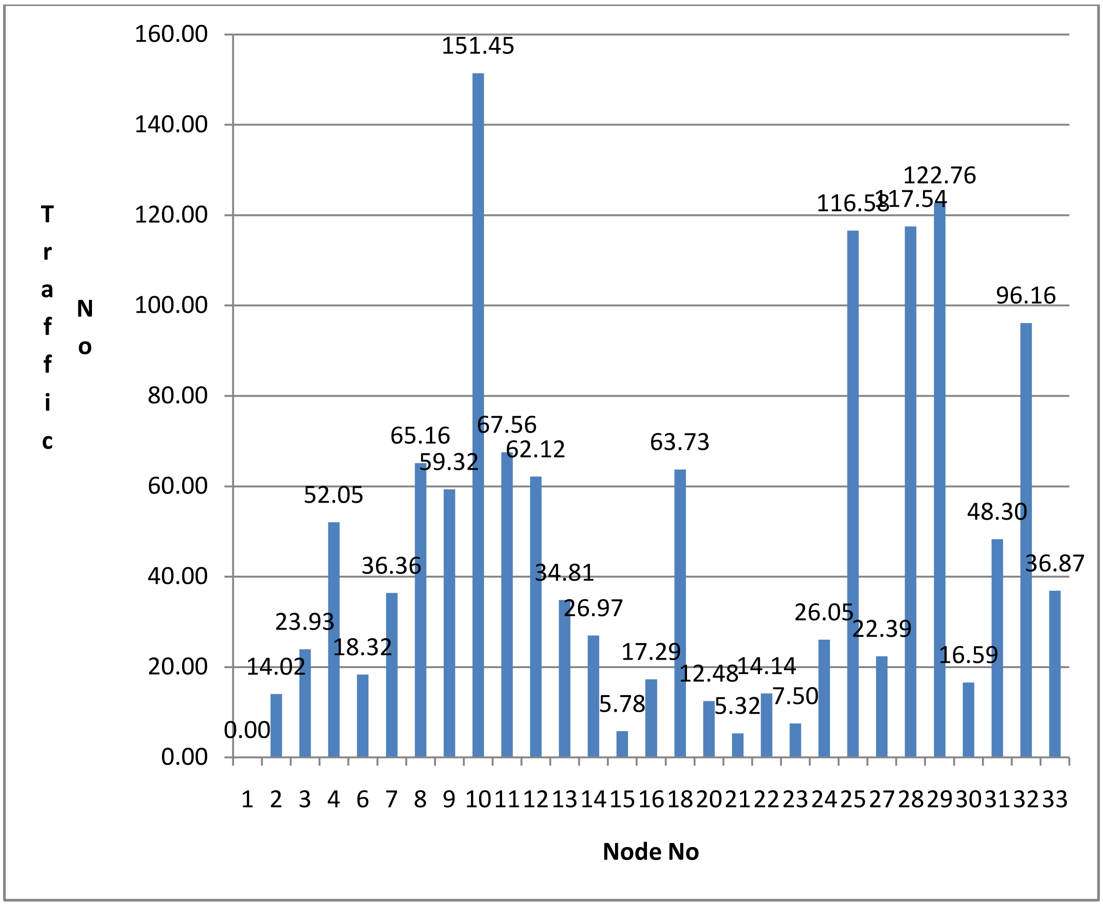


Figure 5.13 Flow from Node 1(English Road Node)

Figure 5.13 describes the flow from node 1(English Road Node) to different nodes. Main features of this node are presented in Table 5.1:

Table 5.2: Main features of the flows from node 1

Major Catchment Areas	Major Destination Nodes
• Keraniganj • Maoya area • Part of Kamrngichor • Around Sadarghat area	• Sciencelab Node • Azimpur Node • Mouchak Node • Jatrabari Node

5.4.2 Flow from Technical junction Node

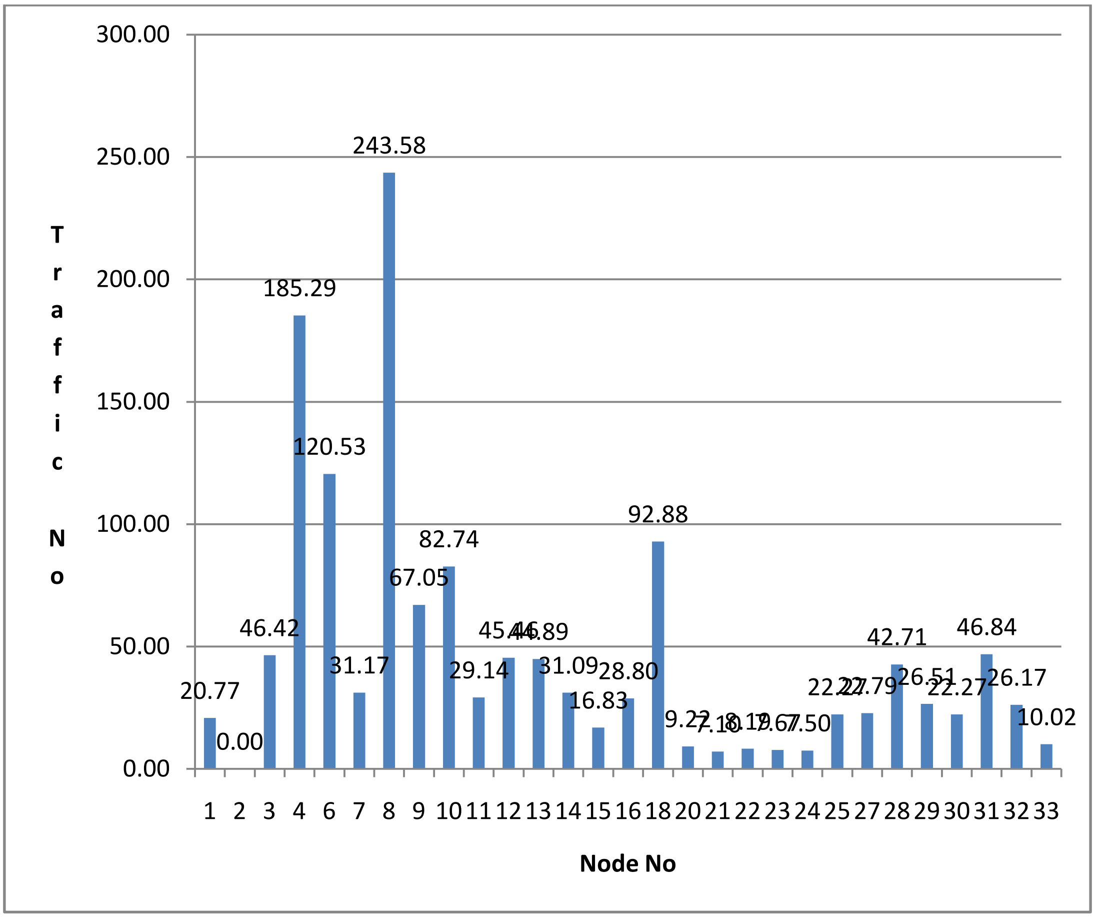


Figure 5.14 Flow from Node 2 (Technical junction Node)

Figure 5.14 describes the flow from node 2 (Technical junction Node) to different nodes. Main features of this node are presented in Table 5.2:

Table 5.3: Main features of the flows from node 2

Major Catchment Areas	Major Destination Nodes
• Savar area • Baipail • Gabtoli • Around Technical college	• Shyamoli Node • Mirpur-10 Node • Agargaon Node

5.4.3 Flow from Mirpur-1 Node

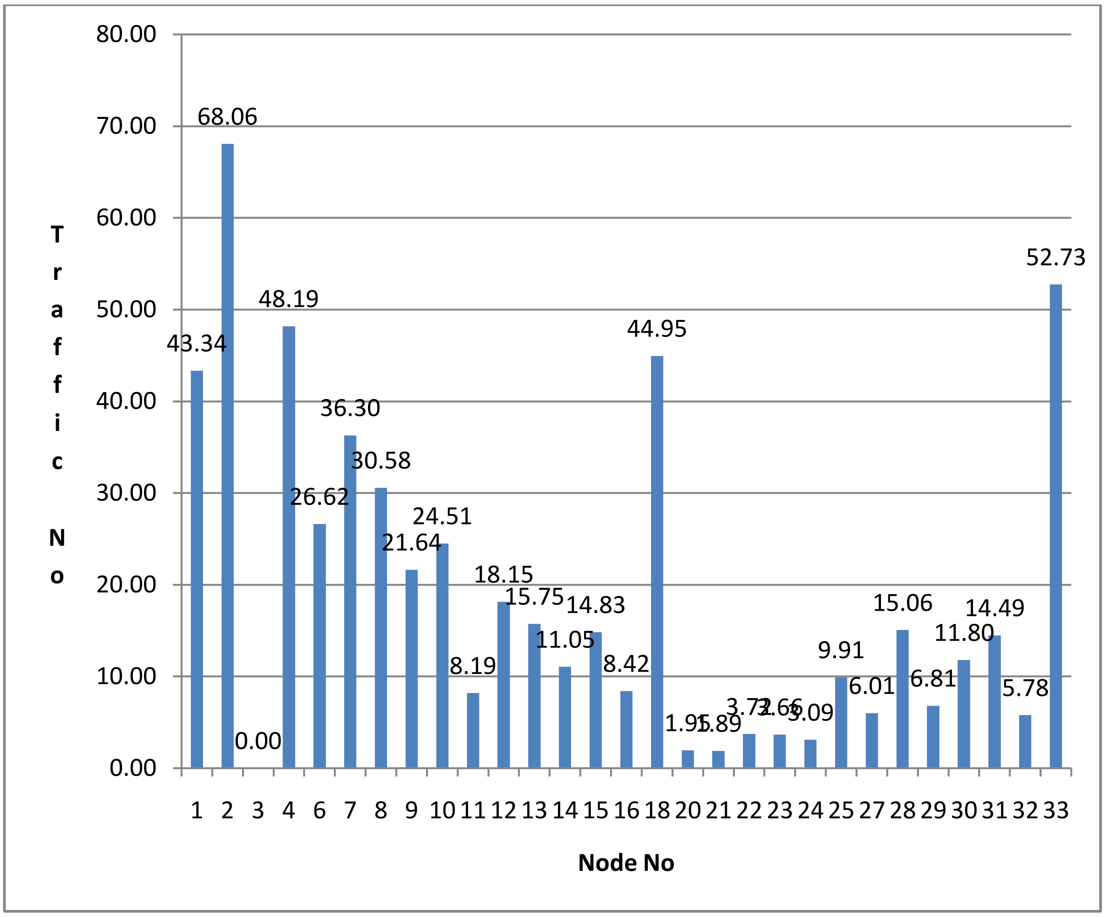


Figure 5.15 Flow from Node 3 (Mirpur-1 Node)

Figure 5.15 describes the flow from node 3 (Mirpur-1 Node) to different nodes. Main features of this node are presented in Table 5.3:

Table 5.4: Main features of the flows from node 3

Major Catchment Areas	Major Destination Nodes
• Around mirpur-1 circle	• Technical junction Node • Bashabo Node • Mirpur-10 Node • Gulshan-2 Node

5.4.4 Flow from Mirpur-10 Node

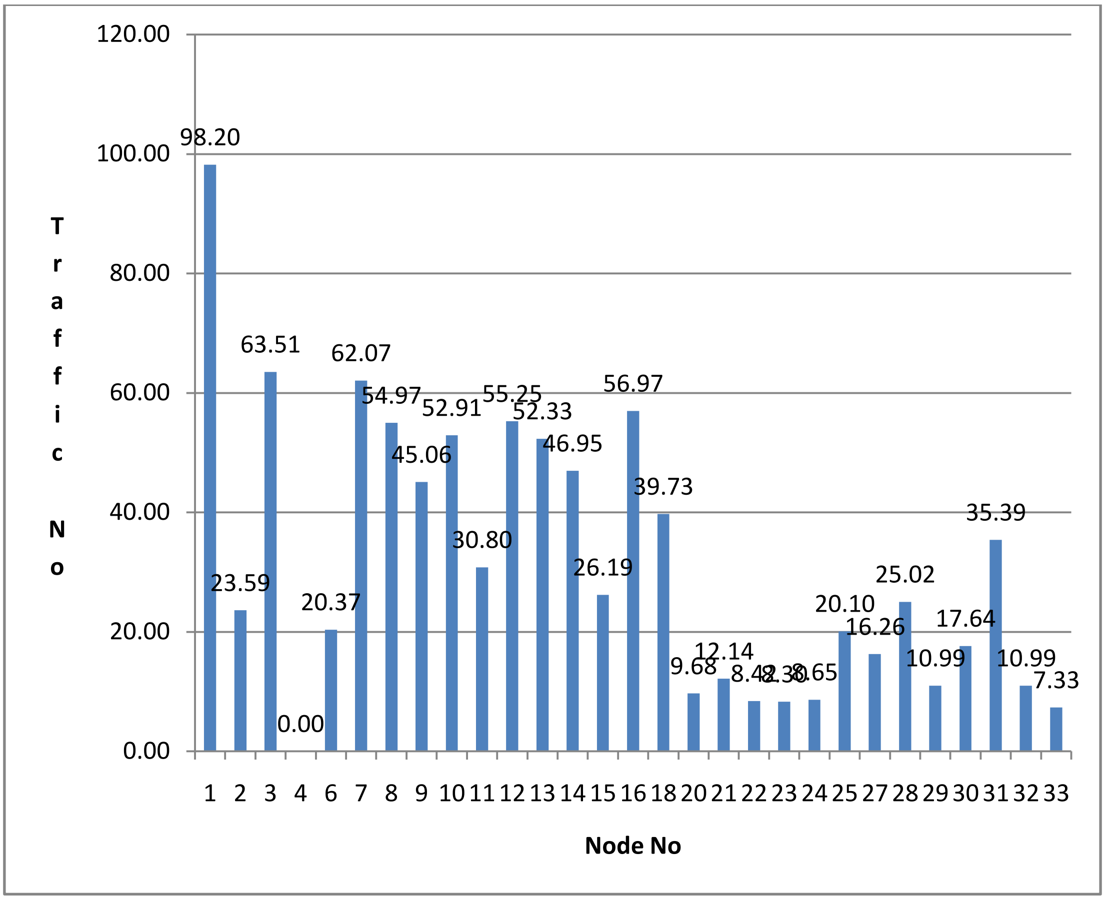


Figure 5.16 Flow from Node 4 (Mirpur-10 Node)

Figure 5.16 describes the flow from Node 4 (Mirpur-10 Node) to different nodes. Main features of this node are presented in Table 5.4:

Table 5.5: Main features of the flows from node 4

Major Catchment Areas	Major Destination Nodes
• Around mirpur-10 circle • Pallabi • Part of Kajipara	• English Road Node (1) • Mirpur-1(3) • Shyamoli Node (7) • Kakoli Node (16) • KawranBazar Node (12) • Asad Gate Node (8) • Farmgate Node (13)

5.4.5 Flow from Agargaon Node

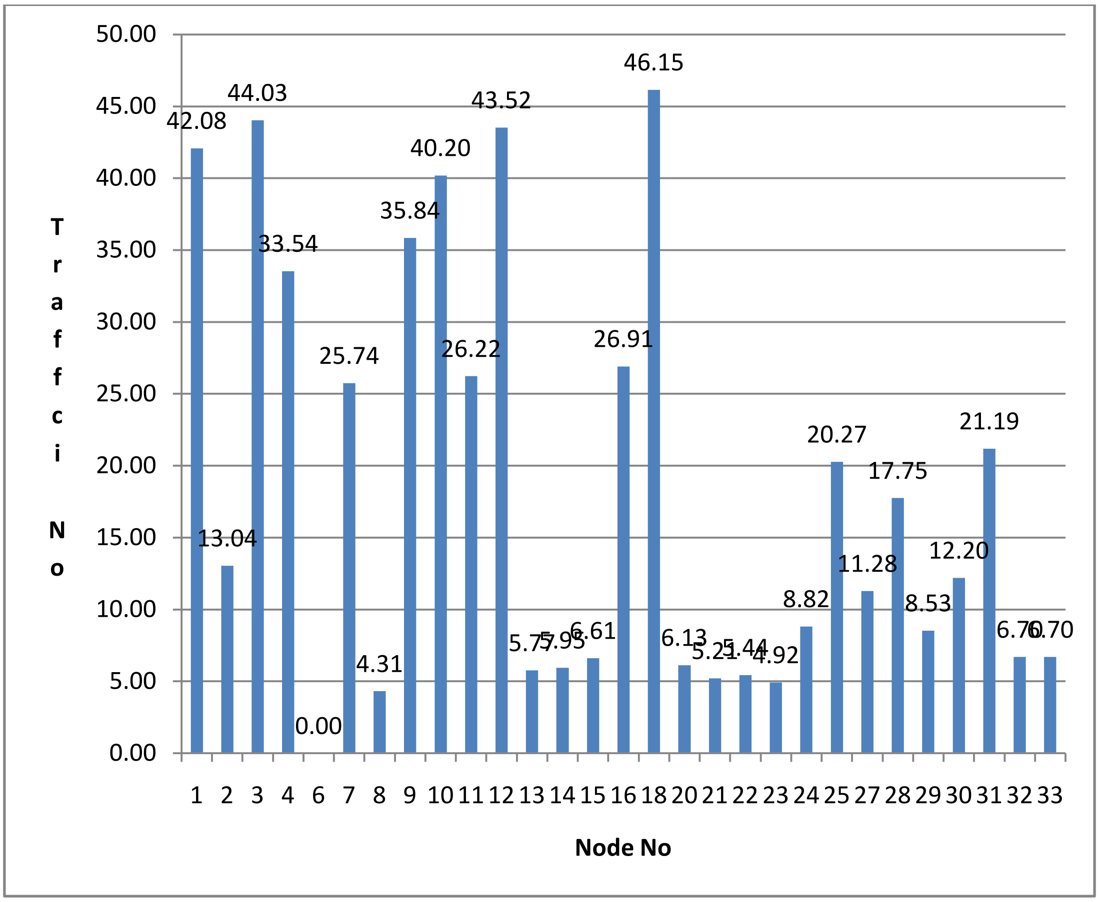


Figure 5.17 Flow from Node 6 (Agargaon Node)

Figure 5.17 describes the flow from Node 6 (Agargaon Node) to different nodes. Main features of this node are presented in Table 5.5:

Table 5.6: Main features of the flows from node 6

Major Catchment Areas	Major Destination Nodes
• Shewrapara • Taltola • Sher a bangla nagor	• Kuril Bishwa road Node (18) • Mirpur-1 Node (3) • English Road Node (1) • Science Lab Node (10)

5.4.6 Flow from Shyamoli Node

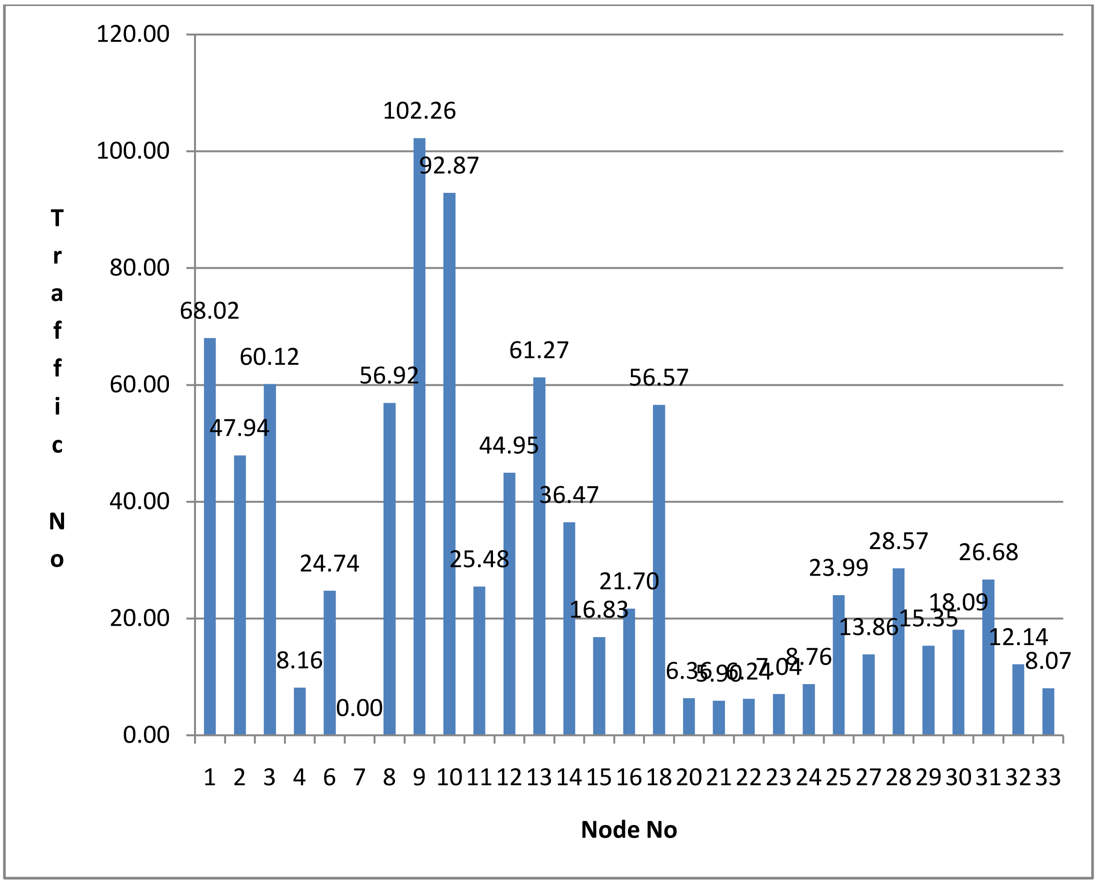


Figure 5.18 Flow from Node 7 (Shyamoli Node)

Figure 5.18 describes the flow from Node 7 (Shyamoli Node) to different nodes. Main features of this node are presented in Table 5.6:

Table 5.7: Main features of the flows from node 7

Major Catchment Areas	Major Destination Nodes
• Shyamoli • Sisu mela • Part of Mohammadpur • Adabar	• Dhanmondi 32 Node (9) • Science Lab Node (10) • English Road Node (1) • Farmgate Node (13) • Mirpur-1 Node (3)

5.4.7 Flow from Asad Gate Node

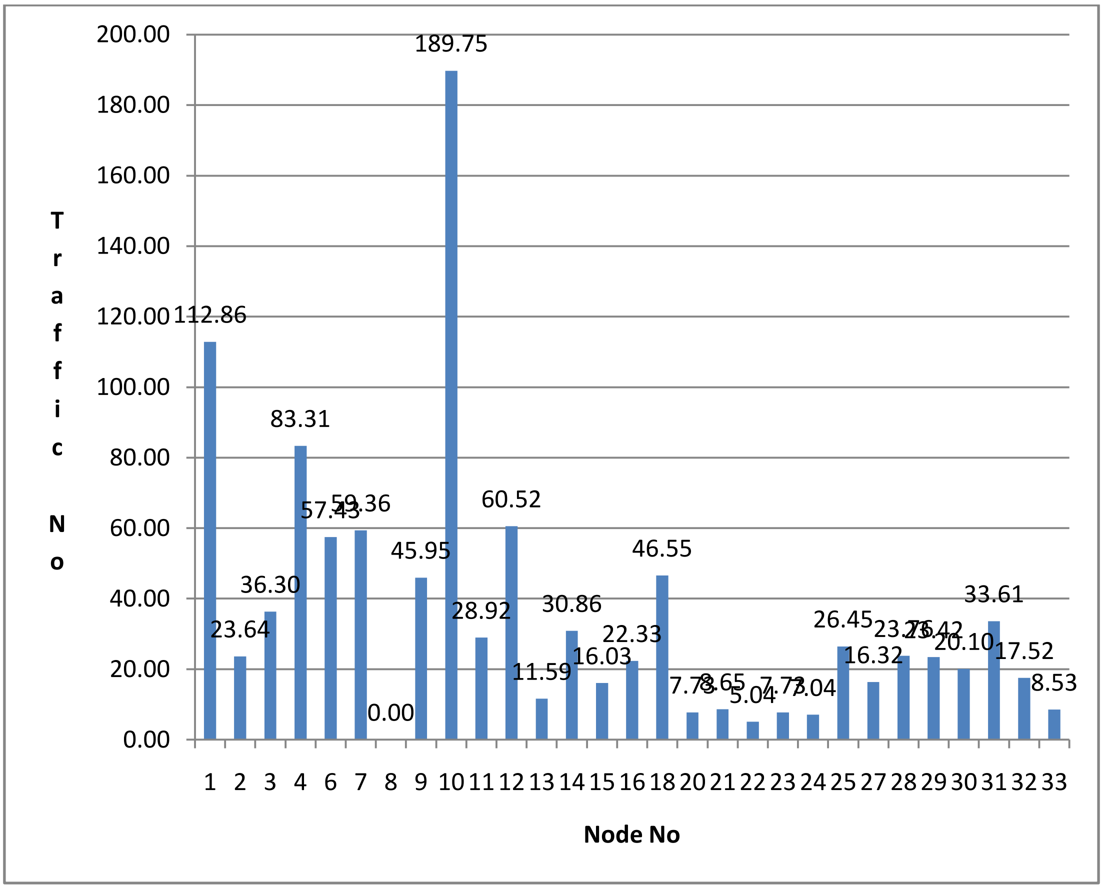


Figure 5.19 Flow from Node 8 (Asad Gate Node)

Figure 5.19 describes the flow from Node 8 (Asad Gate Node) to different nodes. Main features of this node are presented in Table 5.7:

Table 5.8: Main features of the flows from node 8

Major Catchment Areas	Major Destination Nodes
• Ashad Gate area • Lalmatia • Part of Mohammadpur • Part of Sukrabad	• Science Lab Node (10) • English Road Node (1) • Mirpur-10 Node (4)

5.4.8 Flow from Dhanmondi 32 Node

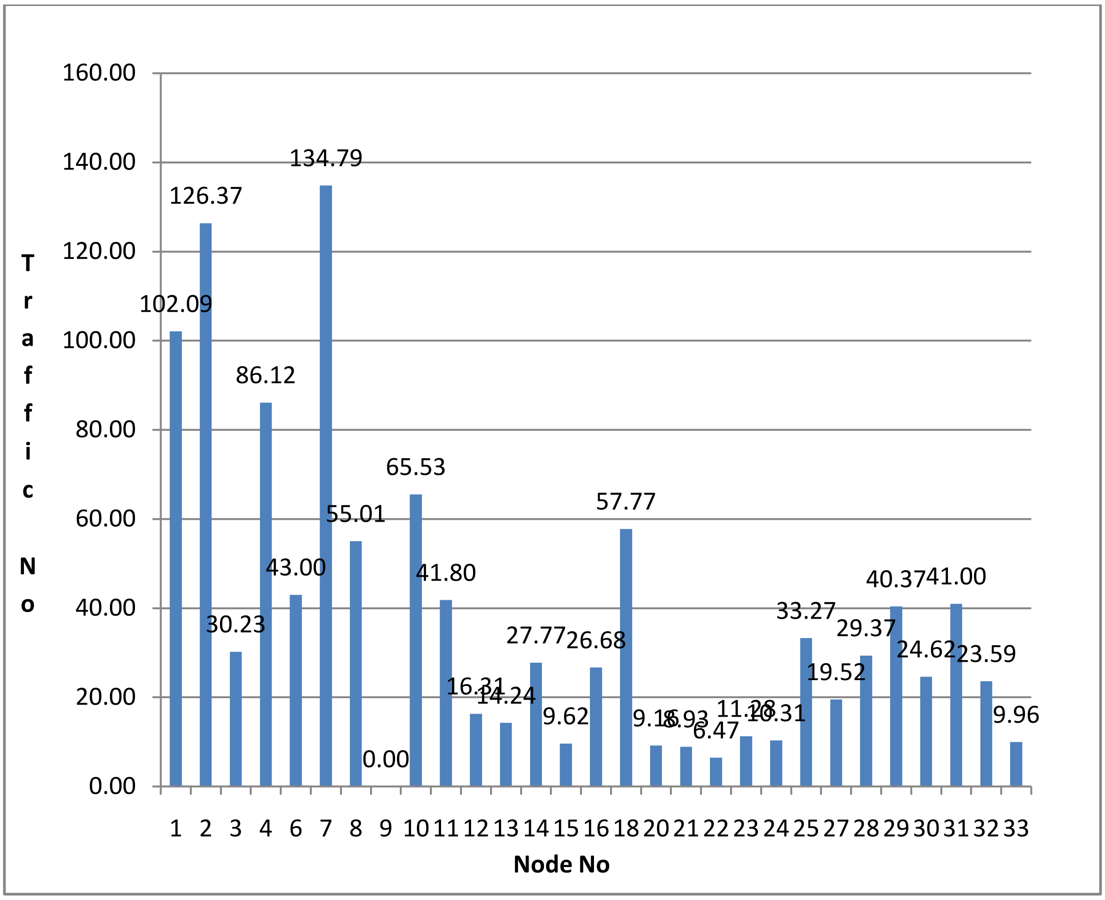


Figure 5.20 Flow from Node 9 (Dhanmondi 32 Node)

Figure 5.20 describes the flow from Node 9 (Dhanmondi 32 Node) to different nodes Main features of this node are presented in Table 5.8:

Table 5.9: Main features of the flows from node 9

Major Catchment Areas	Major Destination Nodes
• Part of Dhanmondi • Shankar • Rayer bazaar • Sukrabad	• Shyamoli Node (7) • Technical junction Node(2) • English Road Node (1) • Mirpur-10 Node (4)

5.4.9 Flow from Science Lab Node

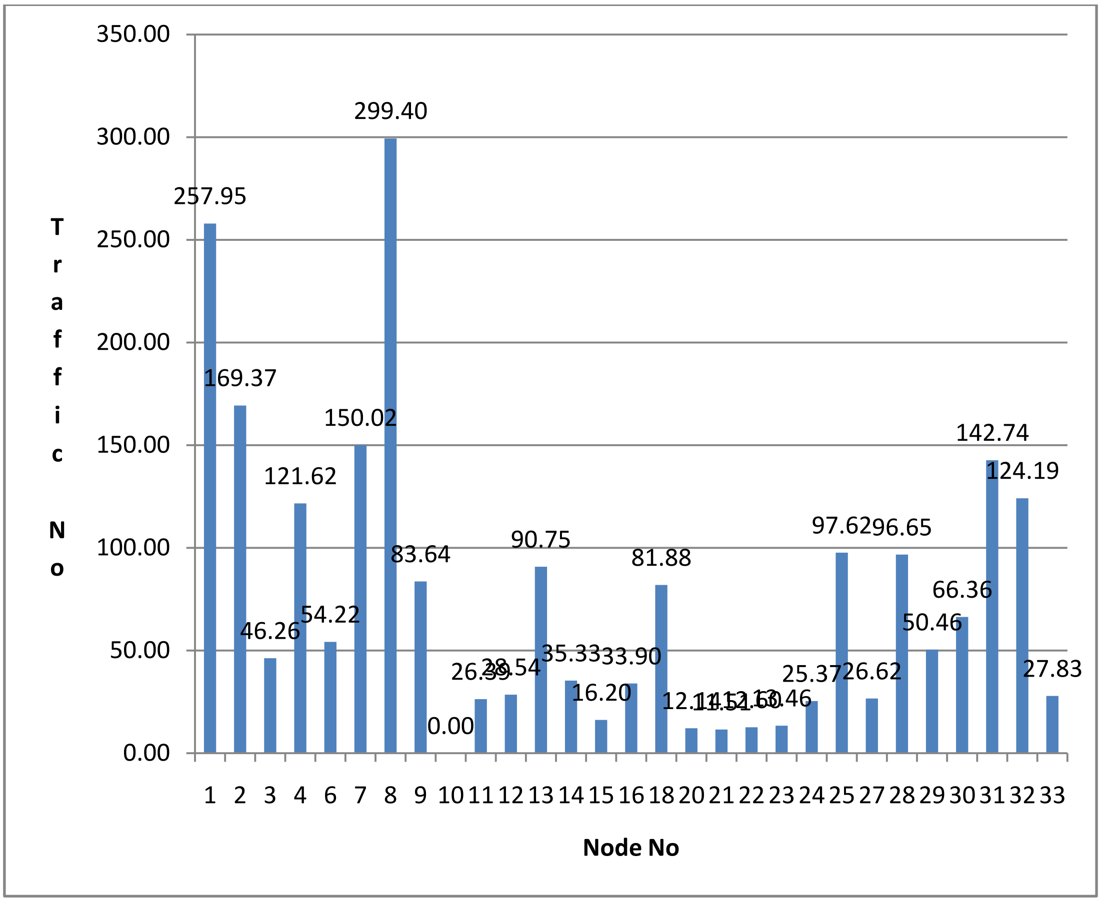


Figure 5.21 Flow from Node 10 (Science Lab Node)

Figure 5.21 describes the flow from Node 10 (Science Lab Node) to different nodes. Main features of this node are presented in Table 5.9:

Table 5.10: Main features of the flows from node 10

Major Catchment Areas	Major Destination Nodes
• Part of Dhanmondi • Kataban • Pilkhana • Hazaribag	• Asad Gate Node (8) • English Road Node (1) • Technical junction Node(2)

5.4.10 Flow from Shahbag Node

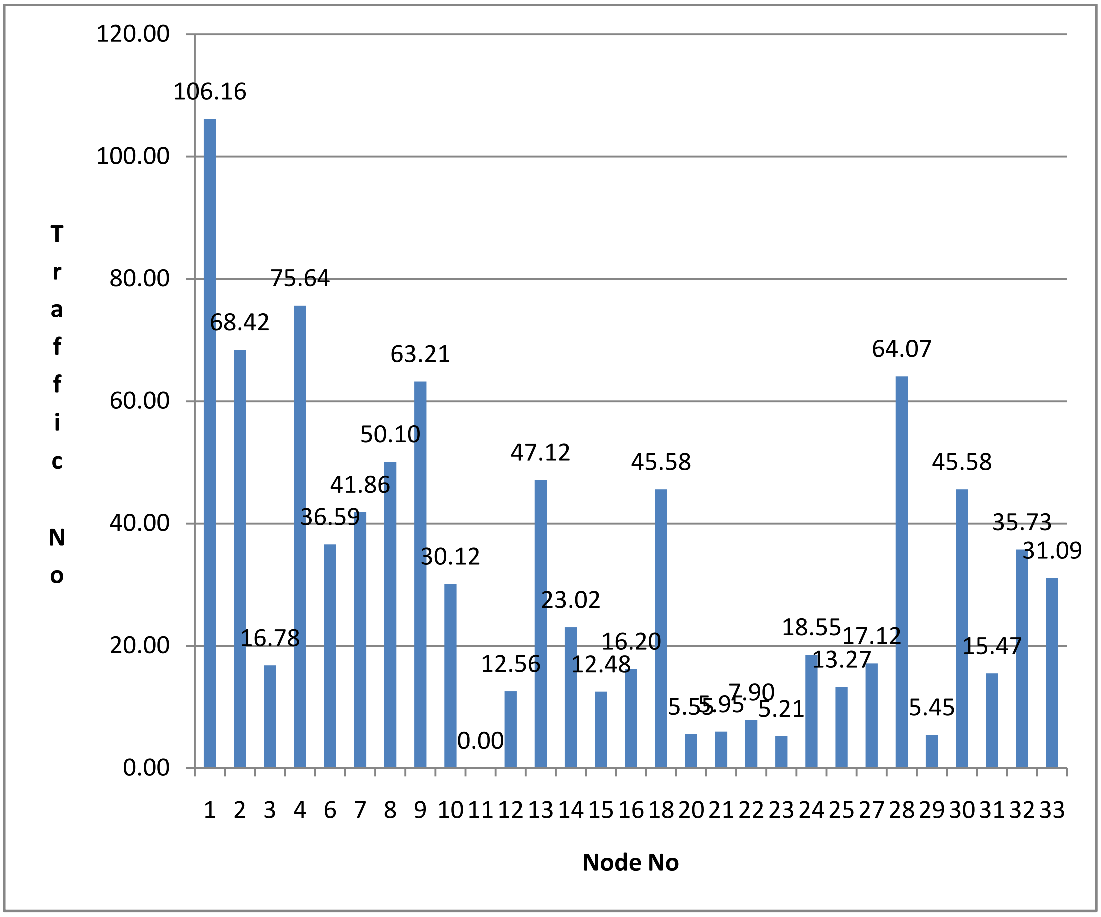


Figure 5.22 Flow from Node 11 (Shahbag Node)

Figure 5.22 describes the flow from Node 11 (Shahbag Node) to different nodes. Main features of this node are presented in Table 5.10:

Table 5.11: Main features of the flows from node 11

Major Catchment Areas	Major Destination Nodes
• Shahbag • Part of University area • Ramna	• English Road Node (1) • Mirpur-10 Node (4) • Technical junction Node(2) • Jatrabari Node (28) • Dhanmondi 32 Node (9)

5.4.11 Flow from KawranBazar Node

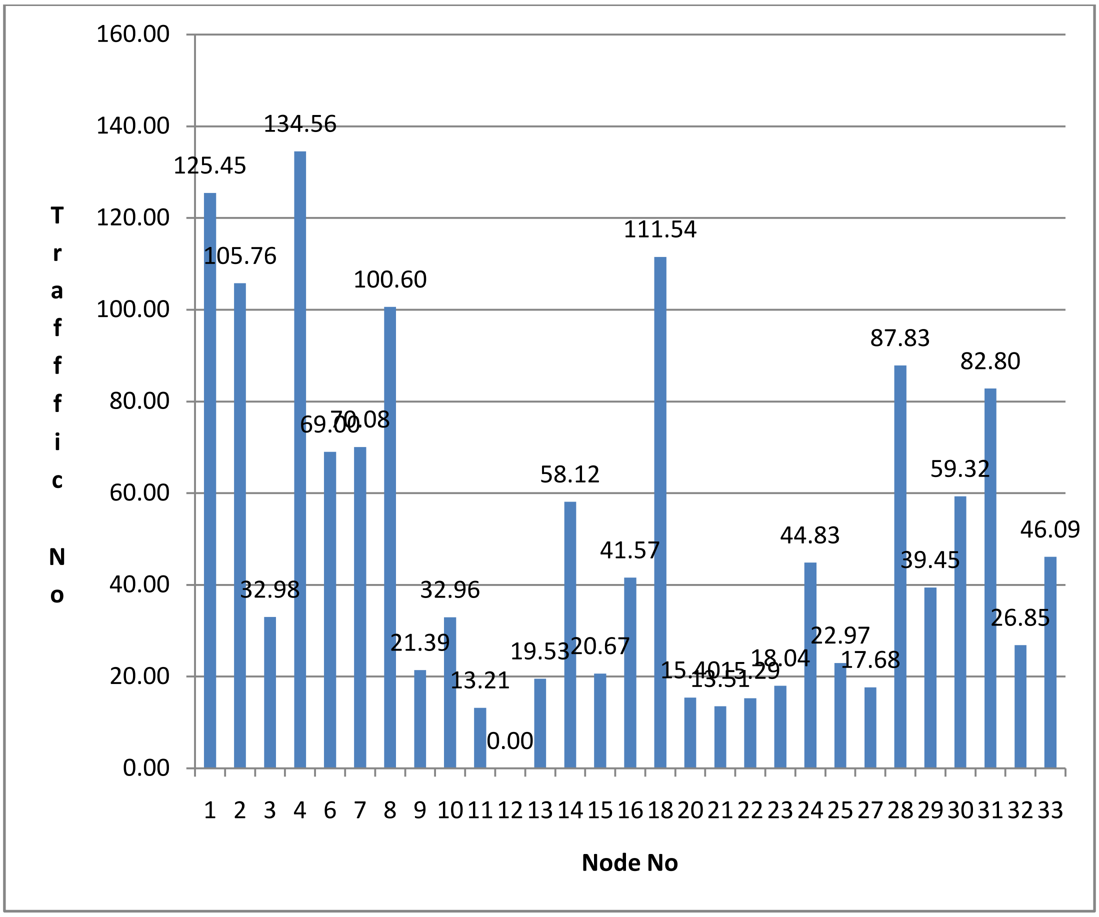


Figure 5.23 Flow from Node 12 (KawranBazar Node)

Figure 5.23 describes the flow from Node 12 (KawranBazar Node) to different nodes Main features of this node are presented in Table 5.11:

Table 5.12: Main features of the flows from node 12

Major Catchment Areas	Major Destination Nodes
• Kawranbazar • Hatirjheel • Kathalbagan	• Mirpur-10 Node (4) • English Road Node (1) • Kuril Bishwa road Node (18) • Technical junction Node(2) • Asad Gate Node (8)

5.4.12 Flow from Farmgate Node

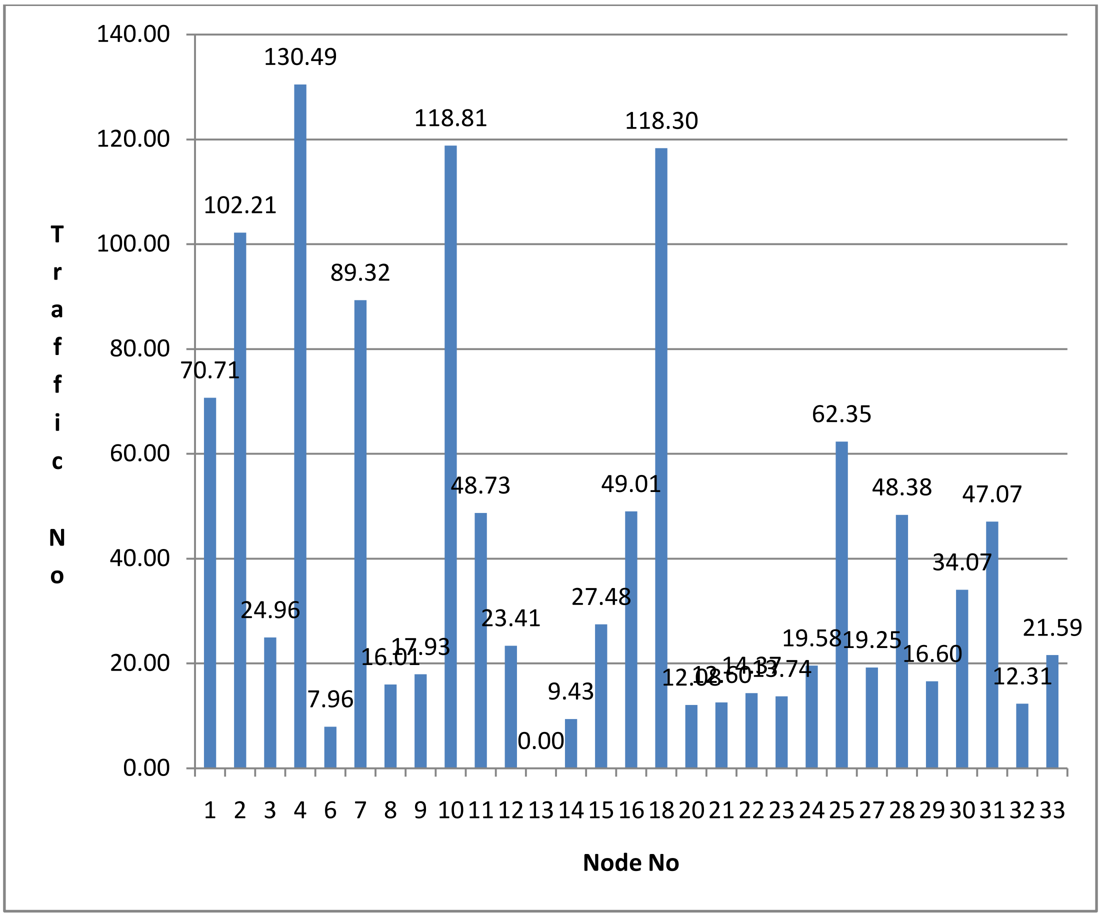


Figure 5.24 Flow from Node 13 (Farmgate Node)

Figure 5.24 describes the flow from Node 13 (Farmgate Node) to different nodes Main features of this node are presented in Table 5.12:

Table 5.13: Main features of the flows from node 13

Major Catchment Areas	Major Destination Nodes
• Farmgate • Monipuripara • Tejkunipara	• Mirpur-10 Node (4) • Science Lab Node (10) • Kuril Bishwa road Node (18) • Technical junction Node(2)

5.4.13 Flow from Mahakhali Node

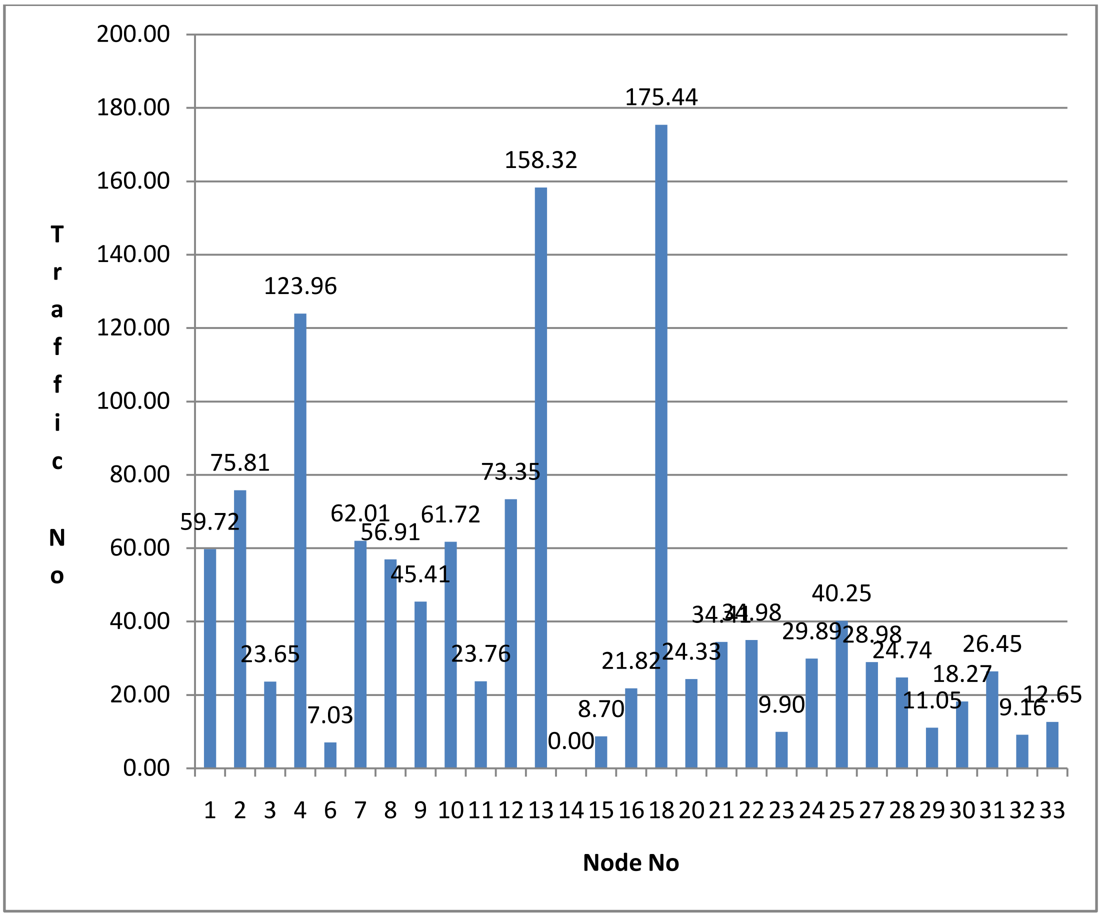


Figure 5.25 Flow from Node 14 (Mahakhali Node)

Figure 5.25 describes the flow from Node 14 (Mahakhali Node) to different nodes. Main features of this node are presented in Table 5.13:

Table 5.14: Main features of the flows from node 14

Major Catchment Areas	Major Destination Nodes
• Mohakhali • Nakhalpara	• Kuril Bishwa road Node (18) • Farmgate Node (13) • Mirpur-10 Node (4)

5.4.14 Flow from Mirpur-14 Node

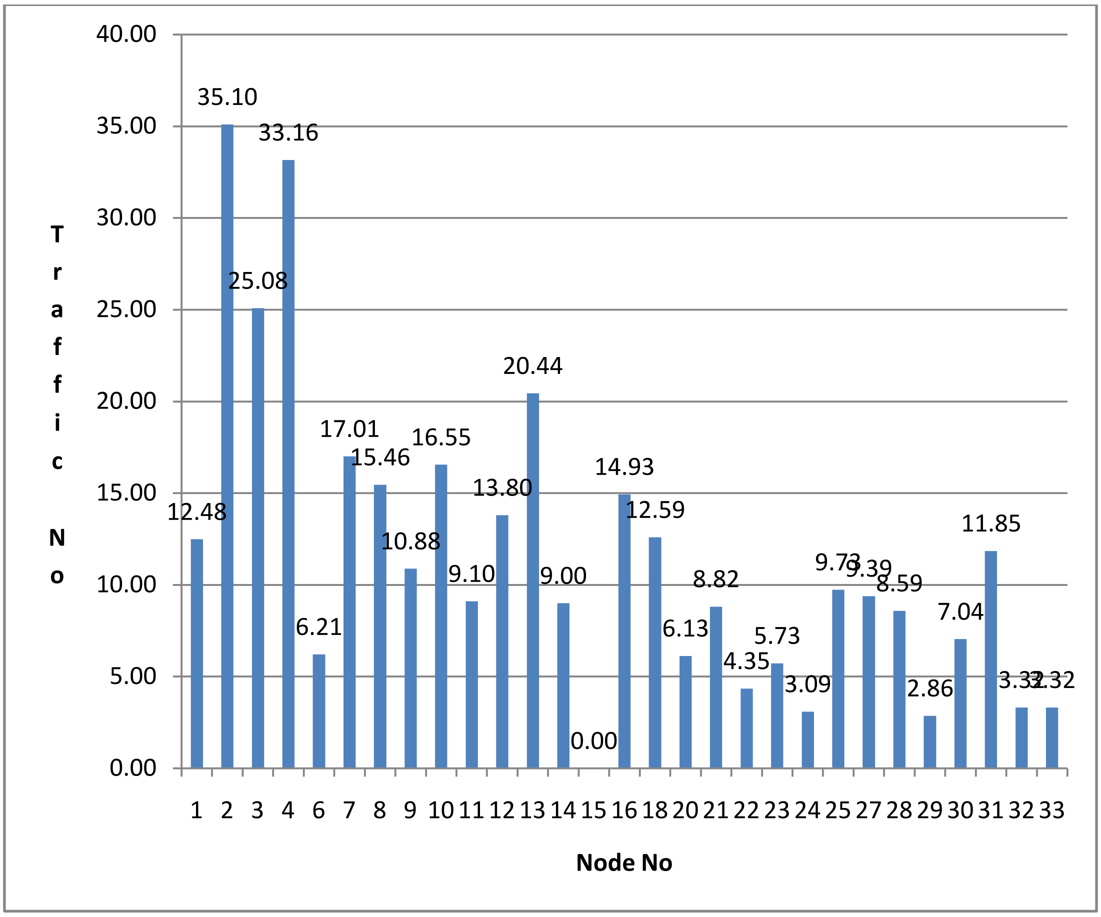


Figure 5.26 Flow from Node 15 (Mirpur-14 Node)

Figure 5.26 describes the flow from Node 15 (Mirpur-14 Node) to different nodes. Main features of this node are presented in Table 5.14:

Table 5.15: Main features of the flows from node 15

Major Catchment Areas	Major Destination Nodes
• Mirpur 14 • Kachukhet • Ibrahimpur	• Technical junction Node(2) • Mirpur-10 Node (4) • Mirpur-1 Node (3) • Farmgate Node (13)

5.4.15 Flow from Kakoli Node

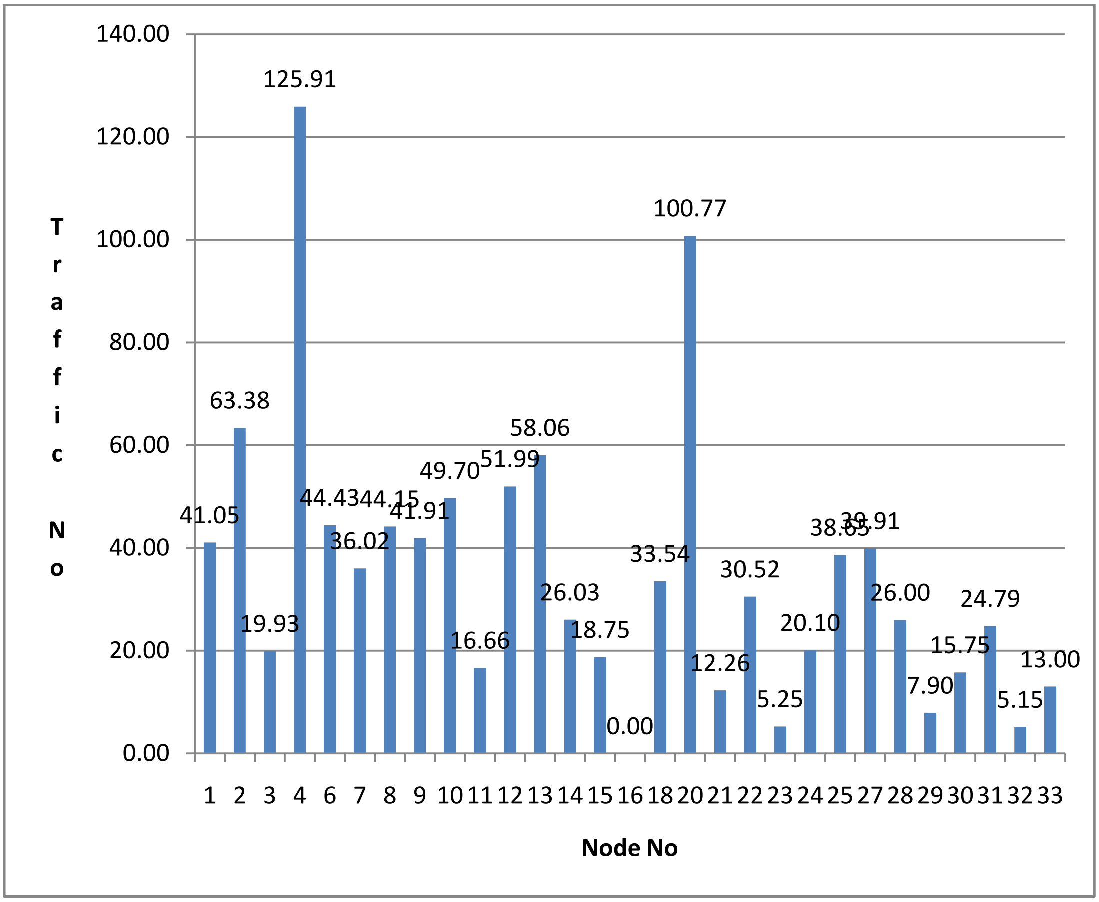


Figure 5.27 Flow from Node 16 (Kakoli Node)

Figure 5.27 describes the flow from Node 16 (Kakoli Node) to different nodes. Main features of this node are presented in Table 5.15:

Table 5.16: Main features of the flows from node 16

Major Catchment Areas	Major Destination Nodes
• Banani • Chairmanabari	• Mirpur-10 Node (4) • Natun Bazaar Node (20) • Technical junction Node(2)

5.4.16 Flow from Kuril Bishwa road Node

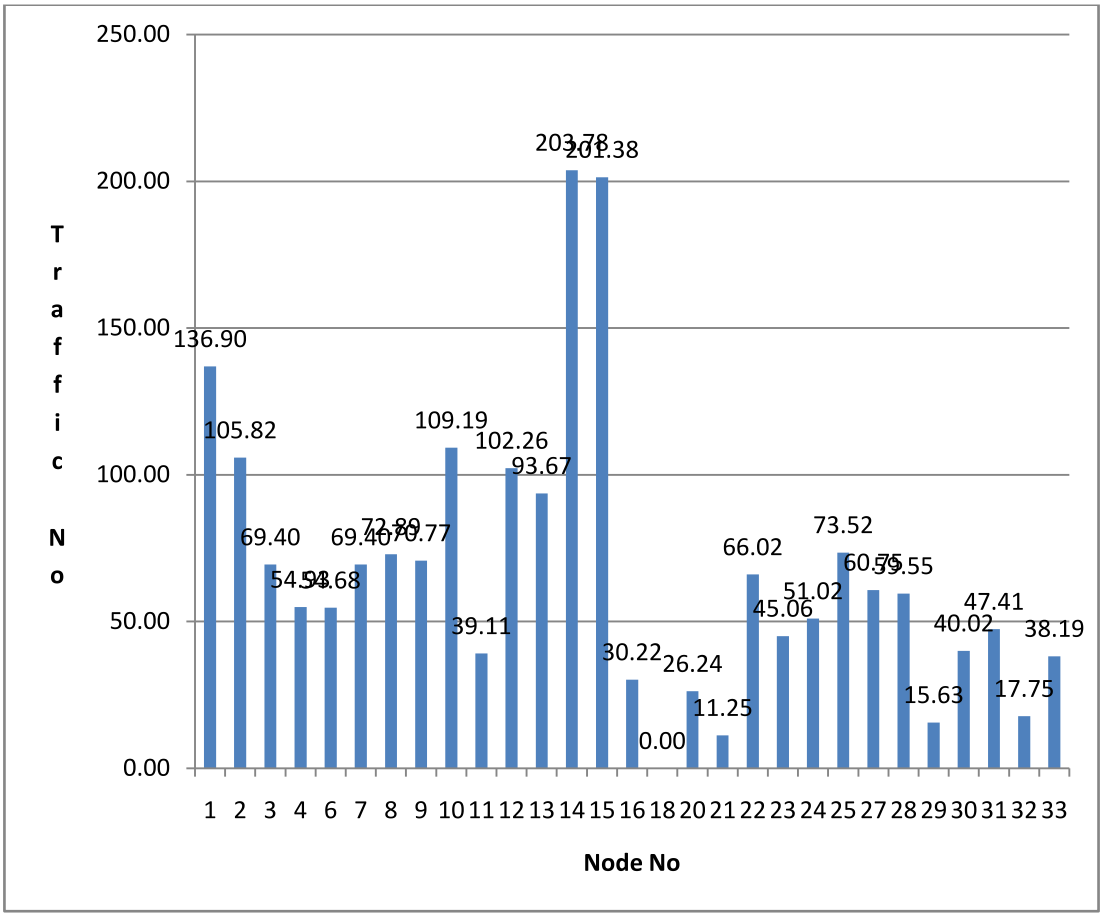


Figure 5.28 Flow from Node 18 (Kuril Bishwa road Node)

Figure 5.28 describes the flow from Node 18 (Kuril Bishwa road Node) to different nodes. Main features of this node are presented in Table 5.16:

Table 5.17: Main features of the flows from node 18

Major Catchment Areas	Major Destination Nodes
• Kuril • Uttora • Airport • Khilkhet • Bosundhara • Nadda	• Mahakhali Node (14) • Mirpur-14 Node (15) • English Road Node (1) • Technical junction Node(2)

5.4.17 Flow from Natun Bazaar Node

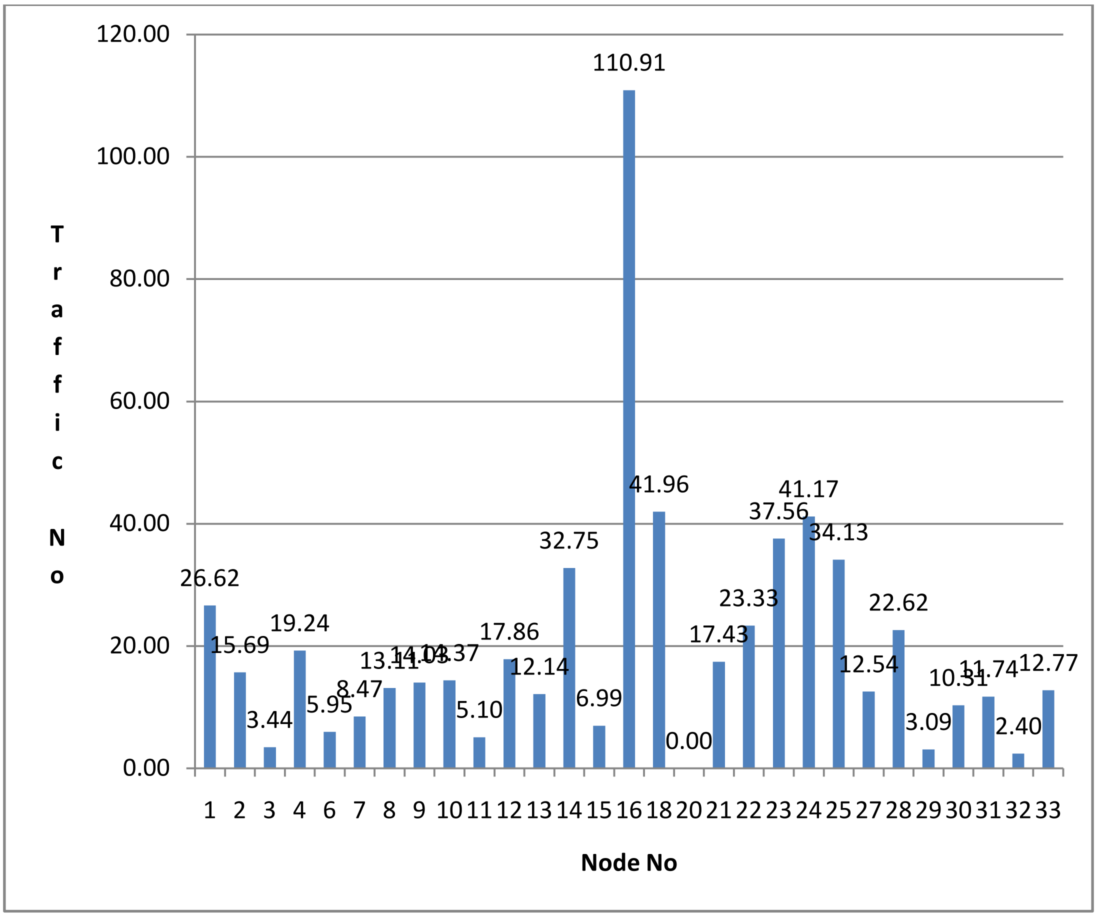


Figure 5.29 Flow from Node 20 (Natun Bazaar Node)

Figure 5.29 describes the flow from Node 20 (Natun Bazar Node) to different nodes Main features of this node are presented in Table 5.17:

Table 5.18: Main features of the flows from node 20

Major Catchment Areas	Major Destination Nodes
• Natun Bazar • Baridhara • Shahzadpur	• Kakoli Node (16) • Kuril Bishwa road Node (18) • Rampura Bridge Node (24)

5.4.18 Flow from Gulshan-2 Node

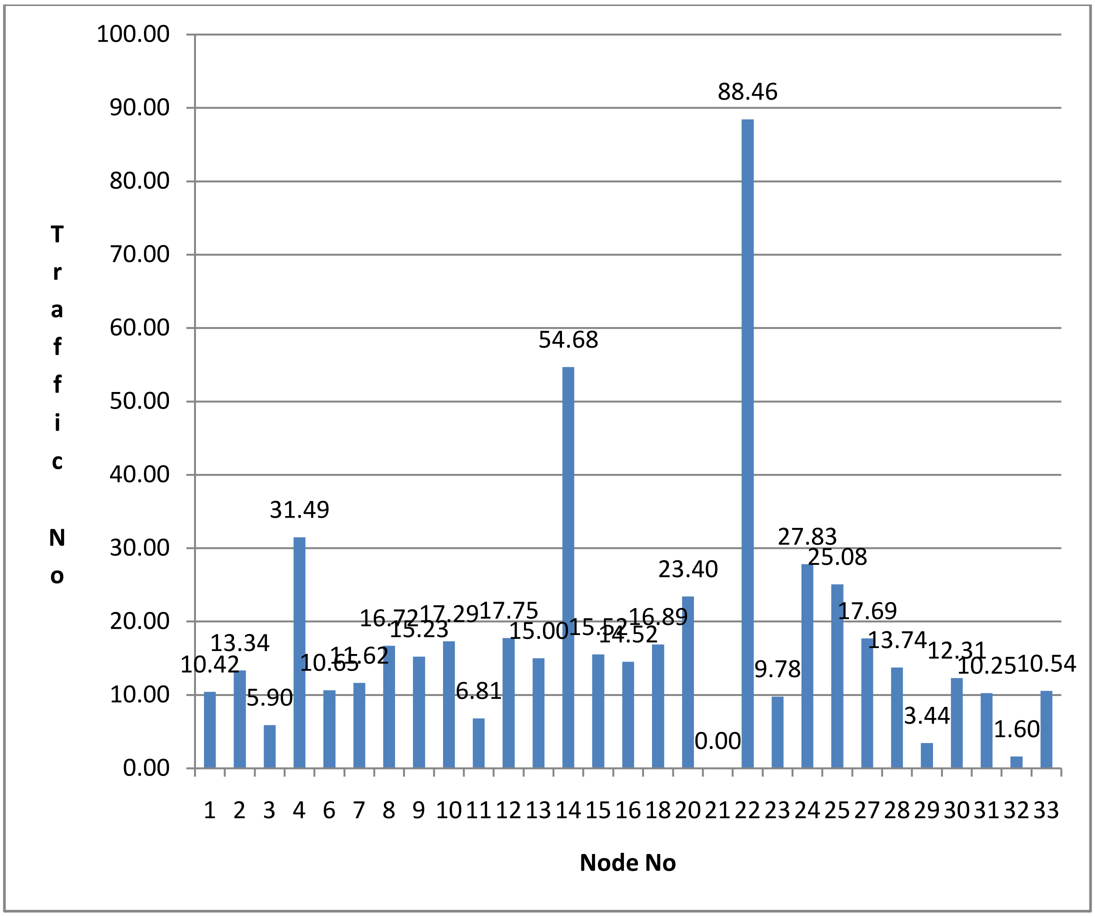


Figure 5.30 Flow from Node 21 (Gulshan-2 Node)

Figure 5.30 describes the flow from Node 21 (Gulshan-2 Node) to different nodes. Main features of this node are presented in Table 5.18:

Table 5.19: Main features of the flows from node 21

Major Catchment Areas	Major Destination Nodes
• Gulshan - 2	• Madhya Badda Node (22) • Mahakhali Node (14) • Mirpur-10 Node (4)

5.4.19 Flow from Madhya Badda Node

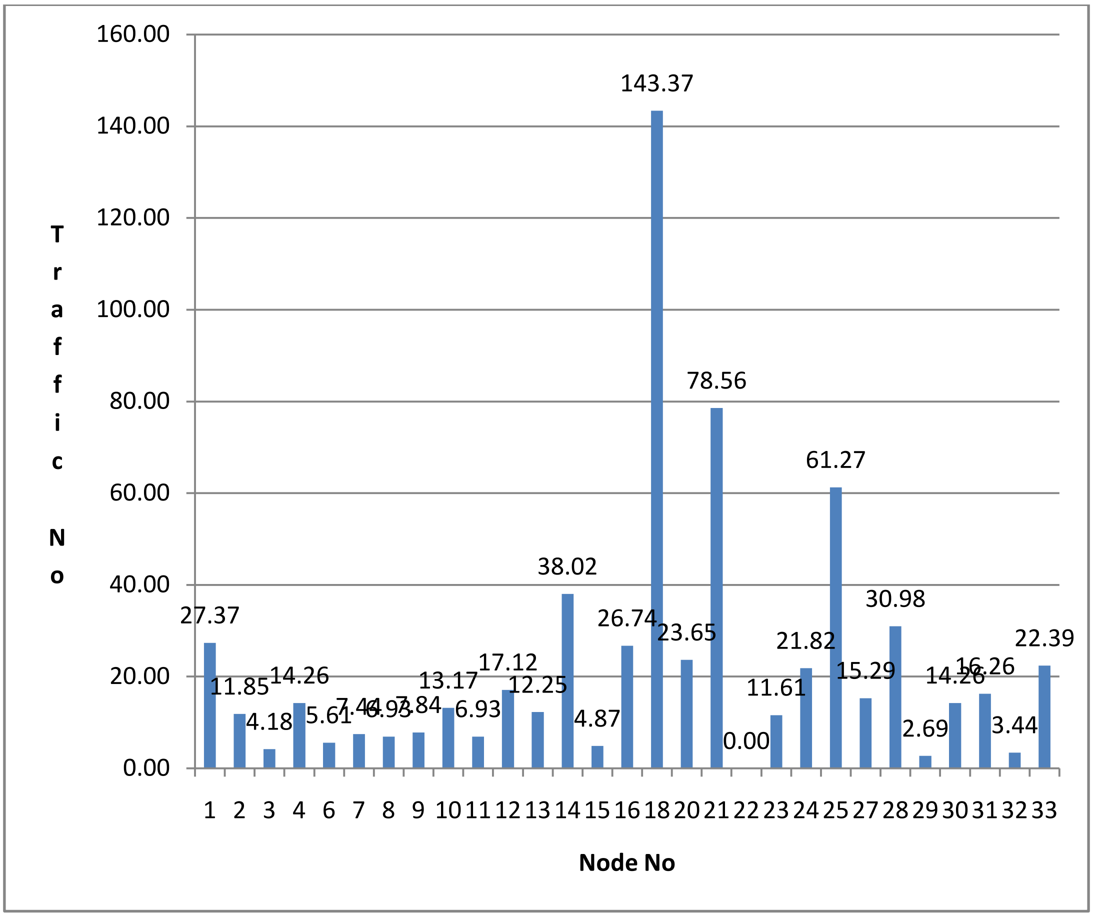


Figure 5.31 Flow from Node 22 (Madhya Badda Node)

Figure 5.31 describes the flow from Node 22 (Madhya Badda Node) to different nodes. Main features of this node are presented in Table 5.19:

Table 5.20: Main features of the flows from node 22

Major Catchment Areas	Major Destination Nodes
• Moddha Badda • Merul Badda • North Badda • Part of Shahzadpur	• Kuril Bishwa road Node (18) • Gulshan-2 Node (21) • Mouchak Node (25)

5.4.20 Flow from Gulshan-1 Node

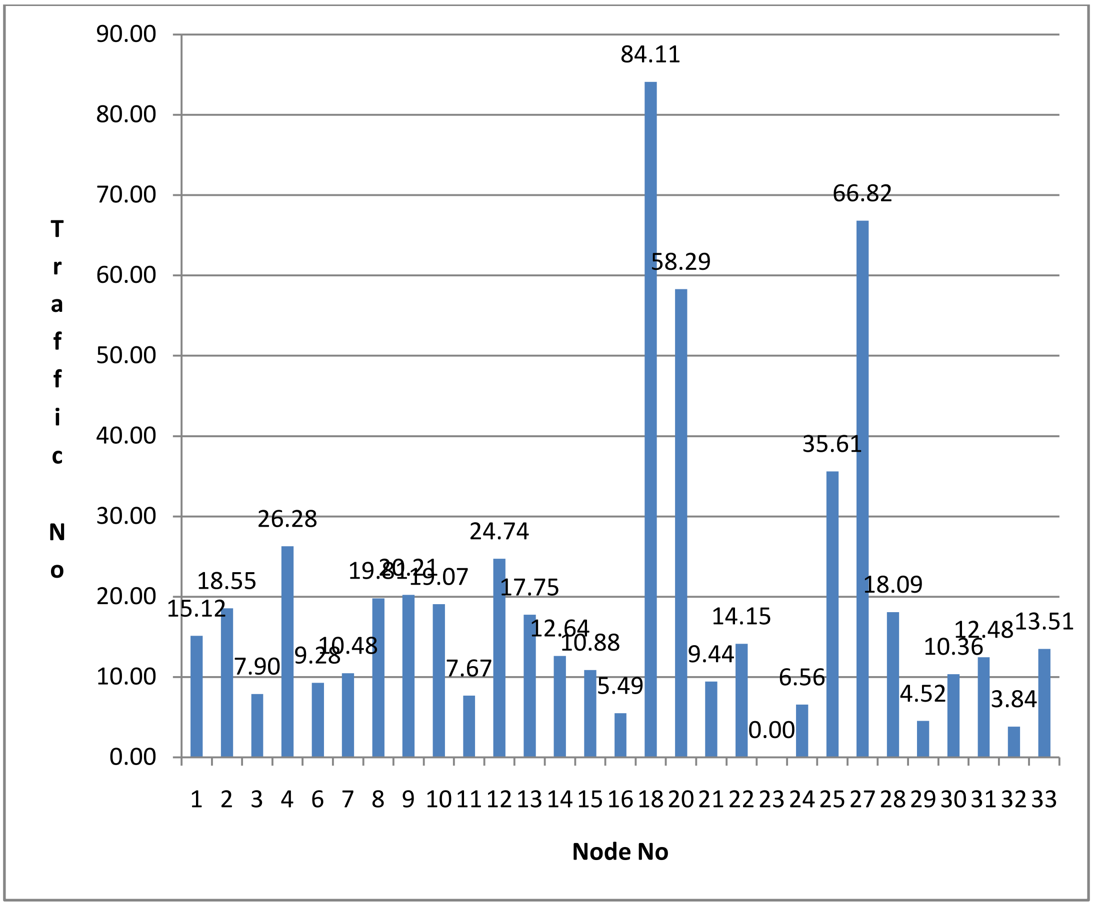


Figure 5.32 Flow from Node 23 (Gulshan-1 Node)

Figure 5.32 describes the flow from Node 23 (Gulshan-1 Node) to different nodes. Main features of this node are presented in Table 5.20:

Table 5.21: Main features of the flows from node 23

Major Catchment Areas	Major Destination Nodes
• Gulshan-1	• Kuril Bishwa road Node (18) • Square of seven Roads Node (27) • Natun Bazaar Node (20) • Mouchak Node (25)

5.4.21 Flow from Rampura Bridge Node

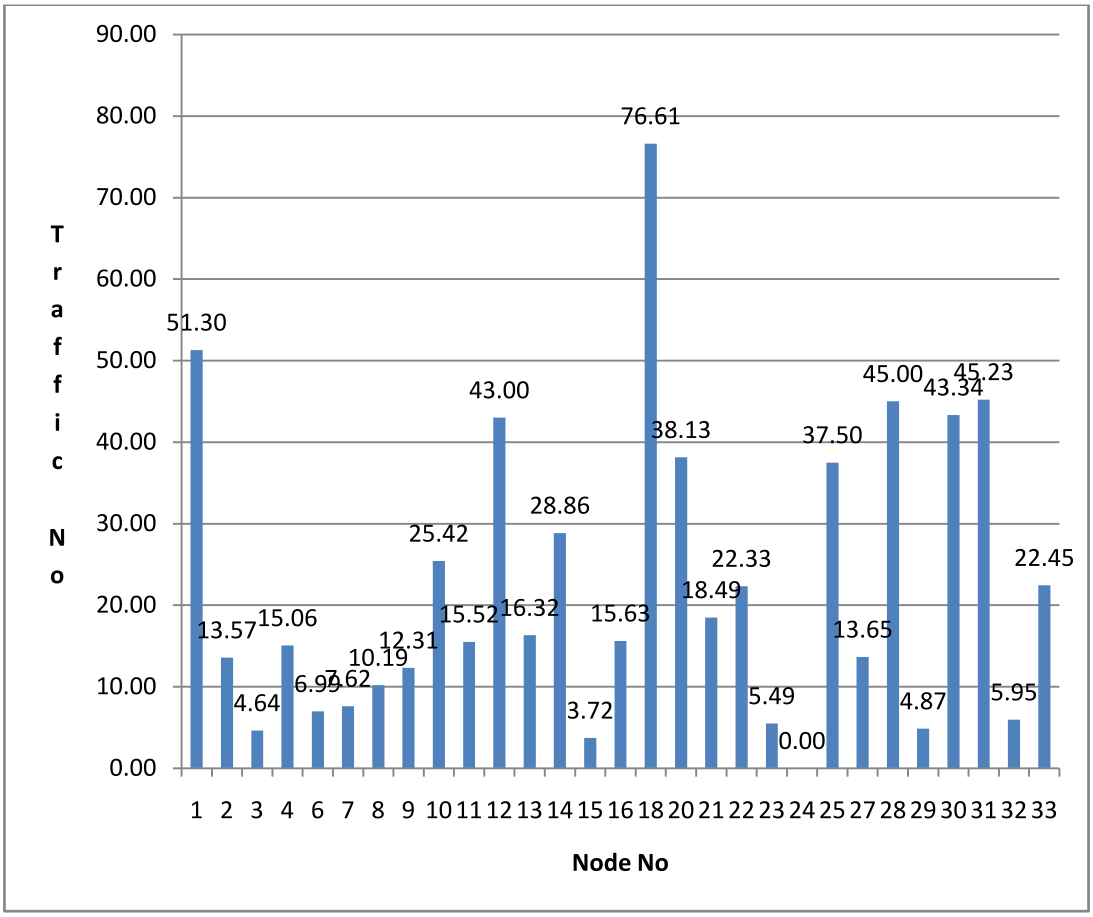


Figure 5.33 Flow from Node 24 (Rampura Bridge Node)

Figure 5.33 describes the flow from Node 24 (Rampura Bridge Node) to different nodes. Main features of this node are presented in Table 5.21:

Table 5.22: Main features of the flows from node 24

Major Catchment Areas	Major Destination Nodes
• Rampura • Banasri	• Kuril Bishwa road Node (18) • English Road Node (1) • Jatrabari Node (28) • ZeroPoint Node (31) • Motijheel Node (30) • KawranBazar Node (12)

5.4.22 Flow from Mouchak Node

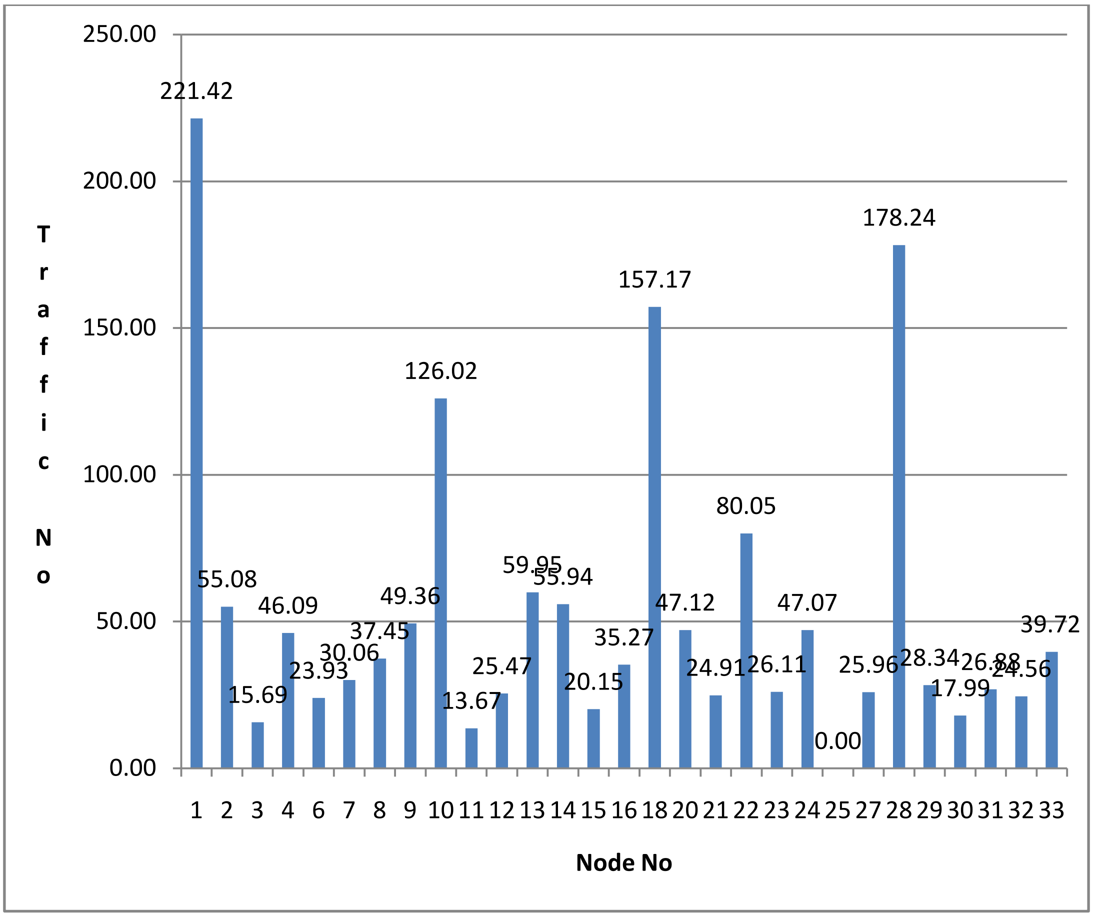


Figure 5.34 Flow from Node 25 (Mouchak Node)

Figure 5.34 describes the flow from Node 25 (Mouchak Node) to different nodes. Main features of this node are presented in Table 5.22:

Table 5.23: Main features of the flows from node 25

Major Catchment Areas	Major Destination Nodes
• Mouchak • Malibagh • Mogbazar • Part of Rampura • Rajarbagh • Shiddeshwari	• English Road Node (1) • Jatrabari Node (28) • Kuril Bishwa road Node (18) • Science Lab Node (10)

5.4.23 Flow from Square of seven Roads Node

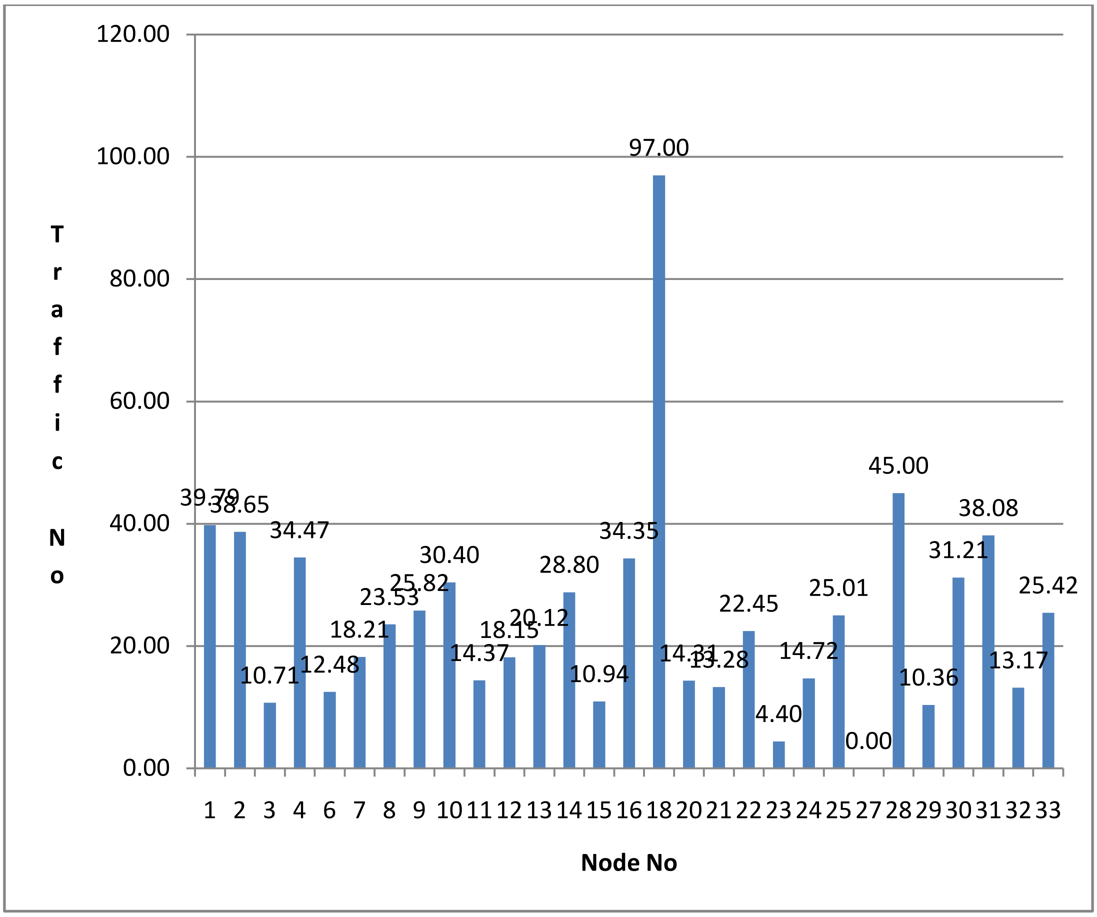


Figure 5.35 Flow from Node 27 (Square of seven Roads Node)

Figure 5.35 describes the flow from Node 27 (Square of seven Roads Node) to different nodes. Main features of this node are presented in Table 5.23:

Table 5.24: Main features of the flows from node 27

Major Catchment Areas	Major Destination Nodes
• Tejgaon	• Kuril Bishwa road Node (18) • Jatrabari Node (28)

5.4.24 Flow from Jatrabari Node

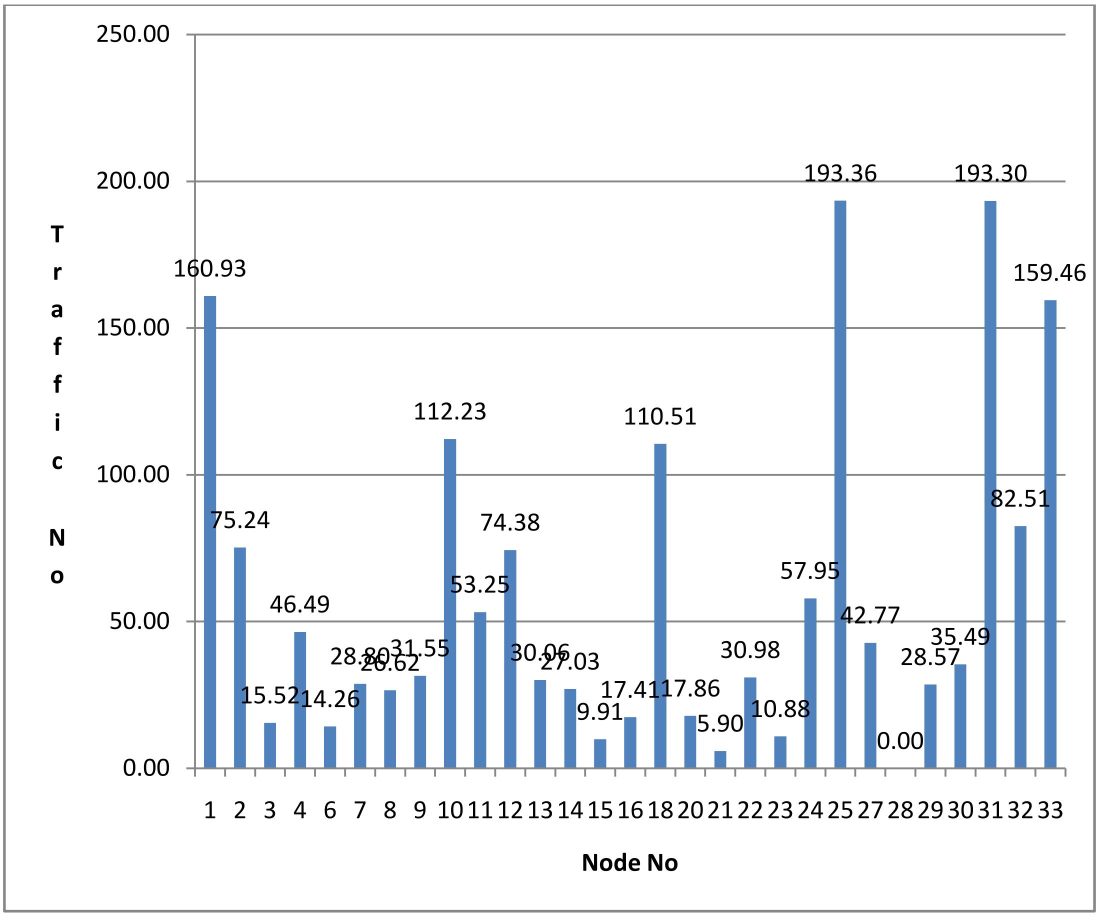


Figure 5.36 Flow from Node 28 (Jatrabari Node)

Figure 5.36 describes the flow from Node 28 (Jatrabari Node) to different nodes Main features of this node are presented in Table 5.24:

Table 5.25: Main features of the flows from node 28

Major Catchment Areas	Major Destination Nodes
• Jatrabari • Sanir akhra • Demra • Dania • Sutrapur • Saidabad	• Mouchak Node (25) • ZeroPoint Node (31) • English Road Node (1) • Bashabo Node (33) • Science Lab Node (10) • Kuril Bishwa road Node (18)

5.4.25 Flow from Azimpur Node

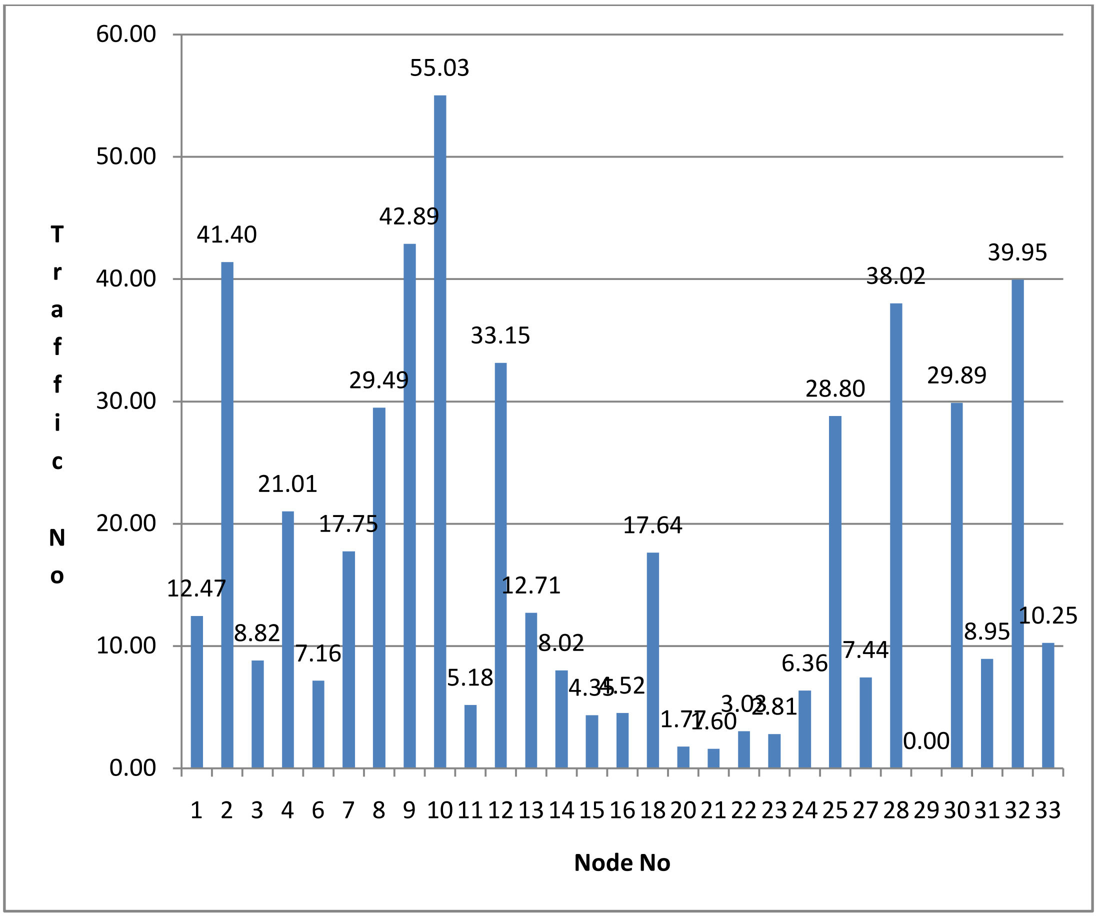


Figure 5.37 Flow from Node 29 (Azimpur Node)

Figure 5.37 describes the flow from Node 29 (Azimpur Node) to different nodes. Main features of this node are presented in Table 5.25:

Table 5.26: Main features of the flows from node 29

Major Catchment Areas	Major Destination Nodes
• Azmpur • Palashi • Part of University area • Nilkhet	• Science Lab Node (10) • Dhanmondi 32 Node (9) • Technical junction Node(2) • Chawk Bazar Node (32) • Jatrabari Node (28) • KawranBazar Node (12)

5.4.26 Flow from Motijheel Node

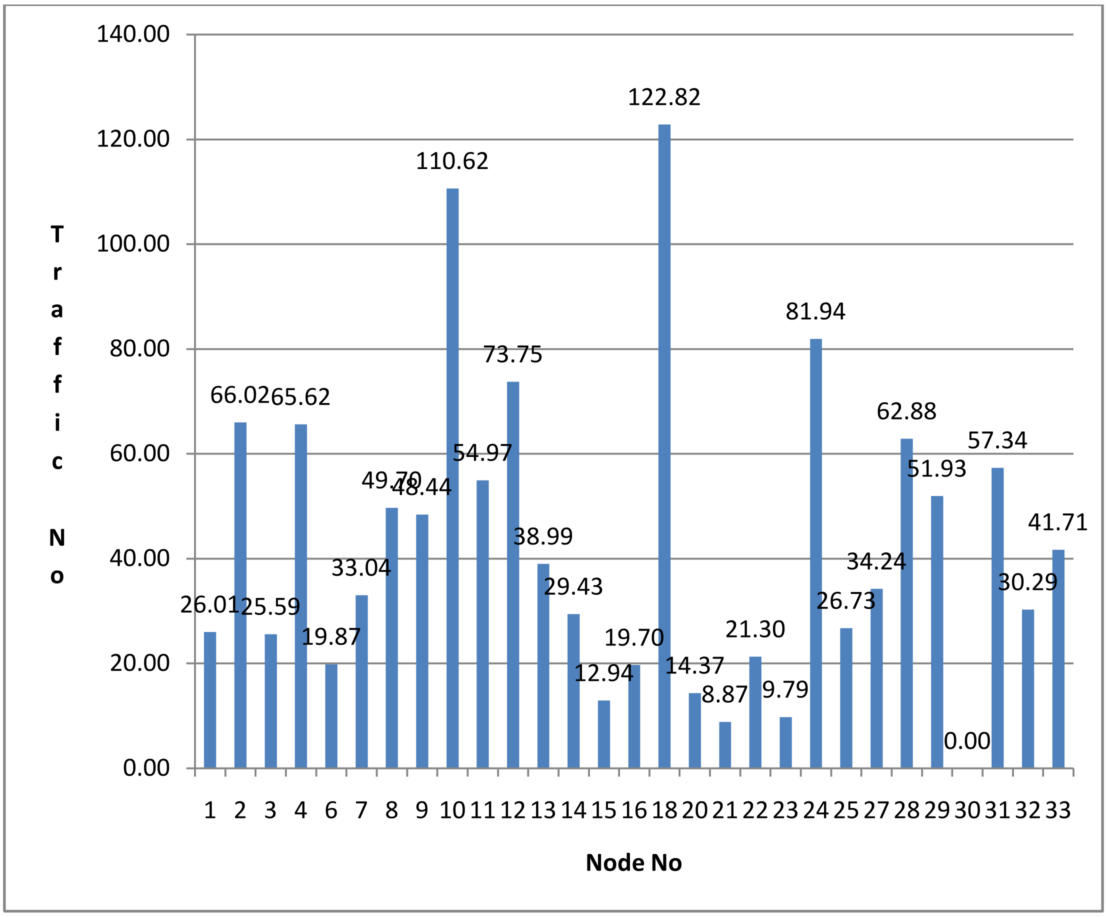


Figure 5.38 Flow from Node 30 (Motijheel Node)

Figure 5.38 describes the flow from Node 30 (Motijheel Node) to different nodes. Main features of this node are presented in Table 5.26:

Table 5.27: Main features of the flows from node 30

Major Catchment Areas	Major Destination Nodes
• Motijheel • Komolapur • Part of Gulisthan • Tikatuli	• Kuril Bishwa road Node (18) • Science Lab Node (10) • Rampura Bridge Node (24) • KawranBazar Node (12) • Mirpur-10 Node (4) • Technical junction Node(2)

5.4.27 Flow from Zero Point Node

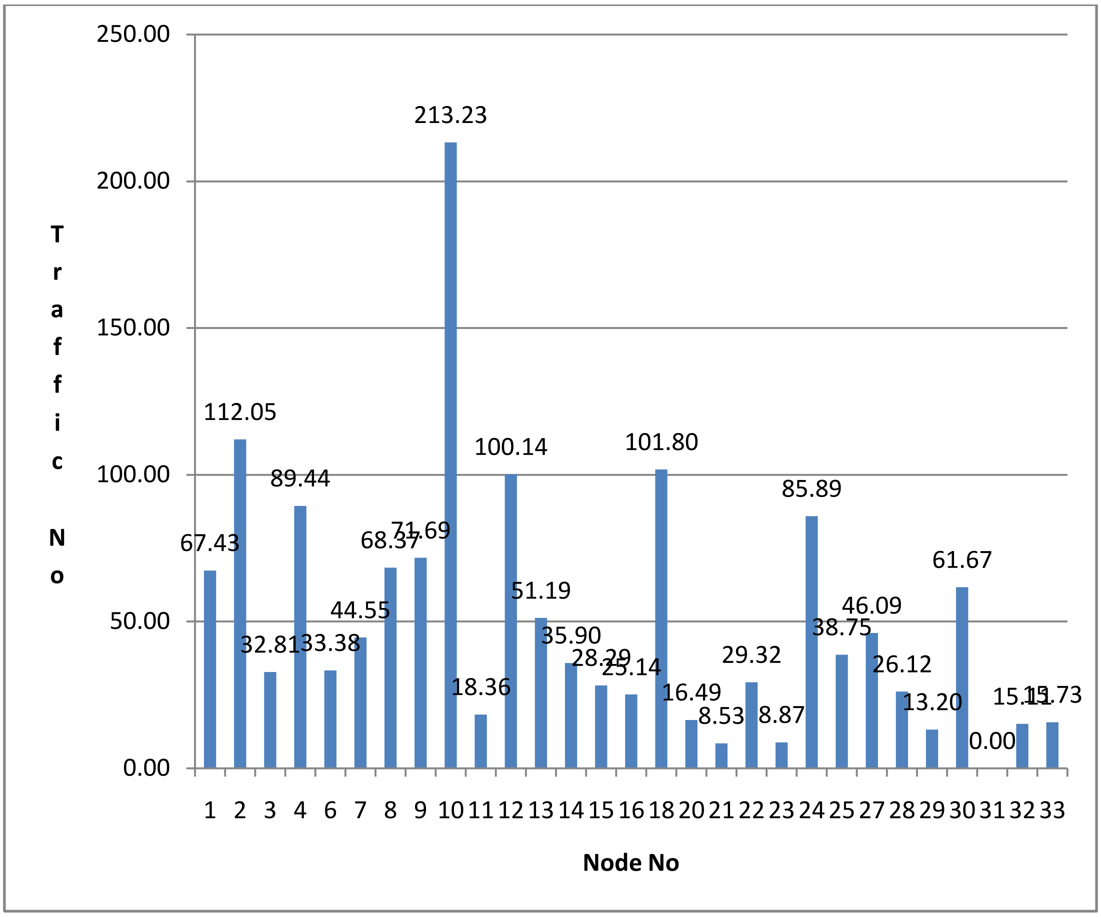


Figure 5.39 Flow from Node 31 (Zero Point Node)

Figure 5.39 describes the flow from Node 31 (Zero Point Node) to different nodes Main features of this node are presented in Table 5.27:

Table 5.28: Main features of the flows from node 31

Major Catchment Areas	Major Destination Nodes
• Gulistan • Paltan • Pressclub	• Science Lab Node (10) • Technical junction Node(2) • Kuril Bishwa road Node (18) • KawranBazar Node (12)

5.4.28 Flow from Chawk Bazar Node

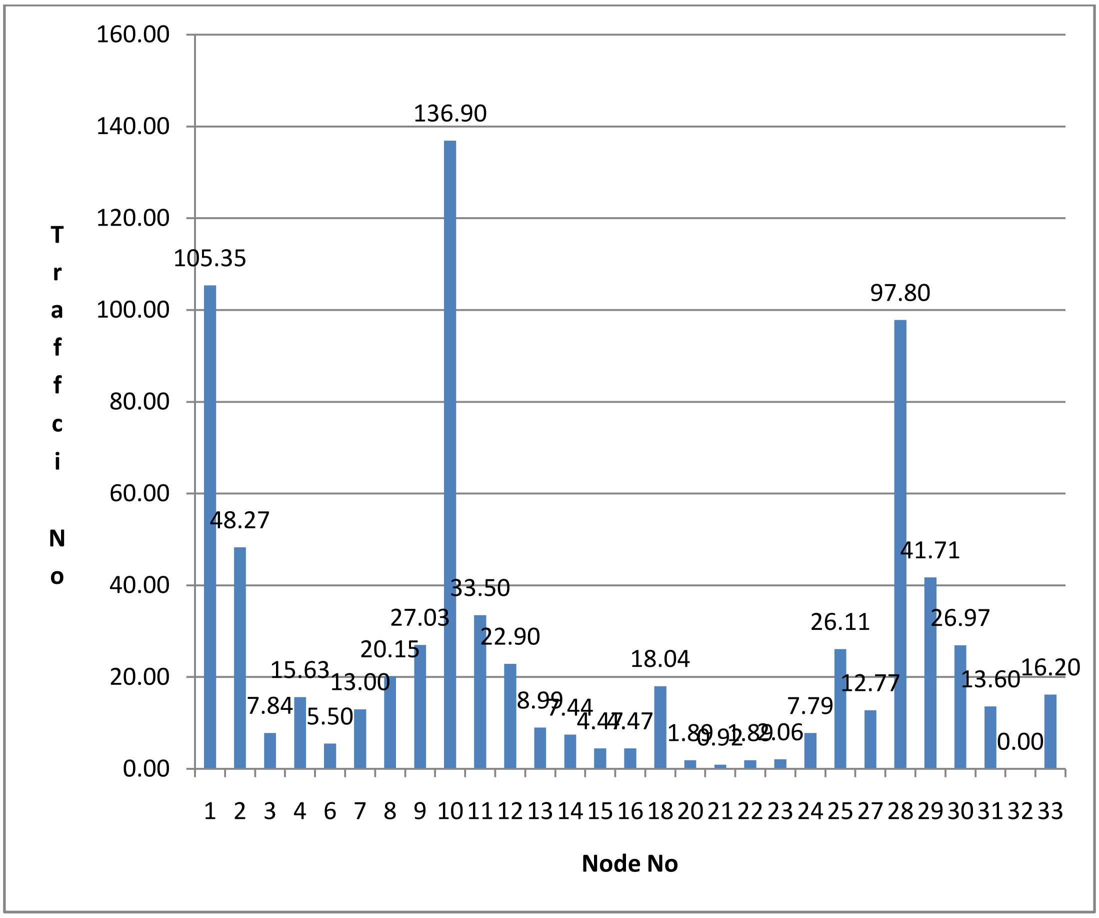


Figure 5.40 Flow from Node 32 (Chawk Bazar Node)

Figure 5.40 describes the flow from Node 32 (Chawk Bazar Node) to different nodes Main features of this node are presented in Table 5.28:

Table 5.29: Main features of the flows from node 32

Major Catchment Areas	Major Destination Nodes
• Chawk Bazar • Begun Bazar • Part of Kamrangir chor • A major portion of Old Dhaka	• Science Lab Node (10) • Jatrabari Node (28) • English Road Node (1)

5.4.29 Flow from Bashabo Node

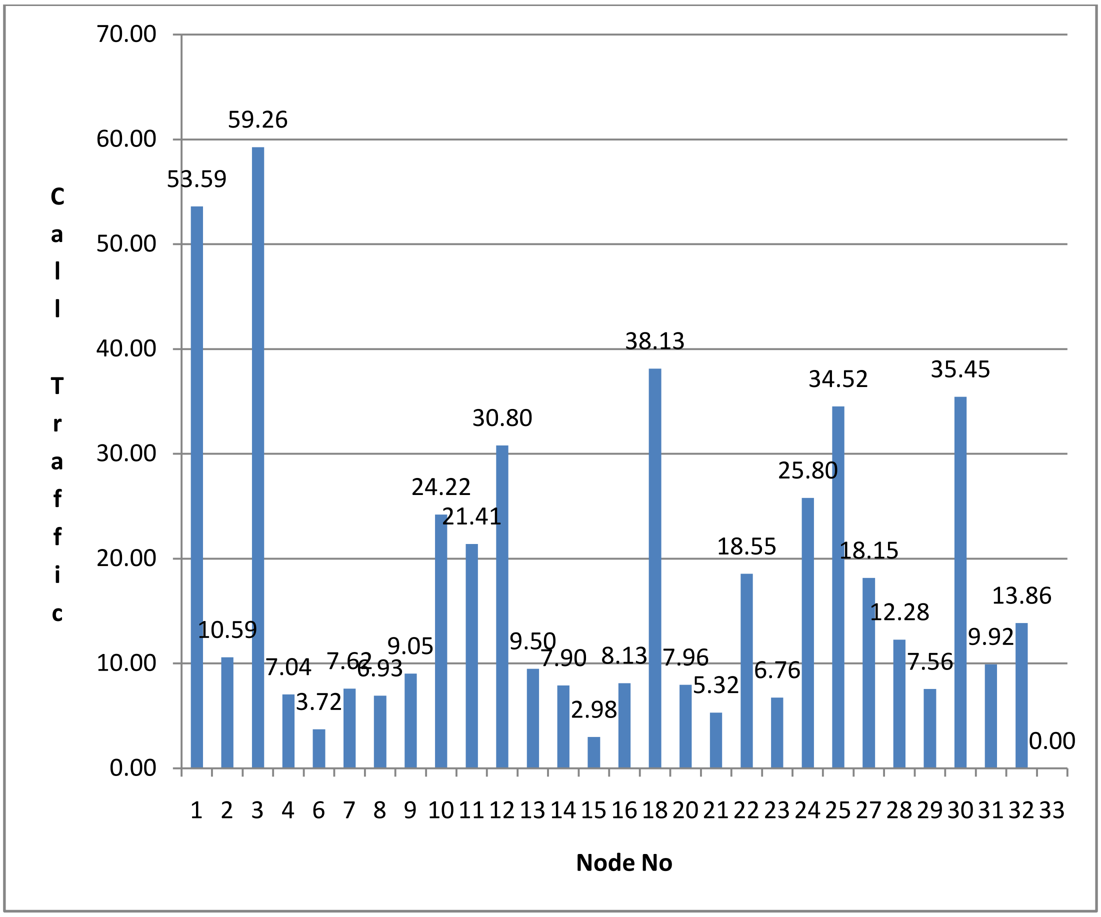


Figure 5.41 Flow from Node 33 (Bashabo Node)

Figure 5.41 describes the flow from Node 33 (Bashabo Node) to different nodes. Main features of this node are presented in Table 5.29:

Table 5.30: Main features of the flows from node 33

Major Catchment Areas	Major Destination Nodes
• Bashabo • Khilgaon • North Shahjahanpur • Sobujbagh	• Mirpur-1 Node (3) • English Road Node (1) • Kuril Bishwa road Node (18) • Motijheel Node (30) • Mouchak Node (25) • KawranBazar Node (12)

5.5 Summary

In this chapter, a case study is described taking Dhaka road network as an example. Mobile Phone data from Grameenphone Bangladesh and video data from thirteen different location of Dhaka city is collected. MySQL and MATLAB are used to process the mobile data and find out a Seed OD matrix. Video data is analyzed using TRAZER to get the traffic count. Finally MITSIMLab is used to minimize the difference between observed and actual traffic counts and find out the scaling factor and thereby the actual OD matrix.

CHAPTER SIX

CONCLUSION

6.1 Overview

The main outcome of the research is to develop a methodology to find the OD matrix using mobile phone CDR data and the result shows that this mobile phone CDR data has an enormous potential to be used as a reliable data source to estimate the OD of an area. This thesis is one of the pioneers to use CDR data in transportation field especially for determining OD matrix.

6.2 Summary of Research

In this thesis, main data source is Mobile phone CDR data collected from Grameenphone Ltd. and the supplementary data source is video recording data of thirteen different locations of Dhaka city road network.

The Dhaka city major road network has been represented using 29 nodes. Each BTS is assigned to a specific node. So if someone calls it gives its position to a definite node. If he /she moves to another node and makes another call, then his/her position is updated to that node. A trip (flow) is recorded, the first node being the origin and the second node being the destination.

The mobile phone CDR data is first analyzed using MySQL to sort and filter out the one day data. Due to computational limitations, only one day mobile phone data is used in this thesis. The reliability of the output can be increased using more data from more days (a week or over the month). MySQL assign the mobile towers to specific nodes and sort it with respect to USERID.

The sorted data is then used by MATLAB to determine the OD matrix. MATLAB is used to analyze the data to check the user wise movement from one node to another with respect to time and store it to a 29x29 matrix. This matrix is termed as Seed OD matrix.

Several others supporting data are generated using different software. Video data is analyzed by software TRAZER to calculate the traffic count. This traffic count is compared with the count output of MITSIMLab. Network of Dhaka is coded using Road Network Editor (RNE) to use it as an input to MITSIMLab.

The Seed OD matrix, road network and the traffic counts are used to run the MITSIMLab. Sensors were placed at those locations where video count data is available. The sensor output file and the count data are then used to find the optimized scaling factor using MATLAB. An optimization based algorithm is used to determine the scaling factor and derive the actual OD matrix.

6.3 Future Research

Though this research has successfully showed a methodology to determine OD matrix but it also has some limitations which need to be addressed in future research. Those are described below:

- In this thesis, only mobile phone voice call data is used. The data would be richer if the Short Message Service (SMS) details data could be included. More research can be done including those types of mobile data which could increase the reliability of the seed matrix.
- One of the major challenges of the research was the handling of the huge data at a time. Unavailability of licensed version of database software in the lab had made the analysis a challenging one. Availability of licensed software will enable using data from additional days.
- Driving behavior of the vehicles used in the software is lane based and for homogeneous traffic. So for traffic like Bangladesh, where the traffic flow is non-lane based and mixed traffic prevail in the flow, a different driving behavior should be used. In future such simulation software may be used where those facilities are available to improve the result.

- Due to unavailability of high configuration computer (RAM greater than 8GB with core i7 processor) with licensed version of windows it was not possible to handle all data by MATLAB or R (statistical software). Using those software could made the data processing easier.
- There is no proper documentation of the road network details of Dhaka such as road width, number of lanes, design speed, safe speed etc. For those thing Google map have been used. If accurate data could be available then the network would be more realistic and simulation errors will be minimized.
- If the mobile data of one month is analyzed then the OD would be more reliable. But this analysis would take several months. So in this M.Sc. thesis only one day data is used to validate the methodology.
- Initially the video of the traffic was recorded for three days (8 hours each day, around 600 hours). The calibration of the traffic count software TRAZER (used for Dhaka traffic for the first time) was found to be much more challenging than anticipated (extremely time consuming recalibration of vehicle database was required for each video clip). Therefore, for this thesis, only 4 hour video of each location on the day 15.7.12 (104 hours) is analyzed. If more video data can be analyzed than the reliability of the count data could be increased.
- A simplified network with only 29 nodes has been used in this study. More detailed network and OD representation can reduce some of the aggregation errors and make the OD more reliable.

REFERENCES

Abrahamsson, T., 1998, “Estimation of Origin-Destination Matrices Using Traffic Counts–A Literature Survey”, Approved by Joanne Linnerooth-Bayer(bayer@iiasa.ac.at) Co-Leader, Risk, Modeling and Policy

Abrahamsson, T. and Lundqvist, L., 1996, “Implementation and application of nested com-bined models for trip distribution, modal split and route choice: a comparative study of different models in the Stockholm region”, Accepted for publication (after revision) in Transportation Science.

Balakrishna, R., Antoniou, C., Ben-Akiva, B., Haris N.Koutsopoulos, H.N., and Wen, Y., 2007, “Calibration of Microscopic Traffic Simulation Models: Methods and Application”, Transportation Research Record: Journal of the Transportation Research Board, No 1999, TRB, National Research Council, Washington, D.C., pp.198-207

Bierlaire, M. and Toint, Ph.L., 1994, “Meuse: An origin-destination matrix estimator that exploits structure”, Transportation Research, 29B.

Barth, M. J., Johnston, E., and Tadi, R. R., 1996, “Using GPS Technology to Relate Macroscopic and Microscopic Traffic Parameters”, Transportation Research Record: Journal of the Transportation Research Board, No. 1520, pp. 89-96.

Bell, M., 1984, “Log-linear models for the estimation of origin-destination matrices from traffic counts”, In Proc. of the Ninth International Symposium on Transportation and Traffic Theory, Delft, The Netherlands.

Bell, M., 1991,“The estimation of origin-destination matrices by constrained generalized least squares”, Transportation Research 25B.

Boyce, D.E., Chon, K., Lee, Y. and LeBlanc L., 1983, "Implementation and computational issues for combined models of location, destination, mode and route choice", Environment and Planning A, 15.

Brenninger-Goethe, K., K. O., Jornsten, J. T. and Lundgren, 1989, "Estimation of origin/destination matrices from traffic counts using a multi-objective programming formulation", *Transportation Research*, **23B**, 257-269

Bricka, S., Zmud, J., Wolf, J., and Freedman, J., 2009, "Household Travel Surveys with GPS An Experiment", Transportation Research Record: Journal of the Transportation Research Board, No. 2105, pp. 51-56.

Calabrese, F., Lorenzo, G. D., Liu L., and Ratti, C., 2011, "Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area", Pervasive Computing, IEEE, 10(4), pp. 36-44.

Candia, J., Gonzalez, M. C., Wang, P., Schoenharl, T., Madey, G. and Barabasi, A. L., 2008, "Uncovering individual and collective human dynamics from mobile phone records", Journal of Physics A: Mathematical and Theoretical, 41(22).

Cascetta E., 1984, "Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator", Transportation Research 18B.

Cascetta, E. and Nguyen, S. 1988, "A unified framework for estimating or updating origin/destination matrices from traffic counts". Transportation Research, 22B

Chu, L., Liu, H.X., Oh, J.S., and Recker, 2003, W., "A calibration procedure for microscopic traffic simulation", IEEE, pp. 1574

Chen, Y., 1994, “Bilevel programming problems: analysis, algorithms and applications”, PhD thesis, report CRT-984, Centr de recherch sur les transports (CRT), University of Montreal, Montreal, Quebec, Canada.

Denault, L., 1994, “Etude de deux methods d'adjustement de matrices origine-destination a partir des flots des vehicules observes (in French)”, Report CRT-991, Memoire D'etudiant, Centr de recherch sur les transports (CRT), University of Montreal, Montreal, Quebec, Canada.

Di, X., Liu, H. X., and Davis, G. A., 2010, “Hybrid Extended Kalman Filtering Approach for Traffic Density Estimation Along Signalized Arterials Use of Global Positioning System Data”, Transportation Research Record: Journal of the Transportation Research Board, No. 2188, pp. 165-173.

Drissi-Katouni, O. and Lundgren, J., 1992.“Bilevel origin-destination matrix estimation using a descent approach”, Technical report LiTH-MAT-R-92-49, Department of Mathematics, Linkoping, Institute of Technology, Linkoping, Sweden.

EMME/2, 1990, “Transportation planning software, INRO Consultants Inc.”, Users' manual, Centr de recherch sur les transports, Universt de Montreal, Montreal, Quebec, Canada.

Erlander, S., Jornsten, K.O. and Lundgren, J.T. 1984, “On the estimation of trip matrices in the case of missing and uncertain data”, Technical report LiTH-MAT-R-84-20, Department of Mathematics, Linkoping University, Sweden.

Erlander, S., Nguyen, S. and Stewart, N., 1979, “On the calibration of the combined distribution/assignment model”, Transportation Research, 13B.

Friedrich, M., Immisch, K., Jehlicka, P., Otterstaetter, T., and Schlaich, J., 2010, "Generating Origin-Destination Matrices from Mobile Phone Trajectories", Transportation Research Record: Journal of the Transportation Research Board, No. 2196, pp. 93-101.

Frignani, M. Z., Auld, J., Mohammadian, A., Williams, C., and Nelson, P., 2010, "Urban Travel Route and Activity Choice Survey Internet-Based Prompted-Recall Activity Travel Survey Using Global Positioning System Data", Transportation Research Record: Journal of the Transportation Research Board, No. 2183, pp. 19-28

Fisk, C.S. and Boyce, D.E., 1983,"A note on trip matrix estimation from link traffic count data", Transportation Research 17B.

Fisk, C.S., 1988, "On combining maximum entropy trip matrix estimation with user optimal assignment", Transportation Research 22B

Fisk, C.S. 1989, "Trip matrix estimation from link traffic counts: the congested network case", Transportation Research 23B.

Florian, M., 1986, "Nonlinear cost network models in transportation analysis", Mathematical programming study, 26

Florian, M., and Chen, Y., 1993, "A coordinate descent method for the bilevel OD matrix adjustment problem", Presented at the IFORS Conference in Lisbon, Portugal, Available in an earlier version as Publication CRT-750 at the CRT, Universit de Montreal, Montreal, Quebec, Canada.

Forrest, T. L., and Pearson, D. F., 2005, "Comparison of Trip Determination Methods in Household Travel Surveys Enhanced by a Global Positioning System", Transportation Research Record: Journal of the Transportation Research Board, No. 1917, pp. 63-71.

Giaimo, G., Anderson, R., Wargelin, L., and Stopher, P., 2010, "Will It Work? Pilot Results from First Large-Scale Global Positioning System–Based Household Travel Survey in the United States", Transportation Research Record: Journal of the Transportation Research Board, No. 2176, pp. 26-34.

Grameenphone Ltd. Bangladesh. http://grameenphone.com, accessed on 15.12.2012

González, M. C., Hidalgo, C. A., and Barabási, A. L., 2008, "Understanding individual human mobility patterns", Nature, Vol. 453, pp. 779–782.

GSM System survey, Prepared by Ericson. ftp://95.173.245.83/Documents/MİKOM/MİKOM/Tezcan%20dokuman/gsm/GSMBasic/sys00r2b.pdf, accessed on 15.12.2012

Gur, Y.J., Turnquist, M., Schneider, M., LeBlanc, L. and Kurth, D., 1980, "Estimation of an origin-destination trip table based on observed link volumes and turning movements", Technical report RD-801034, FHWA, U.S. Department of Transportation, Washington D.C.

Gurusinghe, G. S., Nakatsuji, T., Azuta, Y., Ranjitkar, P., and Tanaboriboon, Y., 2002, "Multiple Car-Following Data with Real-Time Kinematic Global Positioning System", Transportation Research Record: Journal of the Transportation Research Board, No. 1802, pp. 166-180.

Hourdakis, J., Michalopoulos P.G., Kottommannil J., 2007, "Practical Procedure for Calibrating Microscopic Traffic Simulation Models", Transportation Research Record: Journal of the Transportation Research Board, No 1852, TRB, National Research Council, Washington, D.C., pp.130-139

Jan O., Horowitz, A. J., and Peng, Z., 2000, "Using Global Positioning System Data to Understand Variations in Path Choice", Transportation

Research Record: Journal of the Transportation Research Board, No. 1725, pp. 37-44.

Jornsten K. and Nguyen S. 1979, "On the estimation of a trip matrix from network data", Technical Report LiTH-MAT-R-79-36, Department of Mathematics, University of Linkoping, Linkoping, Sweden, (revised, April 1983).

Jornsten K. and Nguyen S., 1983, "Estimation of an OD trip matrix from network data: dual approaches", Technical report LiTH-MAT-R-1983-10, Department of Mathematics, Linkoping University, Sweden.

Kawakami, Lu, H. and Hirobata, Y. 1992, "Estimation of origin-destination matrices from link traffic counts considering the interaction of the traffic modes", Papers in Regional Science, 71.

Kritikal Solutions Ltd., India. http://www.kritikalsolutions.com/products/traffic-analyzer.html, accessed on 15.12.2012

Larsson and Patriksson, M., 1992,"Simplicial decomposition with disaggregated representation for the traffic assignment problem", Transportation Science, 26.

LeBlanc L.J. and FarhangianK., 1982, "Selection of a trip table which reproduces observed link flows", Transportation Research 16B.

Li, H., Guensler, R., and Ogle, J., 2005, "Analysis of Morning Commute Route Choice Patterns Using Global Positioning System–Based Vehicle Activity Data", Transportation Research Record: Journal of the Transportation Research Board, No. 1926, pp. 162-170.

Maher, M. 1983, "Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach”, Transportation Research 17B.

Nguyen, S., 1977, “Estimation of an OD matrix from network data: A network equilibrium approach”, Publication no. 60, Centre de recherche sur les transports, Universite de Montreal, Montreal, Quebec, Canada.

Nguyen, S. 1984,“Estimating origin-destination matrices from observed flows”, Transporta-tion Research, 17B.

Nguyen, S. and Dupuis C., 1983, “An efficient method for computing traffic equilibrium in networks with asymmetric transportation costs”, Publication no. 205, Centr de Recherch sur les Transports (CRT), University of Montreal, Montreal, Quebec , Canada.

Phithakkitnukoon, S., and Ratti, C., 2011, “Inferring Asymmetry of Inhabitant Flow using Call Detail Records”, Journal of Advances in Information Technology, 2 (4), pp. 239-249.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., and Ratti, C., 2010, “Activity -Aware Map: Identifying human daily activity pattern using mobile phone data”, Human Behavior Understanding, 6219(3), 14-25,Springer Berlin / Heidelberg.

Reades, J., Calabrese, F., and Ratti, C. (2009), “Eigenplaces: analyzing cities using the space-time structure of the mobile phone network”, Environment and Planning B: Planning and Design, 36(5), pp. 824-836.

Schönfelder, S., Axhausen, K. W., Antille, N., and Bierlaire, M., 2002 , “Exploring the potentials of automatically collected GPS data for travel behaviour analysis” - A Swedish data source, in J. Möltgen, and A. Wytzisk, (eds.) GITechnologien für Verkehr und Logistik, IfGIprints, 13, 155-179.

Schlaich, J., Otterstätter, T., Friedrich, M., 2010, “Generating Trajectories from Mobile Phone Data”, TRB 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies, Washington, D.C., USA.

Sevtsuk, A., and Ratti, C., 2010, “Does Urban Mobility Have a Daily Routine? Learning from Aggregate Data of Mobile Networks”, Journal of Urban Technology, 17 (1), pp. 41-60.

Sherali, H.D., Sivanandan, R. and Hobeika, A.G., 1994,“A linear programming approach for synthesizing origin-destination trip tables from link traffic volumes”, Transportation Research 28B.

Simini, F., Gonza´lez, M. C., Maritan, A., and Baraba´si, A. L., 2012, “A universal model for mobility and migration patterns”, Nature, Vol. 484, pp. 96–100,

Smith, M., 1979,“The existence, uniqueness and stability of traffic equilibria”, Transportation Research, 13B.

Snickars, F. and Weibull, J.W., 1977, “A minimum information principle, theory and practice”, Regional Science and Urban Economics, 7

Song, C, Koren, T, Wang, P, and Barabási, A. L., 2010, “Modelling the scaling properties of human mobility”, Nature Physics 6: pp. 818–823

Spiess, H., 1987, “A maximum-likelihood model for estimating origin-destination matrices”, Transportation Research 21B.

Spiess, H. 1990, “A descent based approach for the OD matrix adjustment problem”. Publication no. 693 at Centr de recherchesur les transports, Universit de Montreal, Montreal, Canada.

Spissu, E., Meloni, I., and Sanjust, B., 2011, “Behavioral Analysis of Choice of Daily Route with Data from Global Positioning System”, Transportation Research Record: Journal of the Transportation Research Board, No. 2230, pp. 96-103.

Stopher, P. and Shen, L., 2011, “In-Depth Comparison of Global Positioning System and Diary Records”, Transportation Research Record: Journal of the Transportation Research Board, No. 2246, pp. 32-37.

Stopher, P.R., and Zhang, Y., 2011, “Repetitiveness of Daily Travel”, Transportation Research Record: Journal of the Transportation Research Board, No. 2230, pp. 75-84.

Tamin, O.Z. and Willumsen, L.G. 1989,“Transport demand model estimation from traffic counts”, Transportation 16.

Tobin, R. and Friesz, T.Y., 1988,“Sensitivity analysis for equilibrium network flow”, Trans-portation Science 22.

Toledo, T., Ben-Akiva, M. E., Darda, D., Jha, M., and Koutspoulos, H. N., 2004, “Calibration of Microscopic Traffic Simulation Models with Aggregate Dara”, Transportation Research Record: Journal of the Transportation Research Board, No 1876, TRB, National Research Council, Washington, D.C., pp.10-19

Toledo, T., Koutsopoulos H.,N.,Davol, A.,Ben-Akiva, M.,E., Burghout, W., Andréasson, I., Johansson, T., Lundin, C., 2007, “Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study” Transportation Research Record: Journal of the Transportation Research Board, No 1831, TRB, National Research Council, Washington, D.C., pp.65-75

Van Vliet, D., 1982,“Saturn, a modern assignment model”, Traffic eng. control, 23: 578-581.

Van Zuylen, H. and Willumsen, L.G., 1980,“The most likely trip matrix estimated from traffic counts”, Transportation Research 14B.

Wang, J., Dixon, K. K., Li, H., and Ogle, J., 2005, “Normal Deceleration Behavior of Passenger Vehicles at Stop Sign–Controlled Intersections Evaluated with In-Vehicle Global Positioning System Data”, Transportation Research Record: Journal of the Transportation Research Board, No. 1937, pp. 120-127.

Wang, J., Dixon, K. K., Li, H., and Hunter, M., 2006, “Operating-Speed Model for Low-Speed Urban Tangent Streets Based on In-Vehicle Global Positioning System Data”, Transportation Research Record: Journal of the Transportation Research Board, No. 1961, pp. 24-33.

Wardrop, J., 1952, “Some theoretical aspects of road traffic research”, In Proc. Inst. Civil Engr., Part II, pages 325-378.

Willumsen, L.G., 1984, “Estimating time-dependent trip matrices from traffic counts”, In Proc. of Ninth international symposium on transportation and traffic theory.

Wilson, A. G., 1967, “A statistical theory of spatial distribution models”, Transportation Research, 1B: 253-269.

Wolf, J., Hallmark, S., Oliveira, M., Guensler, R., and Sarasua, W., 1999, “Accuracy Issues with Route Choice Data Collection by Using Global Positioning System”, Transportation Research Record: Journal of the Transportation Research Board, No. 1660, pp. 66-74.

Wolf, J., Oliveira, M., and Thompson, M., 2003, “Impact of Underreporting on Mileage and Travel Time Estimates Results from Global Positioning System–Enhanced Household Travel Survey”, Transportation Research Record: Journal of the Transportation Research Board, No. 1854, pp. 189-198.

Yalamanchili, L., Pendyala, R. M., Prabaharan, N., and Chakravarthy, P., 1999, “Analysis of Global Positioning System–Based Data Collection Methods for Capturing Multistop Trip-Chaining Behavior”, Transportation Research Record: Journal of the Transportation Research Board, No. 1660, pp. 58-65.

Yang, H., Iida, Y. and Sasaki, T., 1994,“The equilibrium-based Origin-Destination matrix estimation problem”, Transportation Research 28B.

Yang, H., Sasaki, T., Iida, Y. and Asakura, Y., 1992,“Estimation of origin-destination matrices from link traffic counts on congested networks”, Transportation Research 26B. S. Kawakami, H. Lu and Y. Hirobata (1992), Estimation of origin-destination matrices from link traffic counts considering the interaction of the traffic modes, Papers in Regional Science, 71.

Yang Q. and Koutsopoulos, H. N., 1996 "A microscopic traffic simulator for evaluation of dynamic traffic management systems", Transportation Research C, 4(3):113-129

Appendix

Appendix A

CDR Data Sheet Sample

- The sequence of current record: 1
    - recordType = moCallRecord [Showing that type of contacts voice call or SMS]
    - servedIMSI
        - number:470021400054488 [mobile SIM number]
    - servedIMEI
        - number:353195039571970F[mobile set number]
    - servedMSISDN
        - npi:1
        - nai:1
        - ext:1
        - number:880171*******[number who called it was replaces by an anonymous ID]
    - callingNumber
        - npi:1
        - nai:1
        - ext:1
        - number:880171*******[whom called]
    - calledNumber
        - npi:1
        - nai:2
        - ext:1
        - number:
    - translatedNumber
        - npi:1
        - nai:2
        - ext:1
        - number:
    - connectedNumber = -
    - roamingNumber
        - npi:1
        - nai:2
        - ext:1
        - number:1801080889
    - mscIncomingROUTE
        - rOUTEName:DB85
    - mscOutgoingROUTE
        - rOUTEName:CM33_MGW3002

- globalAreaID = 0x74 F0 20 00 01 FE 53
- basicService
  - teleservice:0x11
- answerTime = 2012-11-14 10:25:13+06:00
- releaseTime = 2012-11-14 10:27:11+06:00
- callDuration = 118 [duration of that call]
- causeForTerm = normalRelease
- gsm-SCFAddress = -
- serviceKey = -
- defaultCallHandling = -
- seizureTime = 2012-11-14 10:25:01+06:00 [time and date]
- setupTime = 2012-11-14 10:24:58+06:00
- alertingTime = 2012-11-14 10:25:02+06:00
- recordingEntity
  - npi:1
  - nai:1
  - ext:1
  - number:8801701000058
- location
  - locationAreaCode:0x00 01
  - cellIdentifier:0xFE 53 [BTS which serve the call each BTS has a definite location that is longitude and latitude]
- changeOfLocation = -
- transparencyIndicator = -
- changeOfService = -
- supplServicesUsed = -
- aocParameters = -
- changeOfAOCParms = -
- msClassmark = 0x53 19 A2
- changeOfClassmark = -
- radioChanRequested = dualFullRatePreferred
- radioChanUsed = halfRate
- changeOfRadioChan = -
- diagnostics
  - gsm0408Cause:144
- callReference = 0x04 2B 01 46 46 B7
- sequenceNumber = -
- additionalChgInfo
  - chargeIndicator:charge
  - chargeParameters:
- recordExtensions = -
- networkCallReference = 0x43 8E 23 00 00

- mSCAddress
  - npi:1
  - nai:1
  - ext:1
  - number:8801701000058
- cAMELInitCFIndicator = -
- fnur = -
- aiurRequested = -
- speechVersionSupported = 0x25
- speechVersionUsed = 0x25
- numberOfDPEncountered = -
- levelOfCAMELService = -
- freeFormatData = -
- cAMELCallLegInformation = -
- freeFormatDataAppend = -
- defaultCallHandling-2 = -
- gsm-SCFAddress-2 = -
- serviceKey-2 = -
- freeFormatData-2 = -
- freeFormatDataAppend-2 = -
- systemType = gERAN
- rateIndication = -
- partialRecordType = -
- guaranteedBitRate = -
- maximumBitRate = -
- modemType = -
- classmark3 = 0xC0 00
- chargedParty = callingParty
- originalCalledNumber = -
- chargeAreaCode = -
- calledChargeAreaCode = -
- mscOutgoingCircuit = 1078
- orgRNCorBSCId = 0x00 DF 00
- orgMSCId = 0x00 69 06
- callEmlppPriority = -
- callerDefaultEmlppPriority
  - value:1
- eaSubscriberInfo = -
- selectedCIC = -
- optimalRoutingFlag = -
- optimalRoutingLateForwardFlag = -
- optimalRoutingEarlyForwardFlag = -

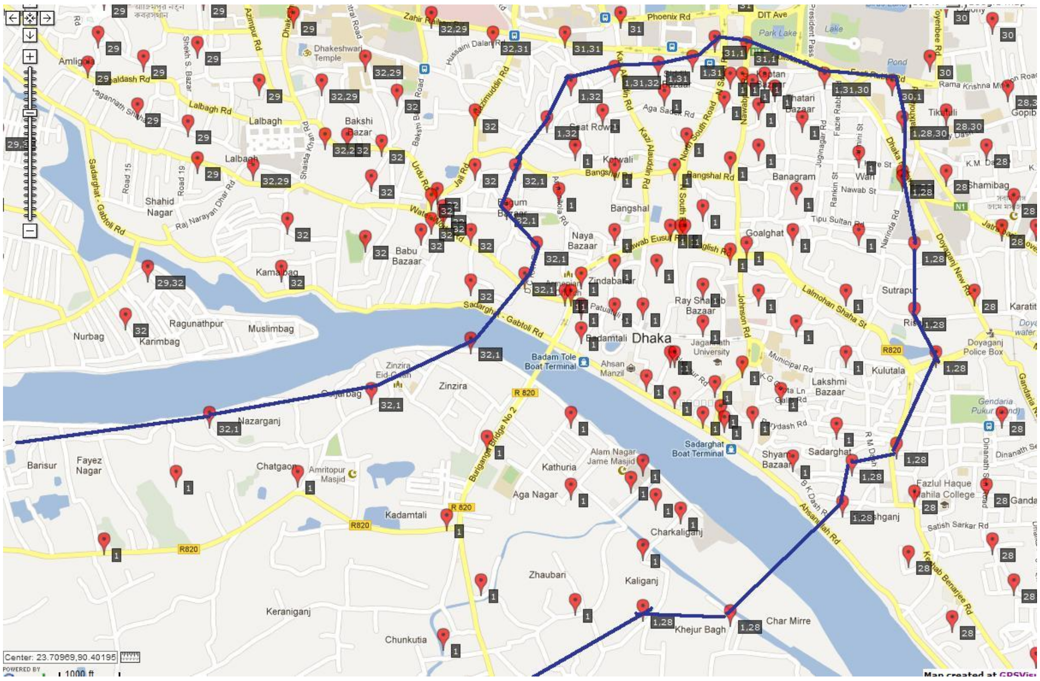
- callerportedflag = numberNotPorted
- calledIMSI
    - number:470020600181423
- changeOfglobalAreaID
        -
        - location:0x74 F0 20 00 01 FE 54
        - changeTime:2012-11-14 10:27:11+06:00
- subscriberCategory = 0x0A
- firstmccmnc = 0x74 F0 20
- intermediatemccmnc = -
- lastmccmnc = 0x74 F0 20
- cUGOutgoingAccessIndicator = notCUGCall
- cUGInterlockCode = -
- cUGOutgoingAccessUsed = -
- cUGIndex = -
- interactionWithIP = -
- hotBillingTag = -
- voiceIndicator = sendToneByOtherMsc
- bCategory = subscriberFree
- callType = outgoing
- resourceChargeIPnumber = -
- groupCallType = -
- groupCallReference = -
- uus1Type = -
- eCategory = -
- tariffCode = -
- disconnectparty = callingPartyRelease
- csReference = -
- csaReference = -
- camelphase = -
- networkOperatorId = -
- typeOfSubscribers = home
- audioDataType = audio
- userType = gsmnormaluser
- recordNumber = 1655612203
- zoneCode = -
- mctType = -
- inapFciBillingInfo = -
- redirectingCounter = -
- ussdCallBackFlag = -
- cAMELDestinationNumber = -
- chargePulseNum = -

- osssServicesUsed = -
- partyRelCause = -
- chargeLevel = chargeBySecond
- locationNum = -
- locationNumberNai = -
- dtmf-indicator = -
- b-ch-number = 1
- ncnpFlag = -
- cARP = -
- accountcode = -
- channelmode = hscsd-nochannel-mode
- channel = hscsd-no-chosen-channel
- specialBillPrefix = -
- calledportedflag = numberNotPorted
- locationroutingnumber = -
- routingcategory = -
- intermediateChargingInd = sINGLE-BILL
- calledIMEI = -
- mscOutgoingROUTENumber = -
- mscIncomingROUTENumber = -
- roDefaultCallHandling = -
- roLinkFailureTime = -
- lastSuccCCRTime = -
- drcCallId = -
- drcCallRN = -
- wPSCallFlag = -
- chargePulses = -
- voBBUserFlag = -
- followMeInd = -
- invokeOfLCLS = -
- vasType = -
- officeName = -
- scpConnection = -
- chargeClass = -
- npa-Nxx = -
- globalCallReference = -
- callRedirectionFlag = -
- nPDipIndicator = -
- aNSIRoutingNumber = -
- lRNSource = -
- callerIPInformation = -

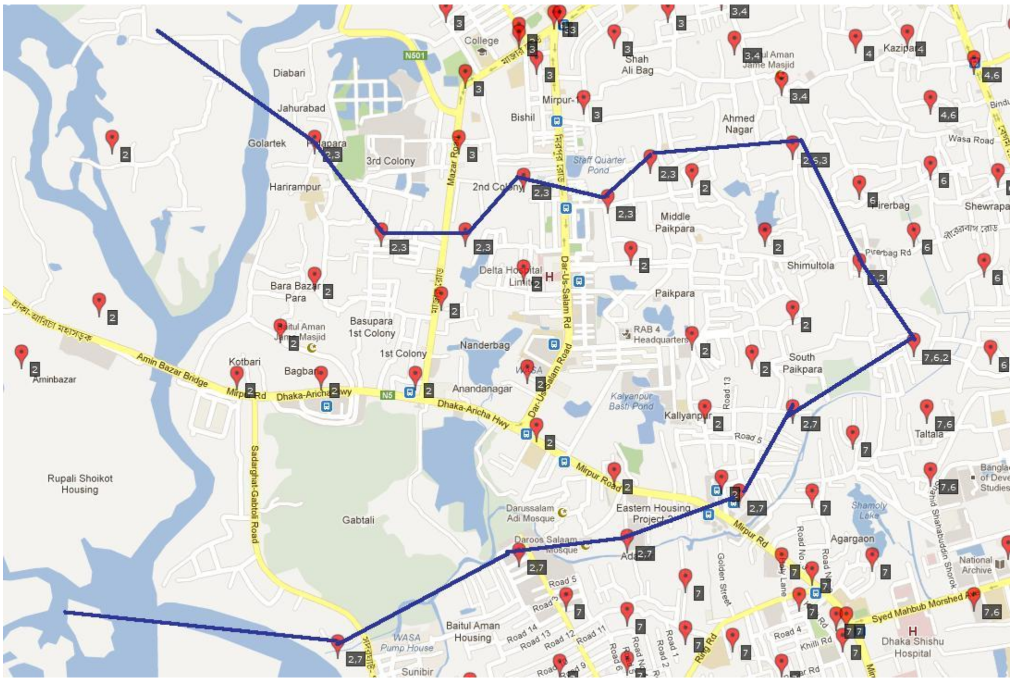
Appendix B

Catchment areas of different nodes

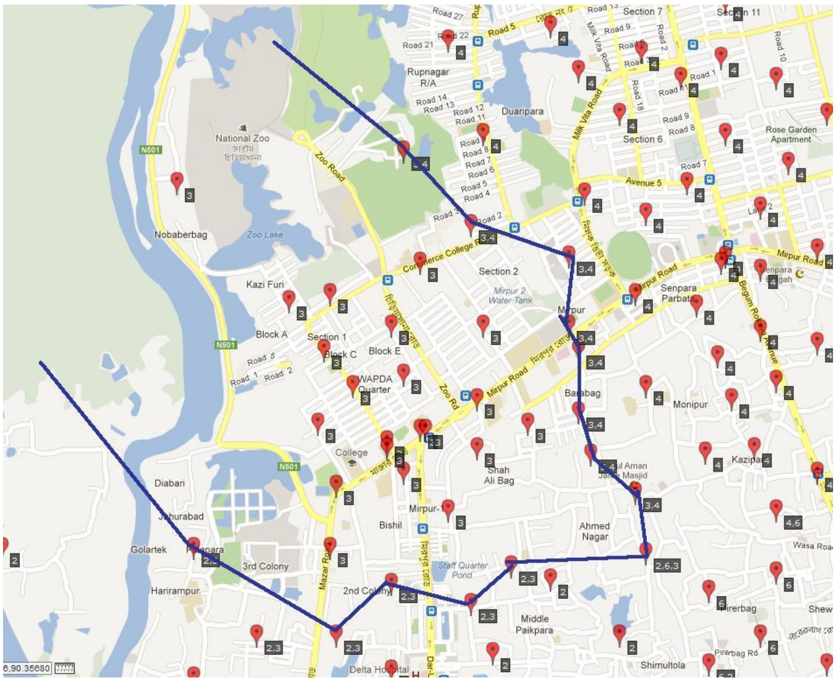
Node 1:



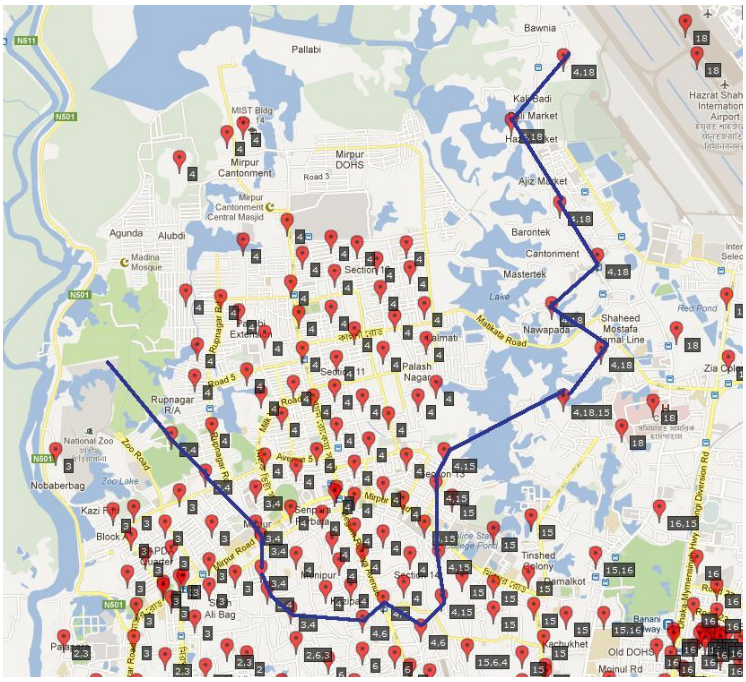
Node 2:



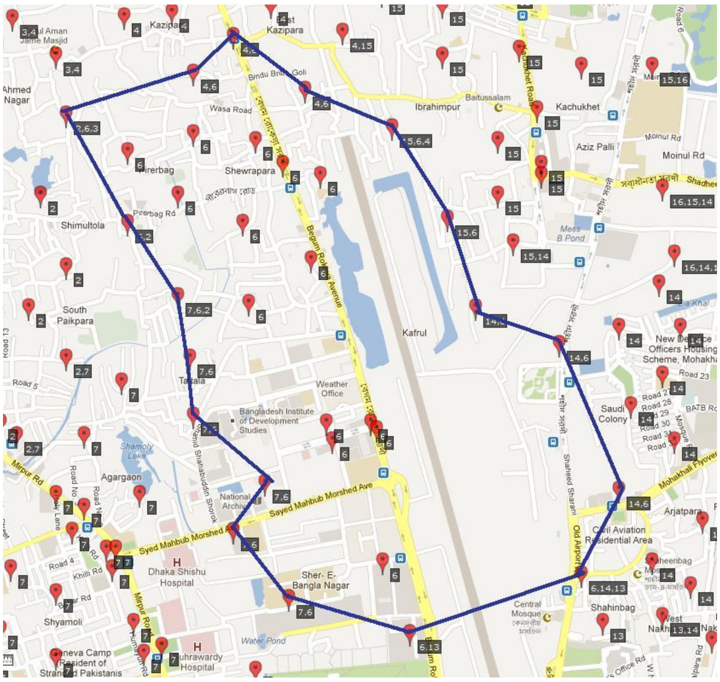
Node 3:



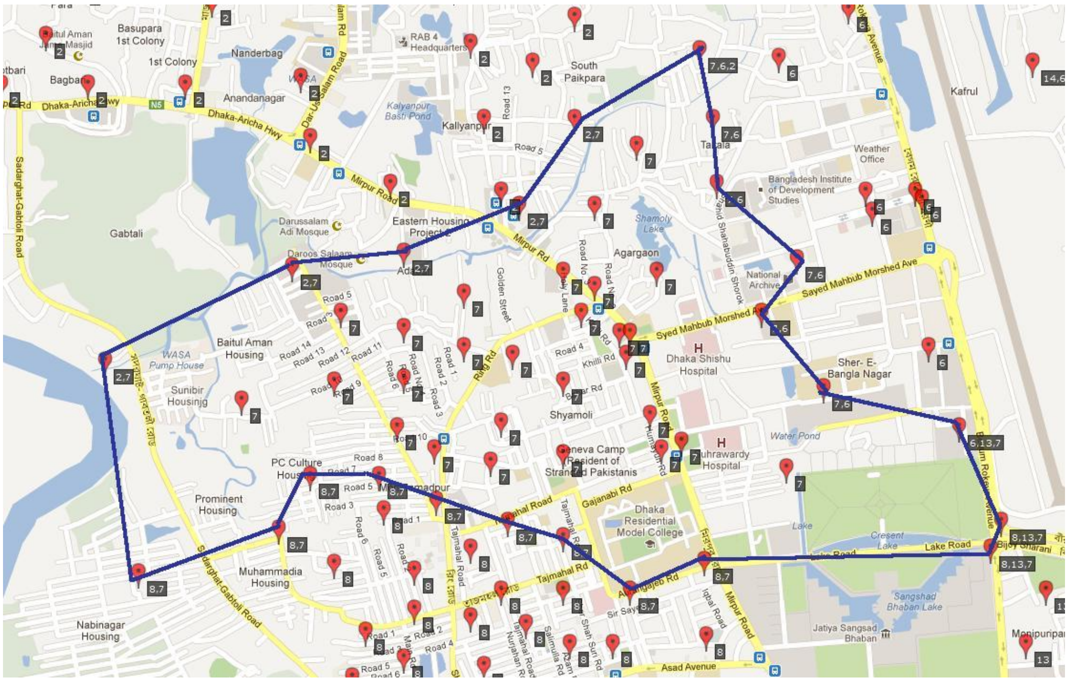
Node 4:



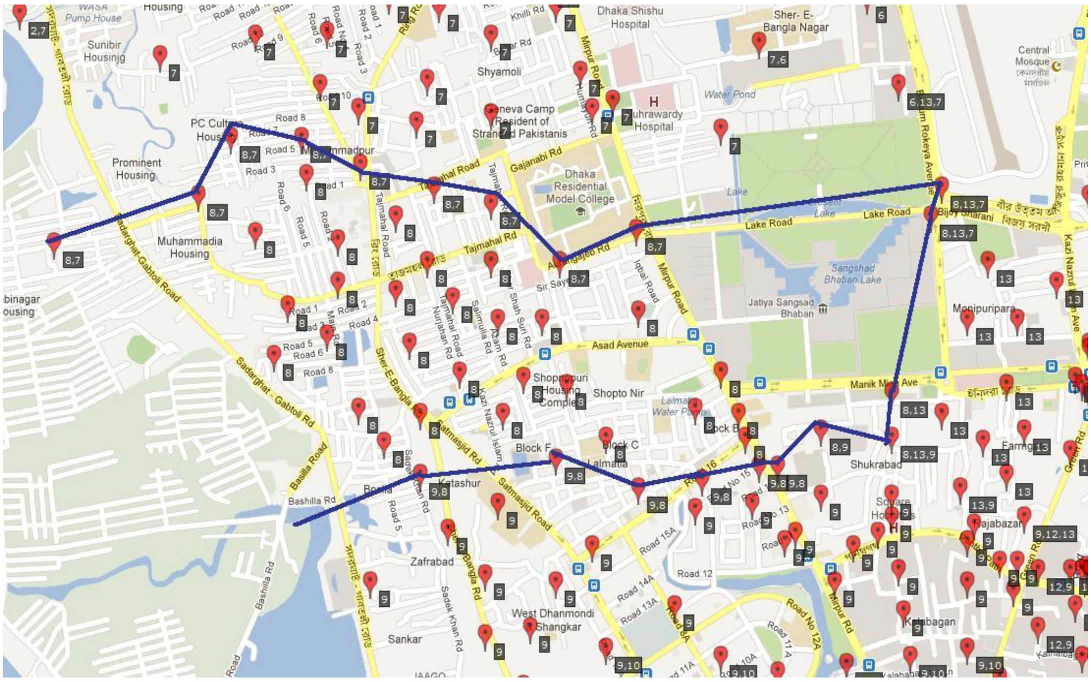
Node 6:



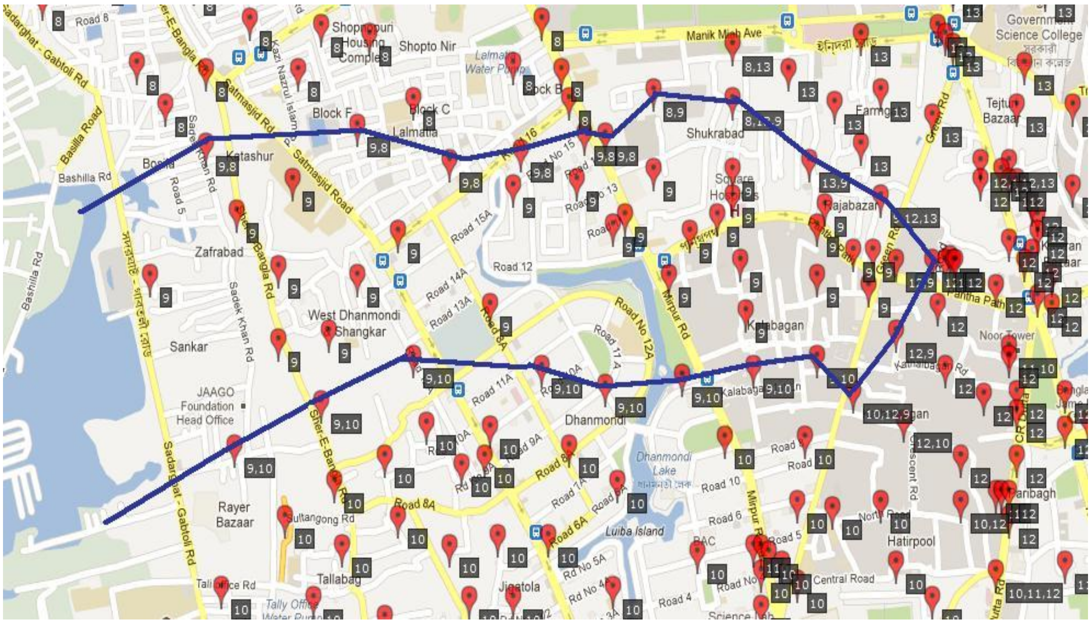
Node 7:



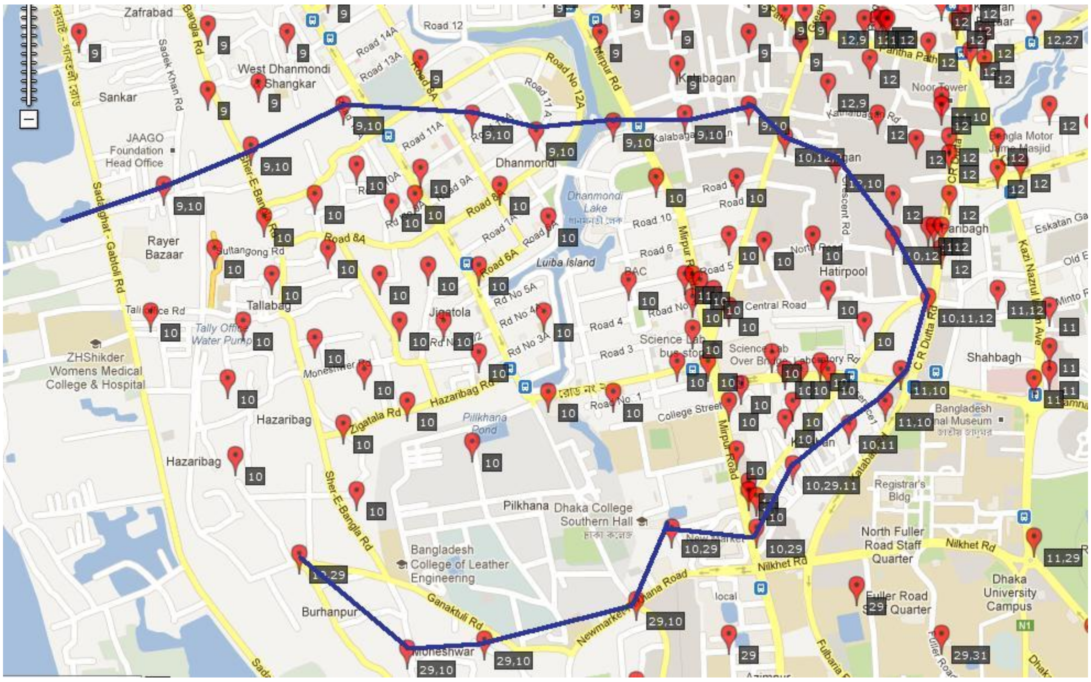
Node 8:



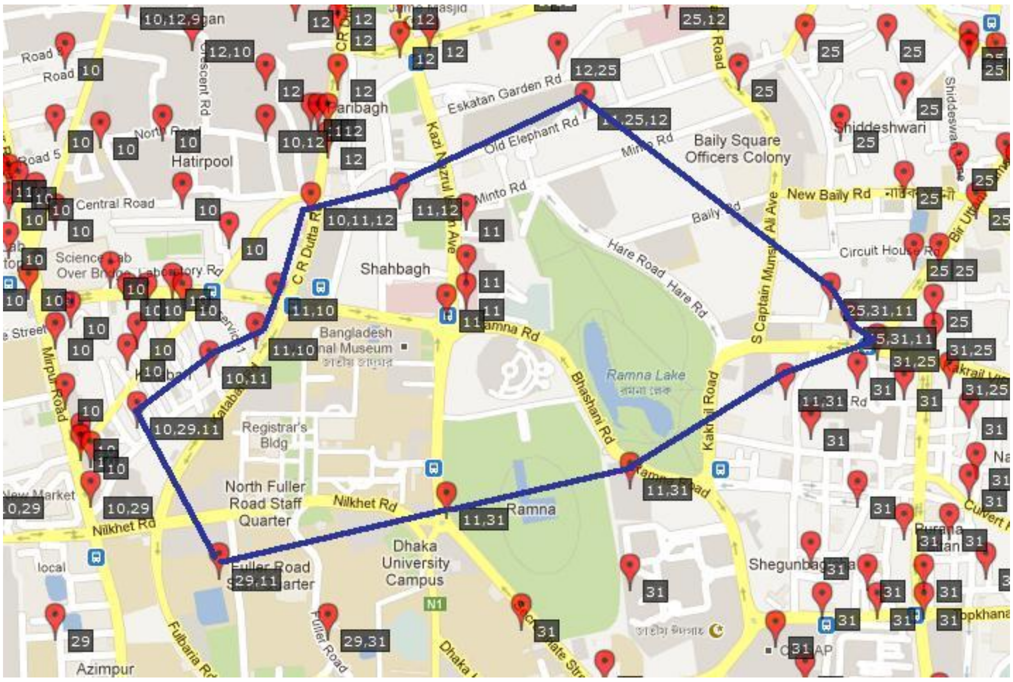
Node 9:



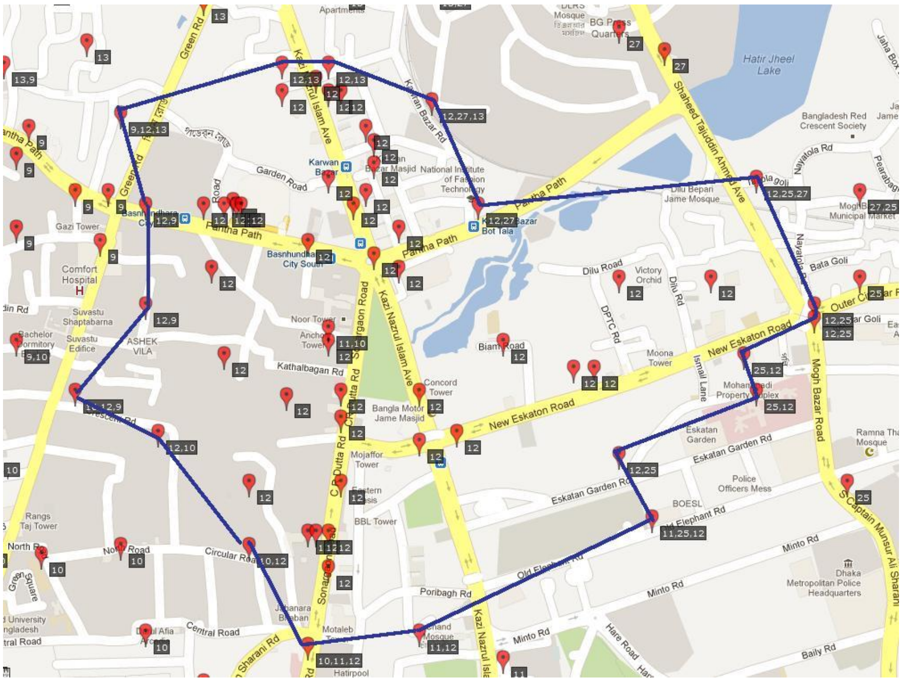
Node 10:



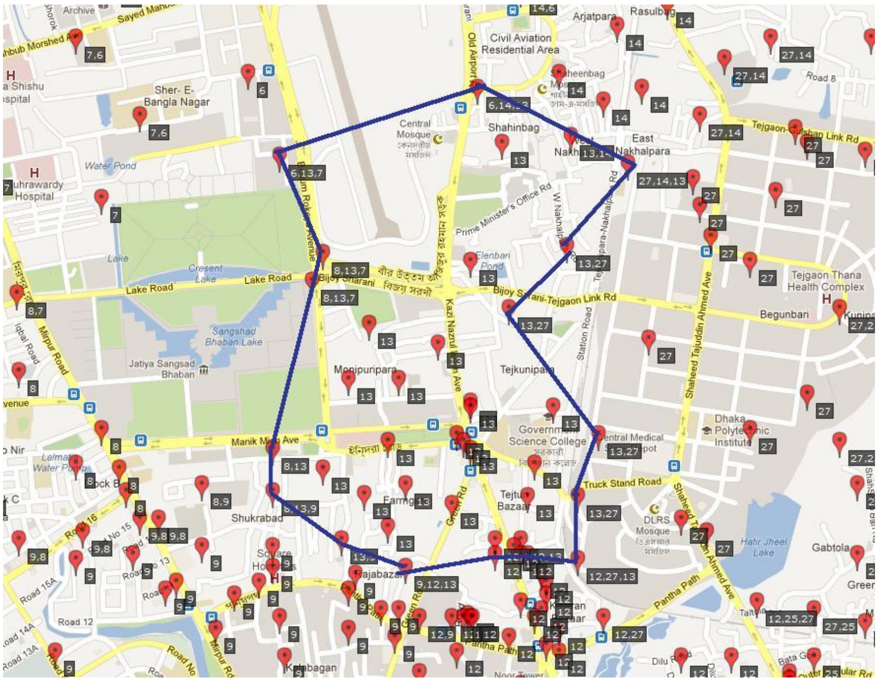
Node 11:



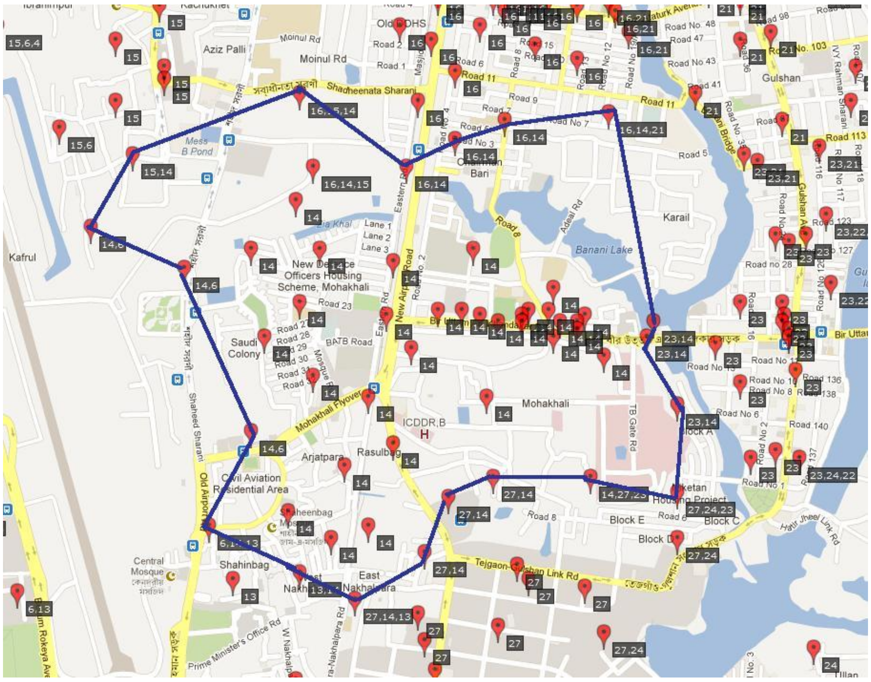
Node 12:



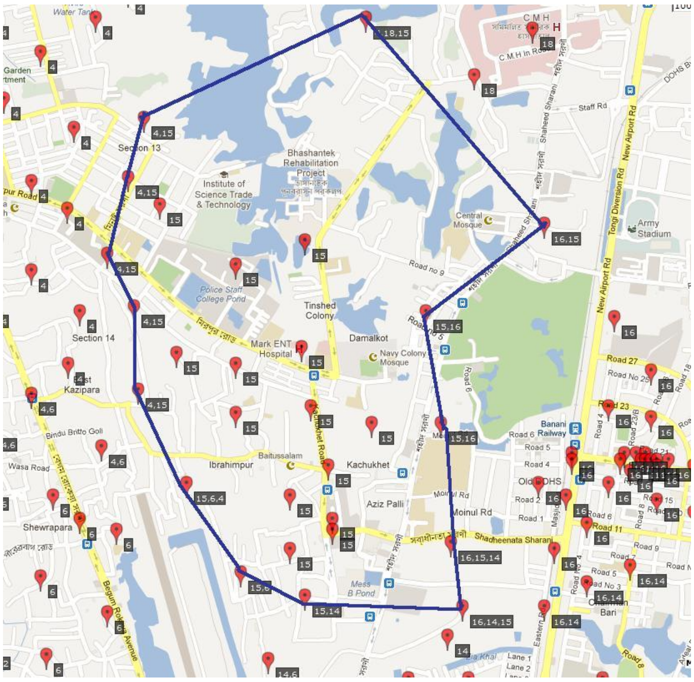
Node 13:



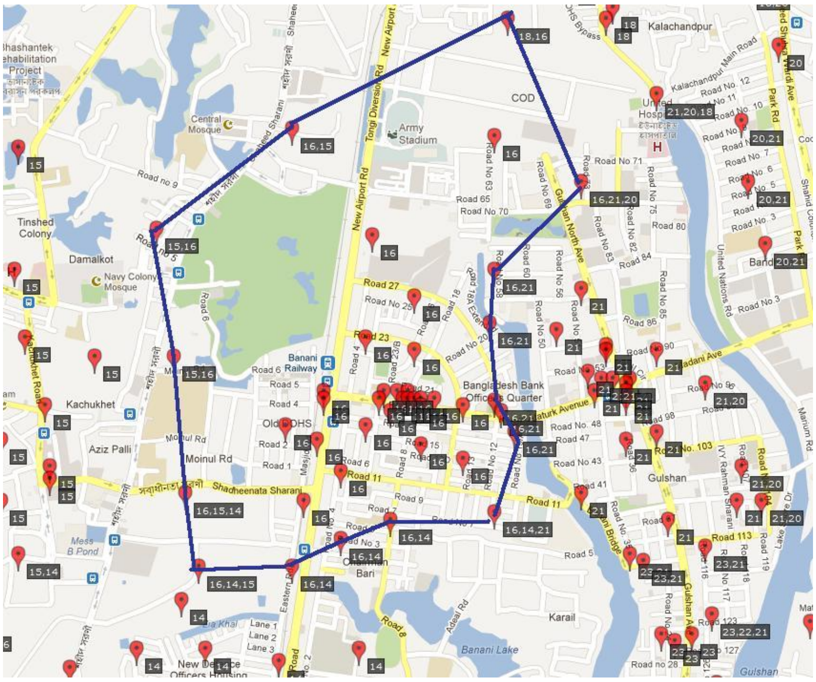
Node 14:



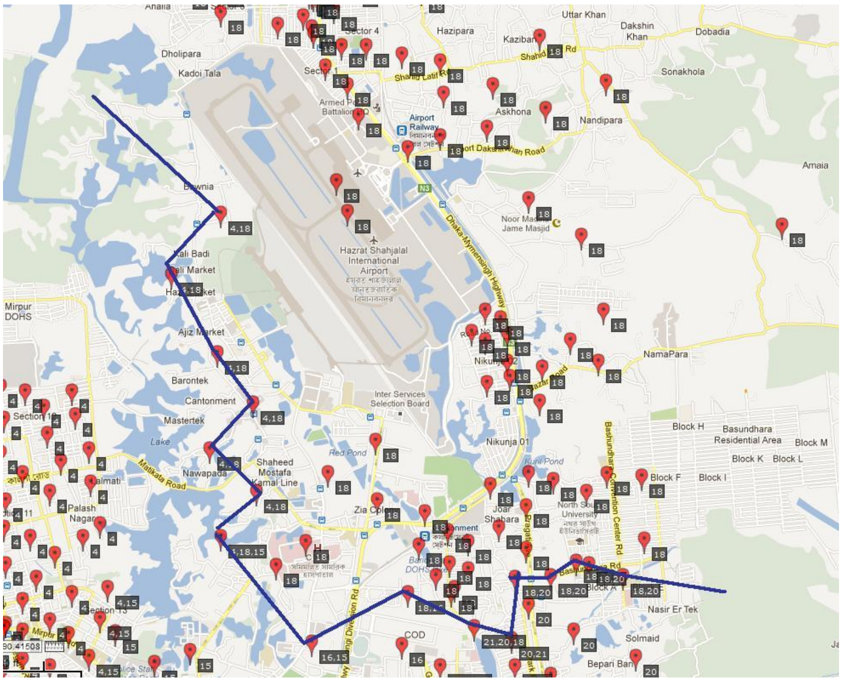
Node 15:



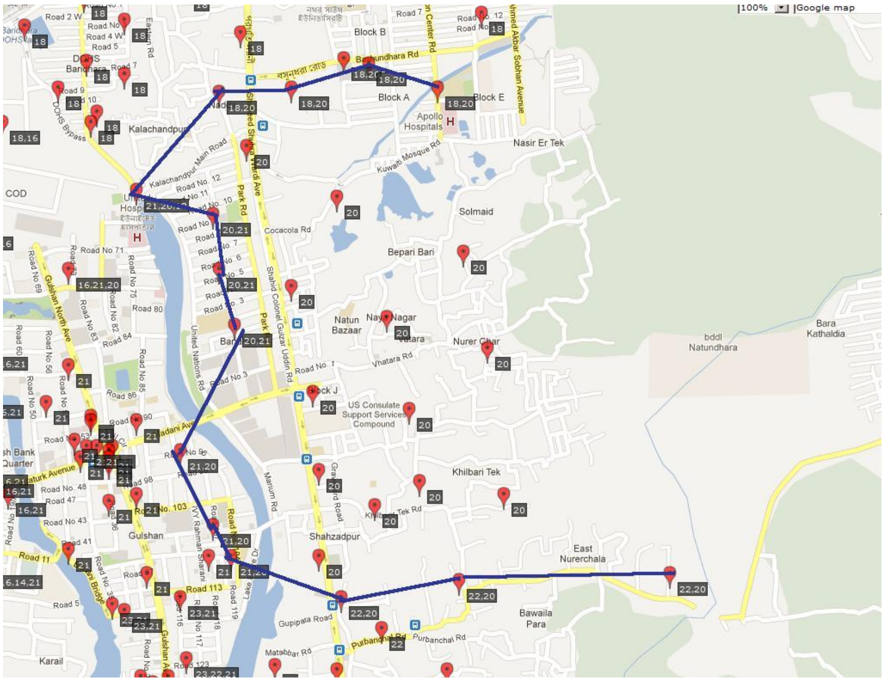
Node 16:



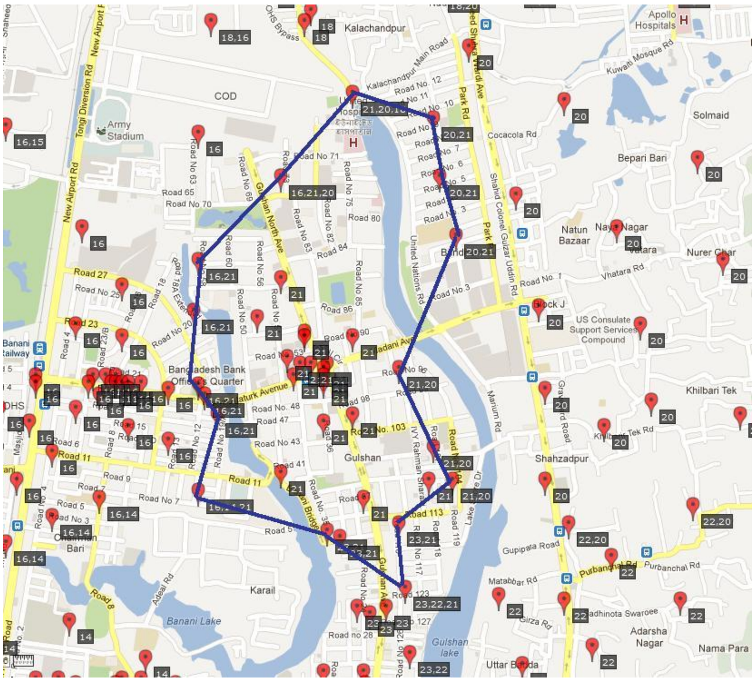
Node 18:



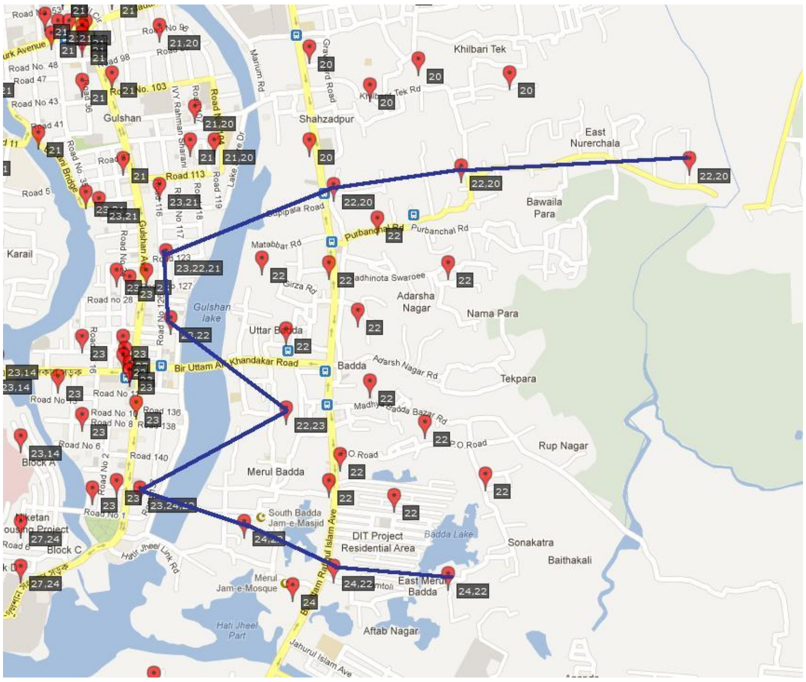
Node 20:



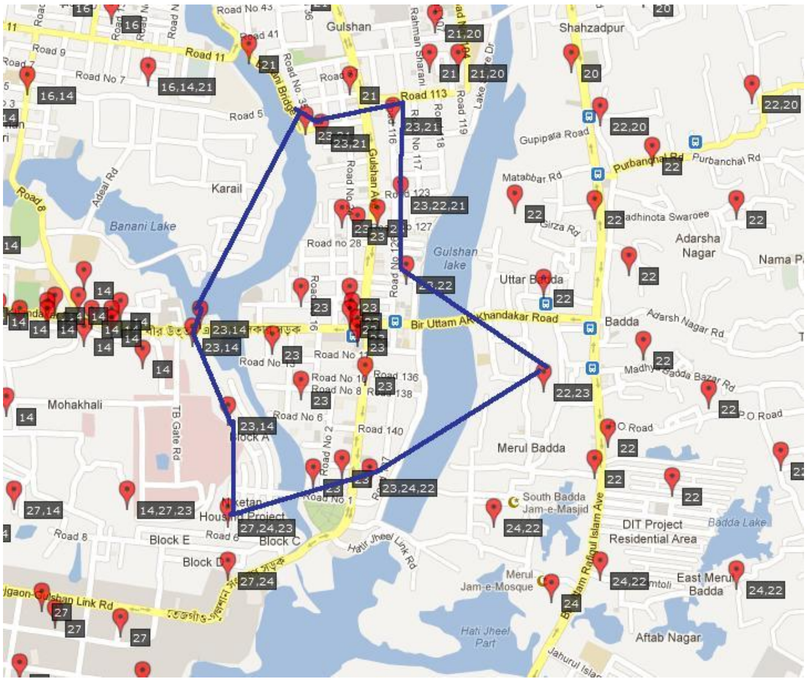
Node 21:



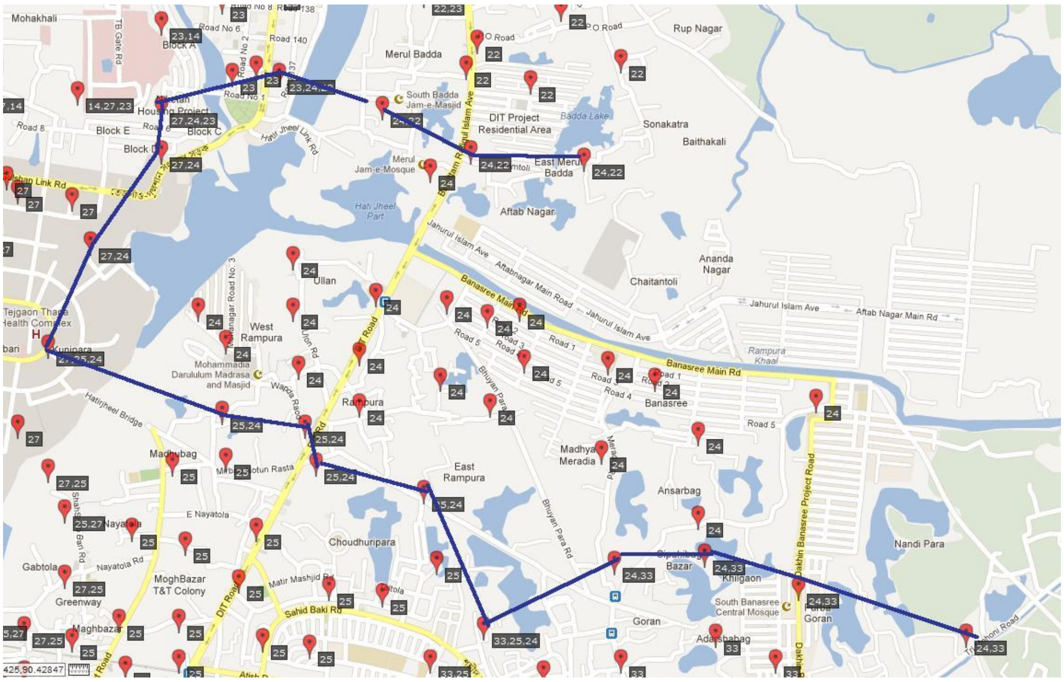
Node 22:



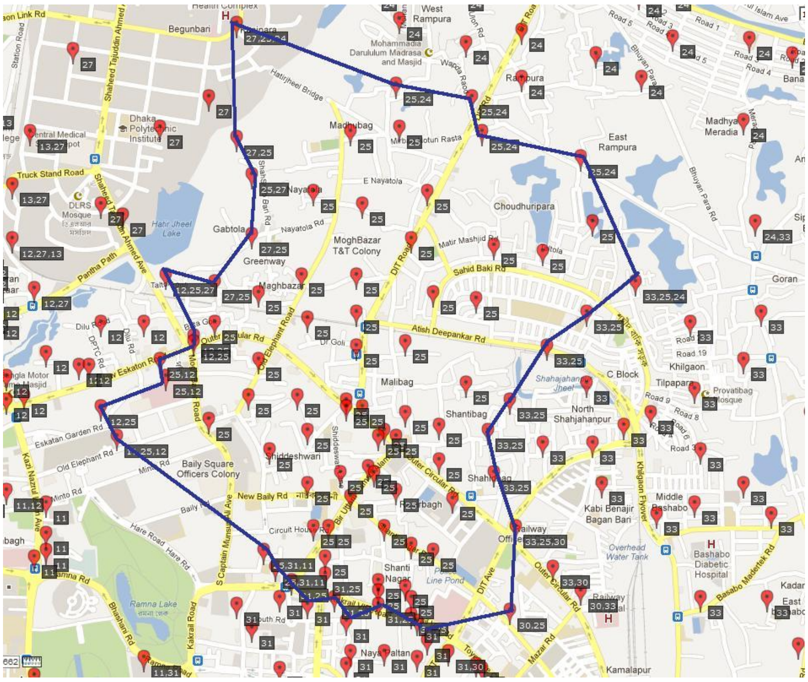
Node 23:



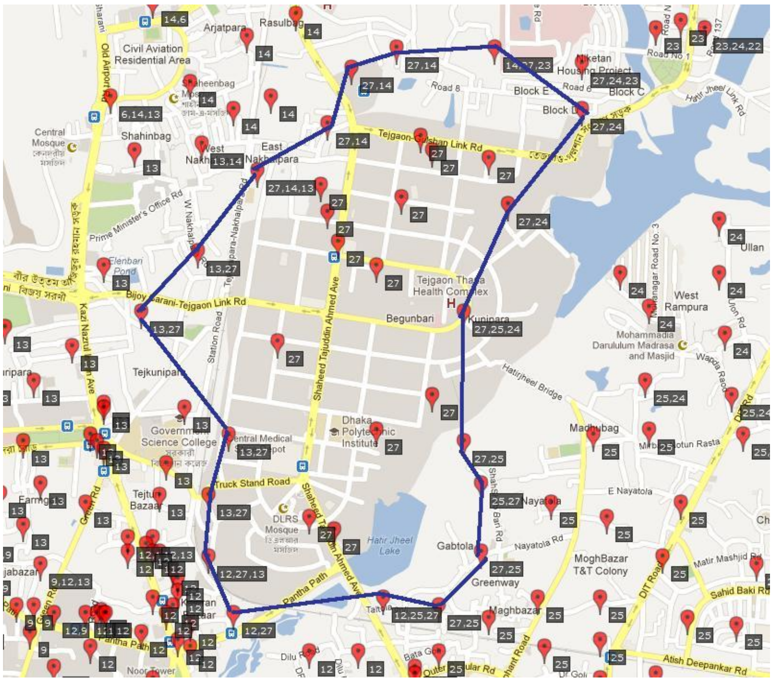
Node 24:



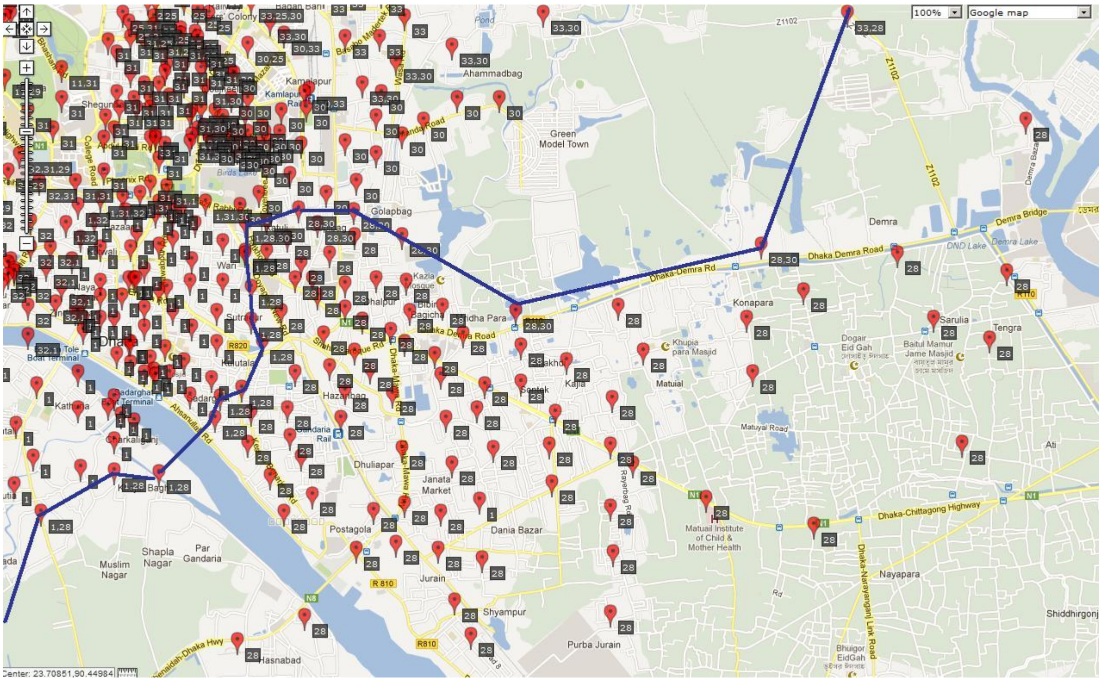
Node 25:



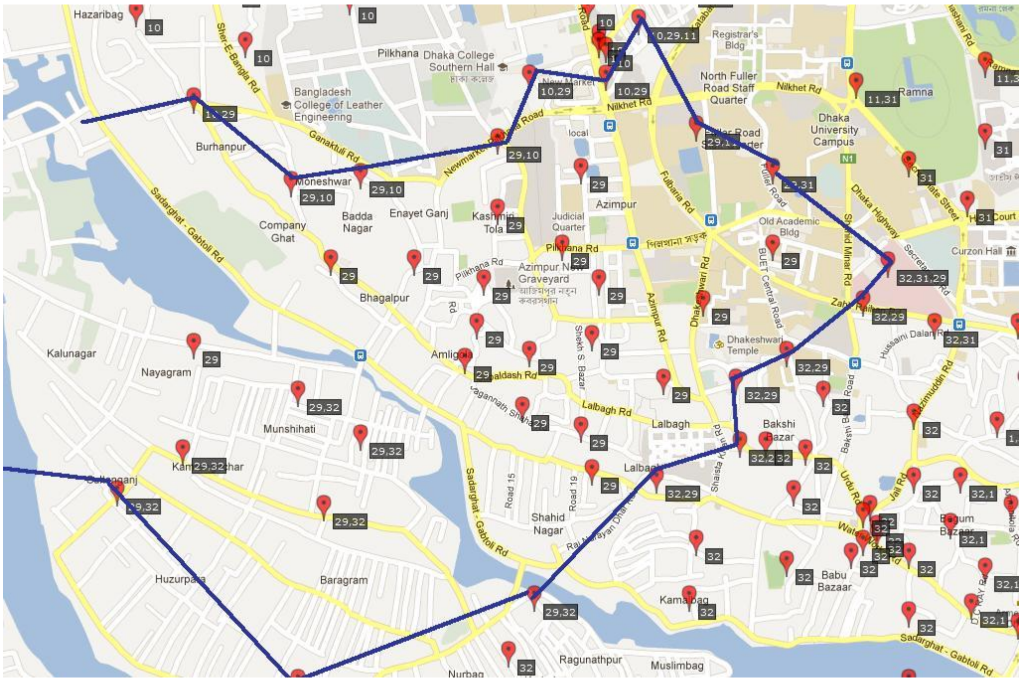
Node 27:



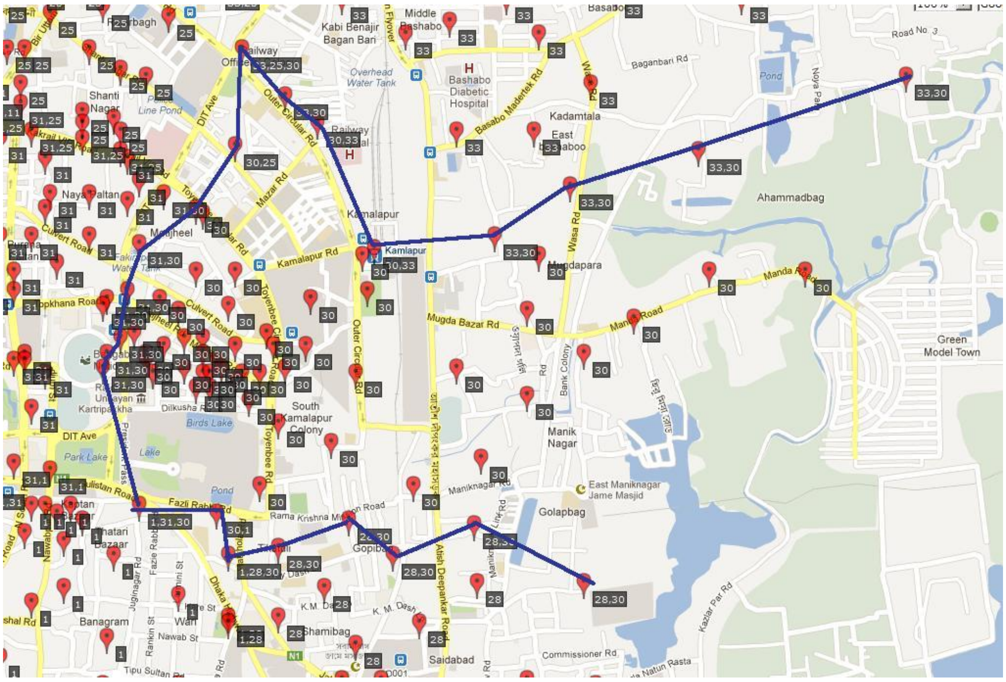
Node 28:



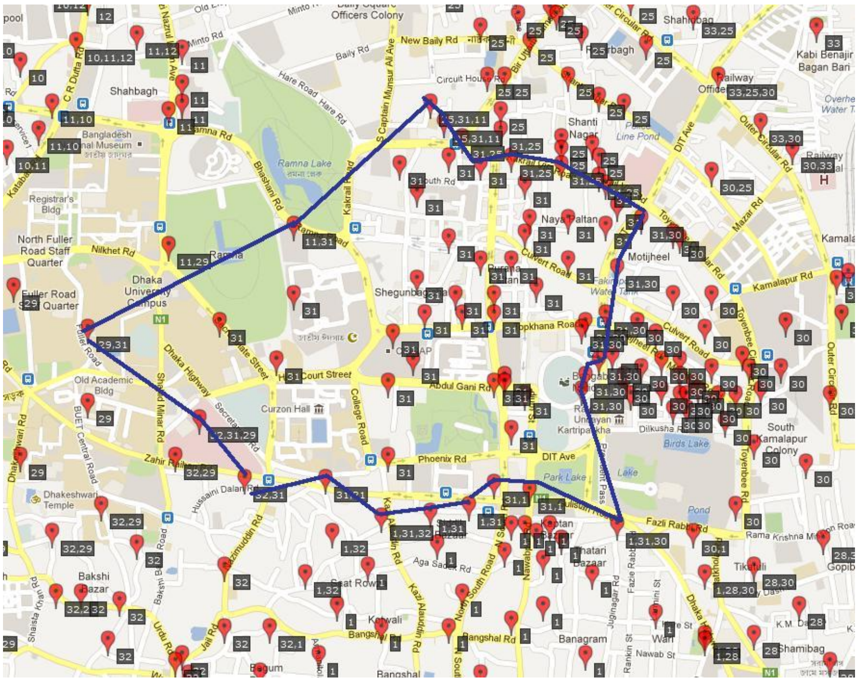
Node 29:



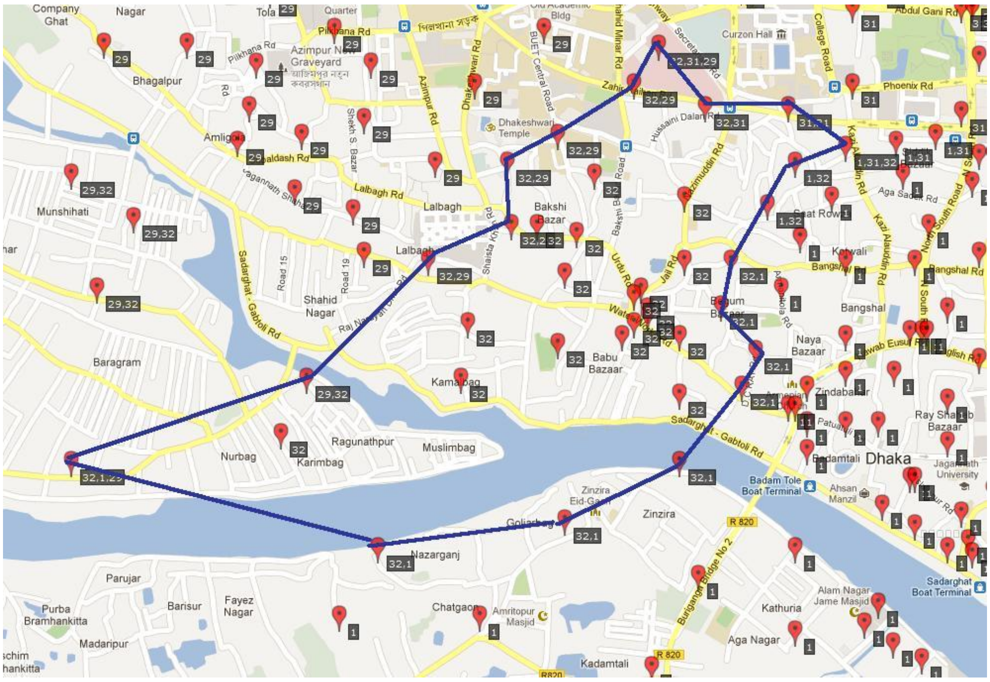
Node 30:



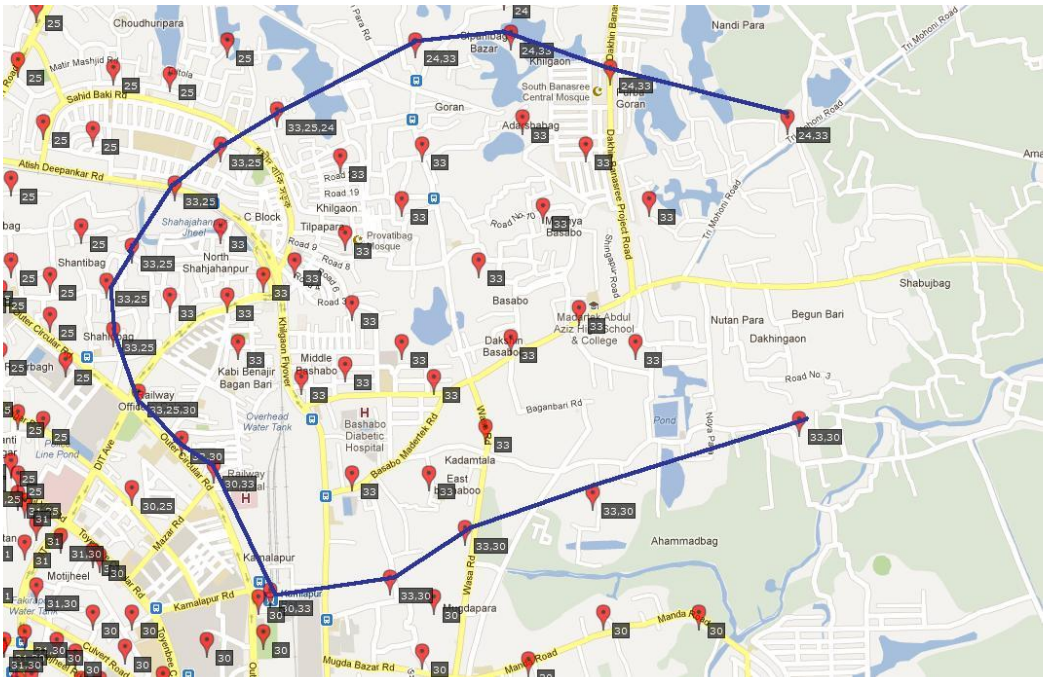
Node 31:



Node 32:



Node 33:



Appendix C

Analysis Codes

<u>C1: Mysql codes:</u>

```
# copy one data to a new table temp1 from the full-data table
CREATE TABLE temp1 LIKE fulldata;
INSERT temp1
SELECT * FROM testdata
where CDATE = '20120715' ;


#create new row for nodes and time_second
ALTER TABLE temp1 ADD (time_second int, node1 int, node2 int, node3 int);


#insert nodes form tower table which contains the nodes corresponding to each tower
update temp1,tower
set
temp1.node1 = tower.Node1,
temp1.node2 = tower.Node2,
temp1.node3 = tower.Node3
where (temp1.LAT = tower.LAT) and (temp1.LON = tower.LON);

update temp1 set time_second = 0;
update temp1 set time_second = (TIME_TO_SEC(CTIME));


#Delete unnecessary columns
alter table temp1
drop CTIME,
drop DURATION,
drop LAT,
drop LON;
```

C2: Matlab Code

```
fid=fopen('July15_C_order.csv');
fid1=fopen('new.csv');
tline = fgetl(fid1);
fclose(fid1);
ww=0;
h=zeros(33,33);
%x=1;
%y=[];
for aa=1:34,
z=[];
c=0;
i=0;
j=0;


a=tline(1:24);
for j=1:1000000,
    tline = fgetl(fid);
    i=i+1;
    if strfind(tline,a)>0
        z(i,1)=c;
    else
         c=c+1;
        z(i,1)=c;
        a=tline(1:24);
    end;
end;
ww=aa*1000000;
b=dlmread('July15_C_order.csv',',',[(ww-1000000),2,(ww-1),5]);
m=[b(:,1:4) z(:,1)];

clear z;
clear b;

%h=[];


j=1;
p=[];
q=[];
```

```
f=0;
g=0;
ag2= size(m);
p(1,:)=m(1,:);
for i=2:(ag2(1,1)),
if m(i,5)== m(i-1,5)
    j=j+1;
    p(j,:)= m(i,:);
else
    ag1=size(p);
if ((ag1(1,1))>1)
q = sortrows(p);
ag=size(q);
    %to eliminate second and third node
    l=1;
    r=[];
for k=1:((ag(1,1))-1),
    if  q(k,3)== 0
        r(l,1)= q(k,2);
        l=l+1;
    else
    end;
end;
o= unique(r);
clear r;
if (length(o))> 1;
for k=1:((ag(1,1))-1),
    if ((q(k,3)> 0) && (q(k,4)> 0) && ((q(k,2)== o(1,1))||(q(k,3)== o(1,1))||(q(k,4)==
o(1,1))))
         q(k,2)= o(1,1);
         elseif (q(k,3)> 0 && q(k,4)> 0 && (q(k,2)== o(2,1)||q(k,3)== o(2,1)||q(k,4)==
o(2,1)))
         q(k,2)= o(2,1);

    elseif (q(k,3)> 0 && q(k,4)== 0 && (q(k,2)== o(1,1)||q(k,3)== o(1,1)))
       q(k,2)= o(1,1);
    else (q(k,3)> 0 && q(k,4)== 0 && (q(k,3)== o(2,1)||q(k,2)== o(2,1)))
       q(k,2)= o(2,1);

    end;
end;
end;
```

```
    clear o;

    %end of node elimination

        for k=1:((ag(1,1))-1),
            if q(k,2)== q(k+1,2)
            else
                %y(x,:)= [q(k,1:2) q(k,5)];
                %y(x+1,:)=[q(k+1,1:2) q(k+1,5)];
                %x=x+2;
                f = q(k,2);
                g = q(k+1,2);
                h(f,g) = (h(f,g)+1);
            end;
        end;
                        else
end;
        p=[];
        q=[];
        j=1;
        p(1,:)= m(i,:);
end;
end;
end;
xlswrite('ODflow',h);
%xlswrite('ODflow_ID',y);
fclose(fid);
```

Appendix D

Combined Analysis Codes

MAIN CALIBRATION.m

```
% This is the main calibration program
echo off;
global COMMAND;
global mainoutput;
flag_simultaneous = 1;
mainoutput = fopen('OutputLog.out','w');
% Input by users
COMMAND = mitsim %input('Command to be executed :','s')
no_days = 1  %input('Days of Data available :')
n_parameter = 4 %input('Max. no. of iterations for parameter calibration :')
tol_parameter = .01   %input('Tolerance allowed for parameter calibration : ')
% Writing the input eneter by user to a Log file
fprintf(mainoutput,'File written on %s\n ',datestr(now));
fprintf(mainoutput,'Command executed : %s\n ', COMMAND);
fprintf(mainoutput,'Days of Data available : %d\n ', no_days);
fprintf(mainoutput,'Max. no. of iterations for parameter calibration : %d\n ',
n_parameter);
fprintf(mainoutput,'Tolerance allowed for parameter calibration : %f\n ',
tol_parameter);
fprintf(mainoutput,'\n ');

% Start of Calibration Program

w = 1;
flag_parameter = 0;
toler_parameter = Inf;

while  flag_parameter < n_parameter & toler_parameter > tol_parameter
  [y1, line_no] = INITIALPARAMETERS; % getting initial set of parameters
  PARAMETERBOX; % Box algorithm
  [y2, line_no] = INITIALPARAMETERS;
  toler_parameter = max(abs((y1 - y2)./(H-G)'));
  fprintf(mainoutput,'Parameter Tolerence : %f\n ', toler_parameter);
  flag_parameter = flag_parameter + 1;
end
fclose(mainoutput);
```

INITIALPARAMETERS.m

```
function [y, line_no] = INITIALPARAMETERS()
% This function returns the inital parameter set.

fid = fopen('od.dat','r');

 y=[];
 line_no = [17 172];

 for i=1:17
    tline = fgetl(fid);
 end;

 y(1)= str2num(tline(11:length(tline)));

  for i=18:172
    tline = fgetl(fid);
 end;

 y(2)= str2num(tline(11:length(tline)));
 fclose(fid);
```

PARAMETERBOX.m

```
% This program sets the input for BOX algorithm and then calls the BOX algorithm
global X N M K G H R F XC ITMAX IC IPRINT ALPHA BETA GAMMA
DELTA;
global mainoutput;
fprintf(mainoutput,'Box algorithm called \n');
INITIAL
M  = N; % No. of implicit constraints

X = zeros(K, M); % K sets of parameters
R = rand(K,N); % Random numbers for generating the sets of parameters
F = zeros(K,1); % Function value associated with each set
XC = zeros(N,1); % Centroid of the points
% Box algorithm
BOX(N,M,K,ITMAX,ALPHA,BETA, GAMMA,DELTA,X,R,F,G,H,XC,no_days);
```

INITIAL.m

```
N = 2;          % No. of parameters to calibrate
K = 3;         % No. of initial points (min N+1)
ITMAX = 5;     % No. of iterations
ALPHA = 1.3;    % Reflection factor
BETA = 0.01;    %Convergence Criteria
GAMMA = 2;      % Covergence Criteria
DELTA = 0.01; % Explicit constraint violation correction
G = [ 0; 0]; % Lower bound of parameters
H = [ 0.01; .1]; % upper bound of parameters
```

BOX.m

```
function BOX(N,M,K,ITMAX,ALPHA,BETA,
GAMMA,DELTA,X,R,F,G,H,XC,no_days)
% This function impleiments the Box algorithm
global COMMAND;
global mainoutput;
tt = 0;
IT = 1
KODE = 0

if (M-N) > 0
   KODE = 1;
end;
for i = 2:K
   X(i,:) = 0;
end;

% Calculate complex points and check against constraints
fprintf(mainoutput,'Inital set of parameters\n');
K1 = 1;
I = 1;
% getting inital set of parametr values
[X(1,:) line_no] = INITIALPARAMETERS;
X(1,:) = CHECK(N,M,K,X,G,H,I,KODE,XC,DELTA,K1);
fprintf(mainoutput,'%f  ',X(1,:));
fprintf(mainoutput,'\n');

% generating (K-1) random sets of parameters
for II =2:K
```

```
    for J = 1:N
        I = II;
        X(II,J) = G(J) + R(II,J)* (H(J) - G(J));
    end;
    K1 = II;
    X(I,:) = CHECK(N,M,K,X,G,H,I,KODE,XC,DELTA,K1);

    fprintf(mainoutput,'%f  ',X(I,:));
    fprintf(mainoutput,'\n');
end
K1 = K;
fprintf(mainoutput,'Inital function values\n');
% GETTING OBJECTIVE VALUE FOR EACH SET OF PARAMETERS
for I=1:K,parameter_set = X(I,:);
    CHANGEPARAMETERS(parameter_set,line_no);
    F(I) = FUNC(N,M,K,X,F,I,line_no,no_days);
    fprintf(mainoutput,'%f  ',F(I));
    fprintf(mainoutput,'\n');
end

KOUNT = 1;
IA = 0;

flag_IEV1 = 1;
Flag_KOUNT = 1;
while (IT - ITMAX) <= 0 & Flag_KOUNT == 1
    fprintf(mainoutput,'Start of Iteration number  %d\n', IT);
    IT = IT + 1;

    % Find Point with Lowest Function Value
    IEV1 = 1;
    for ICM = 2:K
        if (F(IEV1) - F(ICM)) > 0
            IEV1 = ICM;
        end;
    end;
    IEV1

    % Find Point with Highest Function Value

    IEV2 = 1;
    for ICM = 2:K
```

```
    if (F(IEV2) - F(ICM)) <= 0
        IEV2 = ICM;
    end;
end;
IEV2

Flag_KOUNT = 1;
if (F(IEV2) - (F(IEV1)*(1+ BETA))) >= 0
    KOUNT = 1;
else
    KOUNT = KOUNT + 1
    if (KOUNT - GAMMA) < 0
        Flag_KOUNT = 1;
    else
        Flag_KOUNT = 0;
    end
end

% Replace point with highest function value
if Flag_KOUNT == 1
    XC = CENTR(N,M,K,IEV2,I,XC,X,K1); % Centroid of remaining set
    % Applying correction (correction 1) to the highest point
    for JJ= 1:N
        X(IEV2,JJ) = (1.0 + ALPHA)*(XC(JJ)) - ALPHA*(X(IEV2,JJ));
    end
    newX = X(IEV2,:)
    I = IEV2;
    X(I,:) = CHECK(N,M,K,X,G,H,I,KODE,XC,DELTA,K1)
    fprintf(mainoutput,'Iteration %d\n',(IT-1));
    fprintf(mainoutput,'%f  ',X(I,:));
    fprintf(mainoutput,'\n');
    parameter_set = X(I,:);
    CHANGEPARAMETERS(parameter_set,line_no);
    F(I) = FUNC(N,M,K,X,F,I,line_no,no_days)
    fprintf(mainoutput,'Function value: %d\n',F(I));

    % REPLACE NEW POINT IF IT REPEATS AS HIGHEST FUNCTION
VALUE
    flag_IEV2 = 1;
    while  flag_IEV2 == 1 &   (IT - ITMAX) <= 0  & Flag_KOUNT == 1
        IEV1 = 1;
        for  ICM = 2:K
```

```
        if (F(IEV1) - F(ICM)) <= 0
            IEV1 = ICM;
        end
    end

    if (IEV2 - IEV1) == 0
        for JJ = 1:N
            X(IEV2,JJ) = (X(IEV2,JJ) + XC(JJ))/2.0;
        end
        I = IEV2;
        X(I,:) = CHECK(N,M,K,X,G,H,I,KODE,XC,DELTA,K1);
        fprintf(mainoutput,'%f  ',X(I,:));
        fprintf(mainoutput,'%n');
        parameter_set = X(I,:);
        CHANGEPARAMETERS(parameter_set,line_no);
        F(I) = FUNC(N,M,K,X,F,I,line_no,no_days)
        fprintf(mainoutput,'Function value: %d\n',F(I));

        % Find Point with Lowest Function Value (if point of highest value
repeates)
        IEV1 = 1;
        for ICM = 2:K
            if (F(IEV1) - F(ICM)) > 0
                IEV1 = ICM;
            end;
        end;

        % Find Point with Highest Function Value (if point of highest value
repeates)
        IEV2 = 1;
        for ICM = 2:K
            if (F(IEV2) - F(ICM)) <= 0
                IEV2 = ICM;
            end;
        end;
        IEV2

        Flag_KOUNT = 1;
        if (F(IEV2) - (F(IEV1) * (1+BETA))) >= 0
            KOUNT = 1;
        else
            KOUNT = KOUNT + 1
```

```
          if (KOUNT - GAMMA) < 0
             Flag_KOUNT = 1;
          else
             Flag_KOUNT = 0;
          end
       end

       IT = IT+1; % increasing number of iteration (to avoid infinite loop)
       flag_IEV2 = 1;
    else
       flag_IEV2 = 0;
    end
  end

  tt = IEV1
 end
end

% Find Point with Lowest Function Value
IEV1 = 1;
for ICM = 2:K
   if (F(IEV1) - F(ICM)) > 0
      IEV1 = ICM;
   end;
end;
fprintf(mainoutput,'Final value of parameters %f',X(IEV1,:));
fprintf(mainoutput,'\n');

% Running MITSIMLab with the final value
parameter_set = X(IEV1,:)
CHANGEPARAMETERS(parameter_set,line_no);
unix(COMMAND);
```

CHANGEPARAMETERS.m

```
function CHANGEPARAMETER(value,line_no)
% THIS FUNCTION WRITES THE NEW SET OF VALUES TO PARAMETER
FILE

fid = fopen('OD.dat','r');
```

```
fid1 = fopen('OD3.dat','w');

 for i=1:((line_no(1))-1)
    tline = fgetl(fid);
    a = sscanf(tline,'%c');
    fprintf(fid1,'%s',a);
    fprintf(fid1,'\n');
 end;
 tline = fgetl(fid);
 fprintf(fid1,'%s%f\n',tline(1:11),value(1));

  for i=((line_no(1))+1):((line_no(2))-1)
    tline = fgetl(fid);
    a = sscanf(tline,'%c');
    fprintf(fid1,'%s',a);
    fprintf(fid1,'\n');
 end;
 tline = fgetl(fid);
 fprintf(fid1,'%s%f\n',tline(1:11),value(2));

while ~(feof(fid))
    tline = fgetl(fid);
    a = sscanf(tline,'%c');
    fprintf(fid1,'%s',a);
    fprintf(fid1,'\n');
 end;

fclose(fid);
fclose(fid1);
copyfile('OD3.dat','OD.dat','f')
```

CHECK.m

```
function gg = CHECK(N,M,K,X,G,H,I,KODE,XC,DELTA,K1)
% global X N M K G H R F XC ITMAX IC IPRINT ALPHA BETA GAMMA
DELTA;
 KT = 0;
% CONST(N,M,K,X,G,H,I);
gg = X(I,:);
```

```
gg
 for J = 1:N
      if (gg(J) -G(J)) <= 0
       gg(J) = G(J) + DELTA;
      else
       if (H(J) - gg(J)) <= 0
        gg(J) = H(J) - DELTA;
       end
      end
end;
```

CENTR.m

```
function XC = CENTR(N,M,K,IEV2,I,XC,X,K1)
% global X N M K ITMAX IC IPRINT ALPHA BETA GAMA;
'in center function value of X(IEV2)'
X(IEV2,:)
for  J = 1:N,
               XC(J) = 0.0;
           for IL = 1:K1,
                              XC(J) = XC(J) + X(IL,J);
           end
               RK = K1;
   XC(J) = (XC(J) - X(IEV2,J))/(RK-1.0);
end
XC;
```

countspeed.m

```
% This file takes the "sensor.out" file and splits it into "counts.out",
% "speeds.out", "occupancy.out" and "headyway.out" file. It requires a
% file "SensorID.dat", which contains the sensor ids' for which data is
% available
fid = fopen('sensor.out','r');
fid1=fopen('counts.out','w');
fid2=fopen('speeds.out','w');
fid3=fopen('occupancy.out','w');
fid4=fopen('headway.out','w');
[sensor_ids] = textread('SensorId.dat','%d');

for i=1:4,
    tline=fgetl(fid);
```

```
end

Num_sensor_intervals = 0;
while ~(feof(fid))
    tline=fgetl(fid);
    if tline(length(tline))=='{',
        temp = sscanf(tline,'%d');
        tline=fgetl(fid);
        while sscanf(tline,'%c')~='}',
            a=sscanf(tline,'%f%hx%f%f%f%f');
            for i = 1:length(sensor_ids),
                if a(1) == sensor_ids(i),
                    Flag_Counts = 0;
                    Flag_Speed = 0;
                    Flag_Occupancy = 0;
                    Flag_Headway = 0;
                    j = 3;
                    if mod(a(2),2) == 1
                        Flag_Counts = 1;
                        a(2) = a(2) - 1;
                    end
                    if a(2) >= 8
                        Flag_Headway = 1;
                        a(2) = a(2) - 8;
                    end
                    if a(2) >= 4
                        Flag_Occupancy = 1;
                        a(2) = a(2) - 4;
                    end
                    if a(2) == 2
                        Flag_Speed = 1;
                    end
                    if Flag_Counts == 1
                        fprintf(fid1,'%0.2f  %d  %d\n',temp, a(1), round(a(j)));
                        j = j + 1;
                    else
                        tt = 0;
                        fprintf(fid1,'%0.2f  %d  %d\n',temp, a(1),tt);
                    end
                    if Flag_Speed == 1
                        fprintf(fid2,'%0.2f  %d  %d\n',temp, a(1), round(a(j)));
                        j = j +1;
```

```
            end
            if Flag_Occupancy == 1
                fprintf(fid3,'%0.2f  %d  %d\n',temp, a(1), a(j));
                j = j +1;
            end
            if Flag_Headway == 1
                %code for printing
            end


            %           Num_sensor(Num_sensor_intervals) =
Num_sensor(Num_sensor_intervals) +1;
        end
      end
      clear a;
      tline=fgetl(fid);
    end
    tline=fgetl(fid);
  end

end

fclose(fid);
fclose(fid1);
fclose(fid2);
fclose(fid3);
fclose(fid4);

%load counts1.out
%B = sortrows(counts1,[1 2]);
% save tempsensor.out B /ascii;
%dlmwrite('counts.out',B,'\t')
%clear B;
%!rm -f counts1.out
```

FUNC.m

```
function y = FUNC(N,M,K,X,F,I,line_no,no_days)
% This function return the function value. the objective function
% is sum of square of differences between the observed and simulated counts
global COMMAND
```

```
n_iterations = 1; % no. of runs over which averging is done

bigsum = 0;
  unix(COMMAND);
  countspeed;
  [A]=textread('counts.dat','%*d%*d%d');
  [b]=textread('counts.out','%*f%*d%d');
  a=A(1:12);
  bigsum = bigsum + (a-b)'*(a-b);
y = (bigsum/n_iterations); % Av. value of obejctive function
```