ROUTE

A Technical Overview

14TH NOVEMBER 2019

Ipsos

By: Andrew Green, Neil Farrer and David Strnad

Introduction

Simple…but Complex

The measurement of Out of Home audiences is simple in theory yet can be fiendishly complex in practice to carry out.

In theory, to create an accurate estimate of the number and types of people who have an opportunity to see advertising appearing on billboards, bus shelters, station carriages, shopping mall signs, taxi sides, underground station walls, car park panels and all the myriad environments in which Out of Home advertising appears, we need **five** things:

1. A **frame database** containing information on each frame or panel that needs to be measured – including its exact location (latitude and longitude) on a digital map and details of any information about the site that will influence peoples' likelihood to be able to see it;

2. A **count** of the people travelling (either walking, driving, cycling or otherwise moving) along each street, road or route and through every location where OOH frames are sited;

3. An indication of the typical **travel behaviour** of individuals over time;

4. A way to adjust raw counts of people passing frames to a more **realistic** estimate of how many of them will have had a chance to look at the advertising as they pass;

5. A method and system to **process** all these data into usable information.

There are multiple possible ways to approach each of these challenges. But for an accurate and credible estimate of audiences, the range of choice is narrower. When Route was launched in Great Britain in 2013, a number of methodological choices were made. The methods have not changed fundamentally over the years; but they have been adapted and refined as the medium itself has changed and new techniques have been tried and tested.

Route – an Overview

Route was a pioneer in OOH audience measurement, with many 'firsts' to its credit. So was its predecessor, POSTAR, the body managing OOH audience measurement prior to Route. POSTAR focused largely on measuring roadside billboards and bus shelters; Route was designed to capture all the different environments in which OOH advertising appeared, including on trains, taxis and buses, inside underground stations and shopping malls as well as alongside roads.

One task was to map all the poster locations, ultimately capturing the information necessary to estimate how many people would be able to see particular frames. This included the precise geographical

position of each site and physical information affecting its visibility to passers by – such as whether any obstacles obscure the panel from any points, the direction it is facing and its size etc.

These data were collected for around 240,000 roadside frames across the country (the number in September 2019 stood at 391,156 frames covering multiple environments, including 96,000 roadside billboards).

The next challenge was to know how many people were driving, cycling, walking or otherwise passing along every public pathway in the country. 'Public pathways' include not just major roads and city streets, but the corridors inside underground and overground stations, the escalators inside shopping centres, the inside and outside of airports and the interiors of trains. In short anywhere people went and could potentially be exposed to OOH advertising.

The solution to this was to create a **Traffic Intensity Model** (TIM). The TIM used a combination of limited traffic counts, sets of attributes for each kind of road and a modelling process to create these estimates for outdoor environments along routes where people pass in view of a poster. Different approaches were later added to cover indoor environments.

In this document, we review the Traffic Intensity Model methodology in detail and look at how it has adapted to changes in the medium and developments in measurement techniques. In the Appendix, we include a formal mathematical overview of the model.

Other components of the system include a **Travel Habits Survey** which we use to get a picture of the types of people represented in these numbers (e.g. gender, age, education levels etc.) and to understand how individuals travel between different points.

The Traffic Intensity Model

The Traffic Intensity Model (TIM) enables the following:

1. Traffic flow estimates for each link in the network;
2. An assessment of the reliability of each estimate – used in the modelling process;
3. The ability to update the estimate when new or updated data is available;

Data Layers

The model is organised into data 'layers' which are then subject to 'flow models' that transform the less than perfect data into traffic estimates for each road link.

The first layer in the system is a digital map of Great Britain. This is licensed from HERE Technologies, a company which provides mapping and location data. All roads and pedestrian areas of the country are represented in the HERE database. It has also been expanded into indoor locations. Every road on this map consists of what are called *links, sub-links* and *nodes*, where a link is a section of road, a sub-link is a smaller section of road and a node is the start or end of a link.

The second layer of data in the TIM consists of a series of characteristics associated with each of these links. These include the road type, the presence of buses, parking places, traffic rules (e.g. is it a one-way or two-way street? Are there speed limits or other restrictions?) and so on. This is sometimes referred to as the *network topology*.

Links are classified into five types, each of which have different levels of traffic intensity:

Functional Class	Description	Road Types
1	Roads with few, if any, speed changes, controlled access, high volume, maximum speed travel between major metropolitan areas	Motorways
2	Roads with few, if any, speed changes, high volume, high speed travel to and from Functional Class 1 roads	Major Highways
3	Roads that connect Functional Class 2 roads at a high volume with lower mobility	Non-primary A roads
4	High volume of traffic movement at moderate speeds between neighbourhoods	B roads
5	All other roads	Local roads

In the 2019 version of TIM (TIM5) Great Britain had a total of 4,064,181 road links, were further divided into 26,512,186 sub-links. This additional sub-classification enables us to apply traffic flows at a more detailed and granular level, reflecting the ways cars and people move along them.

The table below shows TIM5 links split by functional class and their count, length and average daily vehicular traffic (avg. number of cars per link). FC5 roads, which make up more than three-quarters of the roads by length – only account for 14% of traffic volumes.

Functional Class	No of links	Length of links (in miles)	Average traffic intensity (number of cars per day per link)	Share of Daily Traffic Volumes
1	51,109	7,506	30,751	14.3%
2	172,851	13,532	12,649	19.9%
3	295,639	18,126	10,618	28.5%
4	481,624	34,079	5,363	23.5%
5	3,062,958	229,241	497	13.8%
Total	**4,064,181**	**302,484**	**2,707**	**100.0%**

This information enables what is known as a *directed-graph* to be built so we can understand the way traffic flows. A one-way street obviously allows traffic to flow in only one direction; a two-way road consists of two opposite links in the directed graph.

Links are also classified according to whether most of the traffic will flow *through* them or will *originate* or *terminate* at the link. A car park or residential *cul-de-sac*, for example, will be an origin or destination; an A road running through a city suburb is more likely to be a flow link.

A third layer added to this map is traffic count data from the Department of Transport in the form of Annual Average Daily Flow (AADF) information. Around 22,758 of the 4.1 million road links across Great Britain offer traffic counts.

Just a handful of these offer continuous traffic counts (where vehicular volumes are measured and reported at all times of day and for every day in the year). This, in theory, would represent the ideal dataset for building an OOH measurement system.

In practice, such a dataset would be computationally challenging, demanding the ingestion of massive data volumes from every road in the country every day of the year. The data processing challenge is even greater considering the need to cover more than just roads and more than just cars.

As well as traveling by vehicle, people walk and cycle. They travel along a whole range of routes other than roads where poster advertising features, including on pavements, in corridors inside underground

stations, in airport terminals, on escalators, in and around shopping malls and on train station platforms. They also sit inside train and underground carriages where posters can be seen.

Vehicle Traffic Counts

The Department for Transport compiles traffic statistics using data from around 8,000 road side 12-hour manual counts and continuous data from a few hundred automatic traffic counters.

The street-level road traffic estimates provide the number of vehicles that pass each 'Count Point' location. They are compiled for each junction-to-junction link along Great Britain's major road network, and for a sample of locations on the minor road network. The following statistics are generated:

- *Average annual daily flow*: the number of vehicles that travel past (in both directions) the count point on an average day of the year
- *Average annual daily flow by direction*: the number of vehicles that travel past the location on an average day of the year, by direction of travel
- *Raw counts*: Where a raw count has been conducted at a given location, this provides the number of vehicles that travel past the location on the given day of the count, by direction of travel, for each hour between 7am to 7pm.

Manual traffic counts

The manual counts are conducted on a weekday by a trained enumerator, for a twelve-hour period (7am to 7pm). They are carried out between March and October, excluding all public holidays and school holidays, due to weather and light considerations at the count locations.

Manual counts on major roads: It is not possible to count every single location every year; therefore, the sections of road are surveyed on either an annual basis or on a cycle of every 2 years, every 4 years or every 8 years. The frequency is based on the traffic level. This means not every link in the major road network has a 12 hour count every year.

Manual counts on minor roads: Due to the vast number of minor roads in Great Britain it is not possible to count the traffic on all of them; instead a representative sample of minor road sites are counted each year. The difference in traffic between the two years is then applied to overall minor road estimates to calculate estimates for the latest year.

Automatic traffic counts

The Department for Transport's road traffic statistics team have approximately 300 automatic traffic counters at locations on Great Britain's road network. The locations represent a **stratified panel sample**, providing observations that can be used to estimate in-year traffic variations by road and vehicle type.

Automatic traffic counters are permanent installations embedded in the road, which combine Inductive Loops with Piezoelectric Sensors in a 'Loop – Piezo Sensors – Loop' array, and record information about vehicles passing over them, including vehicle length and wheelbase, to classify vehicles.

The Department for Transport's road traffic statistics also make use of automatic traffic counter data collected and maintained by other organisations. These include:

- Highways England: operates over 10,000 automatic traffic counters on some of the motorways and 'A' roads in England.
- Transport Scotland: operates more than 900 automatic traffic counters on some of the motorways and 'A' roads in Scotland.
- Transport for London: operates over 300 automatic traffic counters on roads in London.

Travel speed data is culled from the Travel Habits Survey (calculated from the time it takes for people to travel from one destination to another captured by the MobiTest meter carried by our panel).

The Modelling Process

All these layers of data are used to estimate traffic volumes for links with no count data. To do this we need to model how traffic *flows* between and amongst all the known count points. Complex regression models are used, with the computation split into two independent phases.

In the first phase, a simulation model is applied to construct an initial flow and a reliability estimate based on the traffic counts, network topology and link classifications just outlined. In the second phase, this initial estimate is balanced to satisfy flow conservation (i.e. the amount of traffic entering an intersection needs to equal the amount of traffic exiting the intersection).

For example, if a count station is located at a T-junction leading from one A road onto another, we will have a count of the number of cars entering the junction. There are two choices available: some vehicles will turn to the left and others to the right. Those turning right will eventually pass another intersection, with further choices available to them – perhaps to carry straight on or to turn.

At some point after one or more intersections, another count station will be passed, enabling an accurate count of vehicles entering and leaving the links either side of it. The job of the flow model is to quantify what happens between these count stations using everything known about the characteristics of the road links in between and to ensure the number entering and leaving each link is identical.

Vehicle Occupancy As we are counting people rather than the vehicles they travel in, we need to estimate the average number of people travelling in each car. For this, Route uses the National Travel Survey (NTS), which yields occupancy levels of between 1.4 and 1.5 depending on the size of the area being measured.

Counting Pedestrians - Overview

Pedestrian flows are handled differently to vehicle traffic flows. The first step is to list what are called 'Points of Interest' (e.g. schools, bus and train stations, shopping malls etc.). These are the places pedestrians tend to gravitate towards – as cars do to road links. For each Point of Interest, we look at census data on the size of the resident and working population in the immediate area. For some POIs, detailed pedestrian count data is also available.

For Route, the country is divided into a little under 230,000 'output' areas, detailing the size and composition of the population and the number of cars owned by each household. We access count data from some 988,000 locations (most of them bus stops). The table below has the details.

Count station reference	Count description	Number of Count points
C001	Output area counts - including daytime counts from Census data	(227,759) *
C002	RODS (Rolling Origin & Destination Survey) data for London Underground (entries and exit) plus Patronage data for DLR	314
C003	LENNON (Latest Earnings Network Nationally Over Night) data for train stations plus Glasgow subway and Light Rail	2,792
C004	CAA (Civil Aviation Authority) data for Airport	31
C005	Volume on HERE links - Buses - bus stops	875,923
C006	Experian data for Supermarket	17,995
C007	Experian data for Shopping centres	1,774
C008	PMRS (Pedestrian Market Research Services) survey data for footfall	14,729
C009	Internal client - Schools	48,118
C010	Leisure facilities	5,599
C011	Point-X data for car-parks	14,790
C012	HERE data for Service stations	138
C013	HERE data for Hospitals	1,380
C014	HERE data for Bus Stations	160
C015	HERE data for Tourist attractions	992
C016	HERE data for Post offices	2,456
C017	Sourced and approved extra counts (bespoke PMRS, MO & LDC counts) - Shopping locations other	753

* *We have adult population data for 227,759 output areas, but the counts are spread over all residential postcodes in the area and are not single POI counts*

The next step is to build a **gravity model**, a technique for estimating the flow of people between any two places based on the size of the group and the distance between the points. This is designed to expand the count data we have to the approximately 4 million total number of pedestrian links.

The ROUTE Approach in Detail

We measure the journeys of both the urban and suburban population. The challenge is to distinguish between pedestrian and vehicular movement, especially in built-up areas, where – whether you are traveling by foot or in a vehicle - speeds tend to be low and the GPS signal can sometimes be lost.

The second-by-second GPS data captured by the MobiTest meter is used in two different ways:

- To determine the sequence of street links along which people travel

- To map their intersections with Points of Interest

The sequence of street links represents each respondent journey. The GPS data are matched with the roadside network to determine which streets/roads were travelled on by the respondent. For each of the street links, we have information about the time they spend on this link and on the average speed of travel. This, together with overall respondent travel patterns and knowledge of any access restrictions related to the different street links, helps us to determine their mode of transport.

Intersections with Points of Interest provide information about likely visits to these locations.

Step 1 - POI visits

We evaluate visits to Points of Interest based on second-by-second GPS traces at and around each location ('buffers' are drawn around Points of Interest to indicate the likely area within which people will be heading to or away from each place). The size of each buffer will vary according to the location type. For example, a major train station or shopping centre will have a much larger buffer than most bus stops. All visits longer than three minutes are initially assumed to represent a stop. Each stop is matched with a location and a Point of Interest. Data for any measured POI are cleaned further by looking at travel behaviour before and after the stop to be as certain as possible that people are, in fact, walking rather than travelling in a vehicle.

Step 2 – Travel survey population assignment

The Travel Survey sample can never be large enough to accurately represent visits to every single Point of Interest in the country (which is clear when we compare Travel Survey data with measured external counts – e.g. of shopping mall footfall). So, a certain amount of data integration is carried out to enable projection of our sample to the population (for example, we segment people according to their typical travel behaviour on weekdays vs. weekends). From this procedure, we generate average daily pedestrian flow estimates for each Point of Interest.

The importance of the travel survey population assignment lies in the opportunity it affords us to evaluate the pedestrian flow ('gravity') to and from Points of Interest and to merge the gravities between multiple POIs close to one another. This latter situation arises when, for example, a bus station

is next to a train station and we need to determine how many people are visiting either one of these in isolation or both.

Step 3 – Footfall counts

We derive pedestrian counts from a range of sources including direct footfall counts (mall footfalls, station usage, etc.) and indirect estimates (ticket sales etc.). For a list of count sources, see the table on P8. The counts are grouped into three categories:

C001 residential pedestrian counts

Residential counts are computed for links connected to houses/address points. The census population of each output area (OA) is assigned to address points on links inside an OA 'polygon' (boundary). Each link and sub-link then receive part of the total population weighted by the number of address points assigned to it.

Links with a higher density of address points generate higher residential population numbers. The gravity is limited to the residential link and links directly connected to it by end nodes. An experimental exercise based on travel survey respondents assigned to links confirmed that a longer gravity extrapolation creates duplication with bus stop gravity or POI gravity.

The residential component represents pedestrian flows close to places of residence. A general factor of 2.5 visits per day is applied to each residential link value as an average. It results from an initial travel survey analysis of average trips made during the day (in and out)

C00x Points of interest

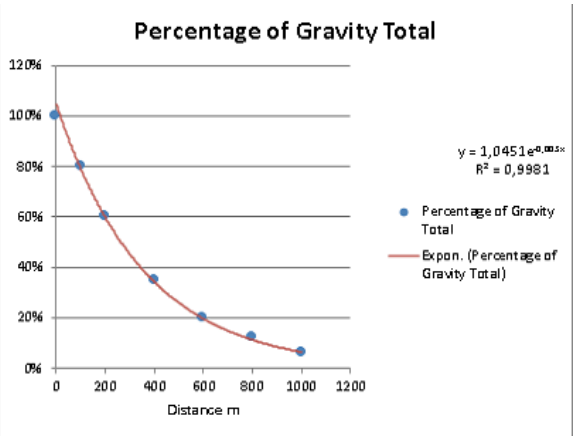
This input is a collection of different data sources representing Points of Interest. Gravity distribution usually means an overlap of different attractors in proximity. An empiric distribution distance and merger of overlapping volumes results from multiple tests and iterations. The final settings deliver results with explainable volumes close to individual POI attractors as well as volumes between attractors.

C005 Bus stops/links turnover

The first version of the model was based on distribution around bus stops. This generated volumes on links with bus stops. In some cases, these links could measure greater distances. We have since started to evaluate all links covered by bus lines to make the bus gravity connected and gap-free. We also enhanced the model by making the finer division of links into sub-links; we have succeeded in achieving a better representation of buses within the local traffic volumes. Short sub-links have enabled us to run more reliable gravity distributions.

Examples - Curves applied in TIM PED at bus stops

Gravity Distance	Percentage of Gravity Total	Probability of sharing attractors	Part contributed to the max intensity on a link
m	%	%	%
0	100%	94%	6%
100	80%	88%	12%
200	60%	80%	20%
400	35%	65%	35%
600	20%	40%	60%
800	12%	20%	80%
1000	6%	0%	100%

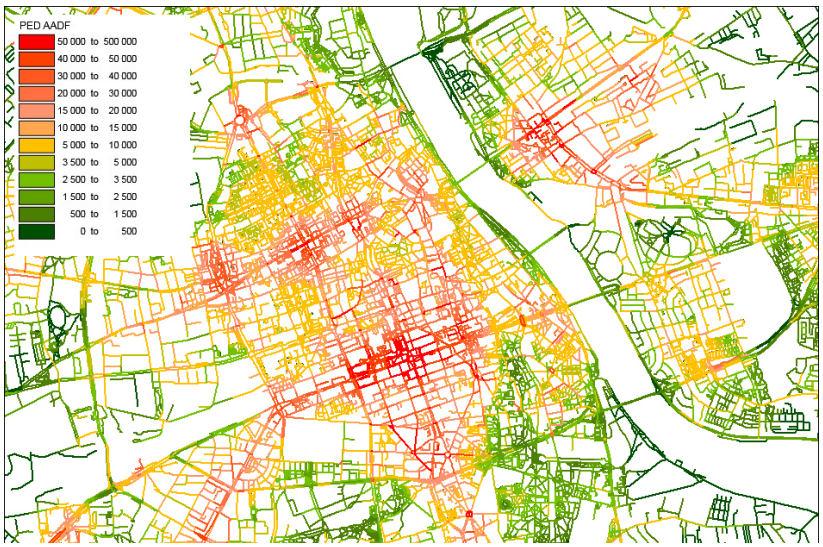


Step 4 - Gravity function and Gravity Processing

The next step concerns links (i.e. routes) connected to Points of Interest. We assume that visitors have been using connected links i.e. travelling along adjacent streets and pavements to access individual POIs. To simulate this, we apply a technique known as gravity distribution, whereby we assign pedestrian volumes to connected links. The intensity (number of pedestrians we count as travelling towards or away from a POI) falls as we get further away from the location. Where two POIs are close together, we take the data we have on the combined area around the two points and make certain assumptions to avoid duplication of pedestrians and overestimation of the final traffic volumes. This is based on learnings from Step 2.

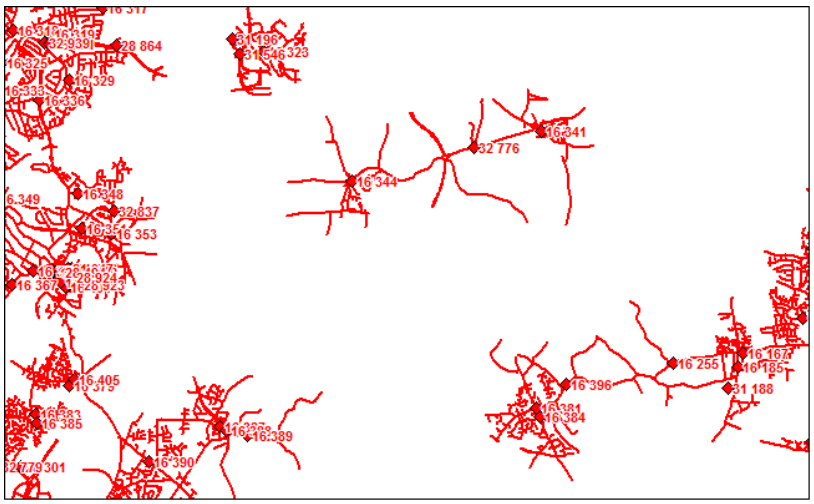
Example – PED model

This image shows a gravity model distribution and how the counts spread in relation to other counts.



In all the TIM PED versions, HERE links comprise the main geography. From TIM 5 onwards, we have divided links into shorter sub-links. TIM PED is based on HERE links which allow pedestrian traffic (AR Pedestrian = "Y"). Other HERE links restrictions are ignored (one-way, tunnels, ramps, bridges) - pedestrians are allowed anywhere.

Example – POI and spider web of gravity links connected to each point



Gravity Processing

Gravity processing of TIM 1 to TIM 6 was based on scripts running on three separate count threads (described step 3):

- Bus stops and links
- Residential
- POIs

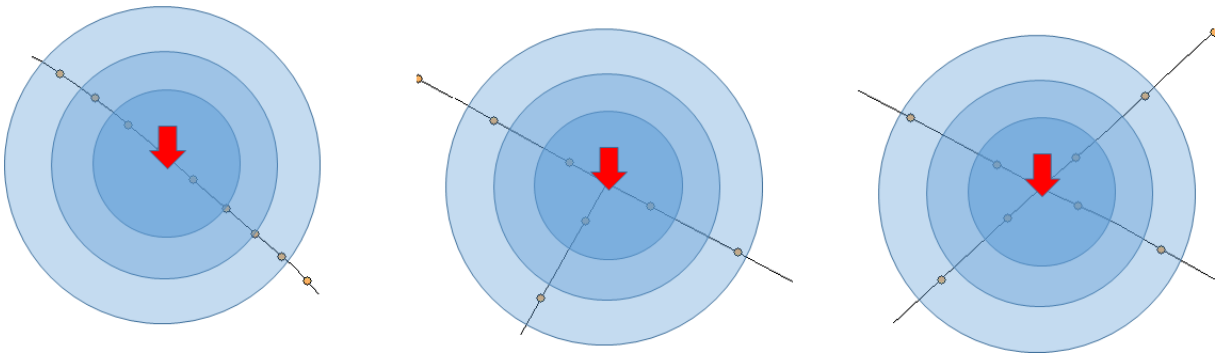
The final gravity was created as a combination of all three components.

Scripts are generated around each POI attractor gravity polygon (boundary) and these are intersected by sub-link centroids. Using centroids (points) instead of sub-links (lines) improves performance. The difference in gravity distribution is very low; the processing speed is approximately 10x faster than by using points.

Curves were converted in a series of steps where each step represents a specific gravity distance. Each distance has a defined gravity factor. This factor considers the number of connected sub-links dividing the volume after gravity application by the number of connected sub links.

This means that locations connected to 4 different sub-links have 2 times faster gravity decrease than a location connected to 2 different sub links.

Example – Gravity in 2 directions, 3 directions and 4 directions



Examples - gravity by distance from home, as measured on the Travel Survey

	Cumulative percent	
Distance travelled from location (m)	Away from home	From home
0	100,00%	100,00%
200	76,08%	77,18%
400	52,54%	56,48%
600	35,85%	42,20%
800	25,36%	32,47%
1000	18,49%	25,33%
2000	4,87%	8,20%
3000	1,68%	2,93%
4000	0,71%	1,15%
5000	0,33%	0,51%
6000	0,17%	0,24%
8000	0,00%	0,00%

TIM Updates

The Traffic Intensity Model (TIM) was first designed by MGE Data in 2008 and implemented in Great Britain the following year and in Croatia three years later. The approach was also implemented in The Czech Republic and Poland in 2015. But the TIM has not stood still.

The various steps in the process of creating and managing the Traffic Intensity Model are complex and challenging. During the process, we have learned a great deal about all the ways in which traffic estimates can be improved – whether it be by ensuring the accuracy of the source data, adjusting some of the model procedures or adding new environments.

As a result, the TIM implemented back in 2009 has been updated roughly every year since – as at the end of 2019, we are using TIM V.5.0, but will shortly launch V.6.0. The main updates to the model in each version have been as follows:

TIM VEH Versions

TIM Version	Main Developments	Date
1	Original version	2009
2	Update to counts, improvements to flow distribution	2012
3	Update to counts, geometry (previously missing road segments filled), extra granularity of links, bespoke counts added as well as local authority counts, update to VA method, speeds	2013
4	Updates to counts	2015
4.2	update to geometry and counts	2016
4.3	update to peds only to include traffic restriction	2017
5	Update to counts, geometry, move to sub-links, update to speeds	2018

TIM PED Versions

TIM Version	HERE MAP Ver.	Geometry	Inputs and updates	Methods
1	2008q2	Links	RESID BUS POI	Initial Gravity v1
2	2008q2	Links	RESID BUS POI	Gravity v1
3	2012q4	Links	RESID BUS POI	Gravity v2
4	2012q4	Links	RESID BUS POI	Gravity v2
4.2	2015q3	Links	POI	Gravity v2
4.3	2015q3	Links	POI	Gravity v2
5	2017q1	Sub links	RESID BUS POI	Gravity v3
6	2018q1	Sub links	RESID BUS POI	Gravity v3

TIM PED updates were historically based on following changes:

- HERE Version
- Input data updates
- All three inputs RESID, BUS POI or selected POI inputs
- Geometry set-up – links and sub-links
- Gravity distribution
    - Gravity V1- initial version, distribution low volumes to very long distances
    - Gravity V2- optimised to shorter distances with minimum impact on final volumes on links (faster processing speed)
    - Gravity V3- optimised to generate distribution results on sub links instead of original links

APPENDIX

TIM

A formal description of modelling and creating the individual vehicle traffic intensity map

Modelling the individual vehicle traffic intensity map

An Informal problem description

The starting point is an accurate and detailed topological map of a road network. At some road links periodic measurements of traffic density are available. The objective of the model is to **estimate the traffic density (1)** at road links where measurements are not available (due to, e.g., cost or other factors). This estimate should also be provided with a **measure of its reliability (2)**. A necessary feature of the estimates is, that they should **match up at the intersections (3)**. That is, the sum of the incoming traffic densities at an intersection should equal the sum of the outgoing traffic densities.

The problem can be divided into two parts: **Estimate calculation** (1), subject to the condition (3) and **Reliability assessment** (2).

Notes to the informal description

(1): Traffic density is measured as *number of vehicles / units in time*. It is important that this is either a pure empirical measurement or a simple arithmetic mean of several empirical measurements. In fact, in the following it will be assumed that given a measured density we can calculate the total number of vehicles travelling on the given road link in a specified amount of time.

(2): The measure of reliability is a real number on a scale ranging e.g. from 0 to 1 with 0 being the lowest and 1 the highest. It is not required that this is the probability of the estimate being valid.

(3): This is a requirement which is valid for the total amount of traffic. It is less clear, that it should also hold for the average amounts. If the average is a simple arithmetic mean, this requirement is still valid. In other cases, it might fail. That is why we require in (1) that the total number of vehicles can be calculated from the given measurements.

Introductory remarks

Problems of this kind are studied in the branch of applied mathematics called Operational Research. This is an established field of mathematics that goes as far back as the 18th century and the work of Ch. Babbage (Sodhi 07). Modern Operational Research arose during World War II with the work of P. Blackett and others in UK and G. Dantzig in USA. Since then it has experienced enormous growth. With the introduction of computers applications of theoretical results have become readily available to the public at large and Operational Research methods have seen wide adoption in business, industrial engineering and other areas. See (Taha 06) or (Wikipedia 09) for more information about Operational Research.

As an established field Operational Research has numerous scientific journals where new results and case studies are published. The following is a brief selection:

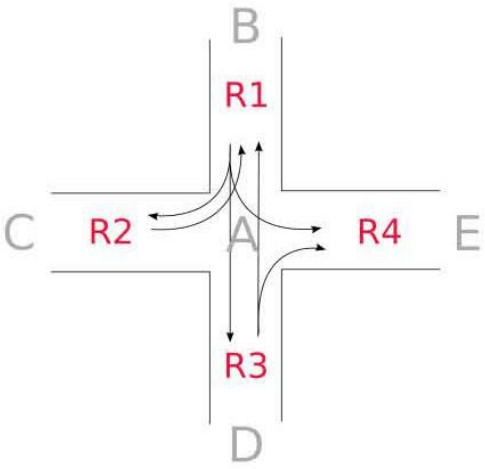
- Information Systems Research
- Mathematics of Operations Research
- Operations Research: Journal of the Institute for Operations Research and the Management Sciences
- Transportation Science

The formal model

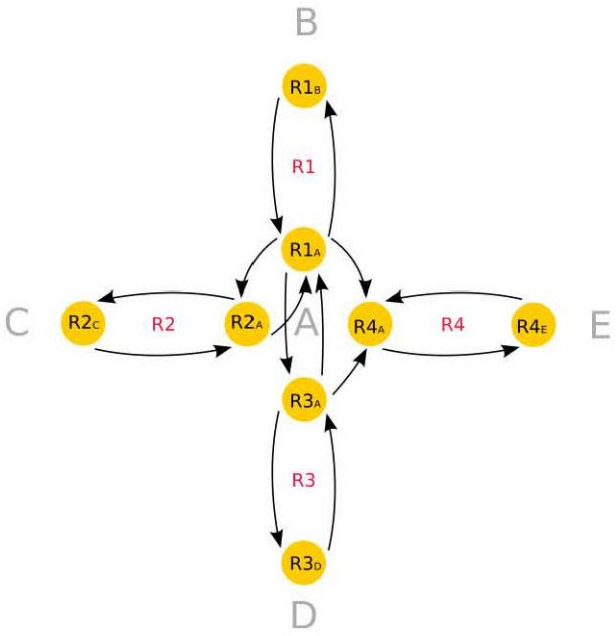
In designing a formal model for the real-world situation, we need to extract the essential features of the situation and make some simplifying assumptions to make the problem tractable.

Feature #1: Graph structure

The underlying configuration of the road network is easily described as a mathematical structure called a directed graph. The first approach where road links correspond to edges of the graph and intersection nodes to vertices of the graph will fail, however, in our specific case. The road network has associated with it certain traffic rules (e.g. no left turns, etc.). Consider the following schema of a road intersection together with arrows indicating allowed flow of traffic:



This schematic situation cannot be modelled by a graph where the edges would be in one-to-one correspondence with roads and vertices with intersections. We must therefore be more careful. Each connection (road link) R between two intersections A, B will be modelled by two graph vertices (RA, RB) corresponding to the two intersections and two edges (RA->RB, RB->RA) corresponding to the connection leaving A and heading to B (RA->RB) and the connection leaving B and heading to A (RB->RA). A vertex RA will moreover be connected by an edge with all vertices TA where T is a road that a vehicle can take after arriving at intersection A via the road R. The previous schematic situation will now look as follows:



This representation is completely general. It should be noted, that the above graph represents a situation where no U-turns are allowed. If a U-turn were allowed when arriving to the intersection A via the road R3, we would need to add the edge R3A->R3A to our graph.

Feature #2: Density Measurements

We assume that the traffic density is measured at single **road links** rather than intersections. Also, we assume that the density is measured in **both directions separately**. The density measurements will be represented by assigning a number to the respective edges in our graphs.

Simplification #1: A static model

Traffic is a naturally dynamic phenomenon where the densities change over time. A completely accurate model would need to take this into account. There are two main obstacles to a dynamic model, however. The first is its inherent **complexity**. Due to the size of the underlying structures (in the order of millions of vertices and tens of millions of edges) it is doubtful, whether constructing and evaluating such a model would be computationally feasible.

A second obstacle is the **static nature of available data**: the density measurements are averages over time intervals considerably larger than would be required to provide data to a dynamic model. It is assumed that, at this granularity, a **static model** will not introduce more error then already present in the available data. For a dynamic modelling approach, see e.g. (Birge & Ho 93), (Davis & Nihan 93) and (Boyce et al. 93).

Estimate calculation

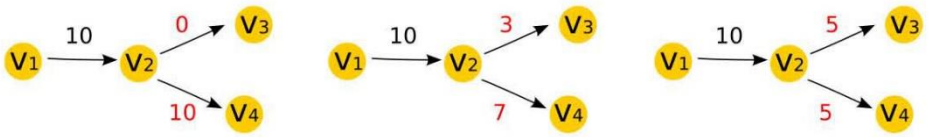
Based on the above model a first approximation to the solution of the is the following formulation:

$$f(e) = m_e, \quad e \in M$$

$$\sum_{e \in in(v)} f(e) = \sum_{e \in out(v)} f(e), \quad v \in V$$

Given a directed graph G = (V,E) find a function f: E → R assigning to each edge a traffic density subject to the following restrictions: where M is the set of edges corresponding to roads, where traffic density data is available. The second equation formalizes the requirement that the estimates **match up at intersections (3)**, that is the sum of inflowing traffic densities equals the sum of out flowing densities.

However, in this formulation the problem will typically have many different solutions. For example, the following picture shows three such solutions for a simple problem with four vertices and three edges (the measured values me are indicated in black, the computed values f(e) in red):



(A) (B) (C)

To choose between these solutions we need some other criteria. A natural criterion arises out of the underlying road network data: we know the size of the respective roads and we know their capacities. In the above example, we might know that the road from V2 to V4 is a highway, whereas the road from V2 to V3 is a small local road.

Thus, solution C is very unlikely: the amount of traffic going along a highway will probably be bigger than the amount of traffic going along a local road. Trying to build this intuition into the model, we may try to assign a cost c_e which is inversely proportional to the size of the corresponding road to each edge in the graph. Also, each road has a capacity – a limit on the amount of traffic that can flow along it. Formalising this, we want to find a solution to the above problem which would satisfy the following restrictions (capacity)

$$f(e) \leq c_e$$

and minimise the total cost:

$$\sum_{e \in E} f(e) \cdot c_e$$

The advantage of this approach is that this is the well-studied **minimum cost network flow** problem with a wealth of efficient algorithms available that find an optimal solution. (the simplex method, the ε-relaxation method, the primal-dual algorithm etc.). See e.g. (Matoušek 06), (Ahuja et al. 93) or (Chandrasekaran) for a list of other network flow books.

This approach has a drawback, which can also be illustrated in the above example. The optimal solution to the minimum cost network flow problem in our case would be (A). However, this is somewhat contrary to our expectations in real life: although cars are more likely to travel along the highway, at least *some* cars are very likely to travel along the local road. Thus, the solution (B) would be a more *realistic* solution. Rather than assigning a cost to a road, we would like to assign a fraction of the traffic to an edge. The solution to the problem would then minimise the following function:

$$\sum_{v \in V} \sum_{e \in out(v)} |f(e) - ifl(v) \cdot p_e|, \quad ifl(v) = \sum_{e \in in(v)} f(e)$$

where p_e is the proportion of traffic travelling along edge e. Due to the presence of absolute values (a more realistic formula would take the square of the absolute value, which introduces another nonlinearity) this is no longer a linear equation and the problem becomes much harder. With a bit of work, we can however reduce this problem to a (still nonlinear) problem that is easier – the so called **convex cost minimum network flow** problem. This is a generalisation of the minimum cost network flow problem, where costs assigned to each edge are not constant, but depend on the amount of flow along the edge, i.e. c_e is a convex function of f(e). Minimising the above sum is equivalent to minimising the cost with the cost function given by the following formula:

$$c_e = |f(e) - ifl(v) \cdot p_e|$$

However, c_e in this formula depends not only on f(e) but also on ifl(v), i.e. it is **non separable**. It turns out that for a given vertices v, it is possible to construct **separable** (i.e. depending only on f(e)) convex functions c'e, such that the sum of the c'_e is very close to the sum of the non-separable c_e.

The convex cost minimum network flow problem can be solved efficiently and, in particular, can be parallelised, see e.g. (Beraldi et al. 01). Although exact solutions are not necessarily obtained, this does not matter since the cost function itself is, by the nature of the problem, not known exactly. Moreover, the flow constraints conditions (match-up at intersections) are always satisfied **exactly**.

If the above reduction turns out to be too simplifying, the more complex and computationally challenging algorithms described in (Nielsen & Zenios 93) or (Ourou et al. 00) will need to be used.

Estimate reliability

The proposed model for calculating the reliability of estimated traffic density is based on the assumption that the further we are from a point where a measurement was made, the less reliable the estimate is. The estimate calculation will give us traffic densities on single roads – edges of the graph – this is called *arc-flow* in the literature. This can be decomposed to what is called a **path-flow**. In other words, given a path through the road network, we know how many cars had actually travelled across this path.

Actually, we do not get traffic densities for *any given* paths, but the paths for which we do get the densities will cover the whole graph. This is all we need for our reliability estimate. We can now proceed as follows. Given an edge R in the graph we want to calculate the reliability of the estimated value f(R) of traffic density. We first look at all the paths which contain R and which have a density estimate. For each such path P, we determine the proportion z_P of traffic it contributes to the total traffic f(R) through R. We then find the distance d_P from R along the path P to a point, where a measurement was made. The reliability of f(R) will then be the weighted sum

$$\sum_{\{P:R\in P,P\in Path\}} z_P \cdot k^{(d_P)}$$

where k is some constant between 0 and 1 representing the drop-in reliability when traversing a single road.

The proposed model is fairly simplistic and could be elaborated upon. For example, the constant k could vary from road to road. Alternatively, we could assign to each measurement point a probability distribution (say Gaussian) and then inductively derive the probability distribution for the other edges. However, such a computation would still be rather arbitrary, while potentially giving the user a false impression of what it actually is.

Creating the Traffic Intensity Map

Available data

AADF Counts.

At a fraction of the roads periodic traffic counts are available. The coverage of the whole network is roughly 1% with a higher coverage of major roads. All the traffic counts are prepared using a well-defined methodology.

Network Topology

The road network is available as an undirected graph with each link corresponding to a road and

each node corresponding to an intersection[1]. Moreover, each link has several attributes connected to it. For the topology network, the attribute specifying whether it is a one-way or a both-way road is most important. Traffic restrictions (i.e. no U-turns, no right-turns, etc.) are also available as a separate table.

We use this information to build a *directed* graph incorporating all traffic restrictions into the topology. This is done in two steps. First, we build a directed graph (each both-way road corresponds to two opposite links in the directed graph). Next, we add an origin and destination node for each link. We connect the destination nodes of a link L1 with the origin nodes of a link L2 if they were originally connected to the same node and the traffic restrictions allow traffic to pass from L1 to L2.

This modification of the graph has the following important property: each node has either at most one incoming link or at most one outgoing link. This property makes it possible to *uniquely* reconstruct path flow from link flow.

Finally, we classify links according to whether the majority of flow along them is expected to be "through flow" (TF) or "originating/terminating flow" (OTF). In graph-theoretic terms, the OTF links correspond to (full) subnetworks without cycles (i.e. trees). The OTF links are grouped into areas (corresponding to a single tree subnetwork) having a unique connection to the rest of the network. The areas are then collapsed (or represented by) the single link connecting them to the "through flow" part of the network. These connecting links are initialized with a minimal traffic intensity based on residential survey data (cars/household, number of households, ...) and POI data.

Link Classification

Each link in the network is classified into five classes according to the importance/usage of the link. This classification is based on empirical observation, construction & traffic characteristics (number of lanes, maximum speed, capacity) and other inputs. The classification divides the network into several subnetworks. It is assumed that the flow on these subnetworks is largely independent in the sense that any inter-network flow is balanced – the amount of traffic leaving a subnetwork is equal to the amount of traffic entering the subnetwork.

Deliverables

- Traffic flow estimate for each link of the network (in the future, flow estimate for paths through the network will be available)
- Reliability assessment of each traffic flow estimate
- Possibility to update the estimate when new/updated data is available

Outline of possible approaches

Traffic is a naturally dynamic phenomenon where traffic flow changes throughout the day. A completely accurate model would need to take this into account. There are two main obstacles to a

[1]Actually, there are nodes which do not correspond to any intersection in the strict sense. Rather they correspond to a place where some important characteristic of the road changes.

dynamic model, however. The first is its inherent **complexity**. Due to the size of the underlying structures (in the order of millions of vertices and tens of millions of edges) it is doubtful, whether constructing and evaluating such a model would be computationally feasible. A second obstacle is the **static nature of available data**: the AADF counts are averages over time intervals considerably larger than would be required to provide data for a dynamic model. It is assumed that at this granularity a **static model** will not introduce more error then already present in the available data. We have investigated two possible approaches to modelling the traffic flow.

Origin-Destination model

This is the standard model for traffic networks. It is well covered in the literature. The input into this model is an OD matrix. Given an OD matrix, there are algorithms for computing a traffic assignment which gives rise to the input matrix. Typically, this assignment is required to satisfy an equilibrium condition. For an overview see e.g. (Boyce et al. '88).

The advantage of this approach is the available theory behind the model and the large amount of experience with these models.

There are disadvantages, however. A first disadvantage is the computational complexity of the model. It is not clear that it would be feasible to model the whole network (on the order of 10s of millions of edges and millions of nodes). This would have to be worked around by splitting the network into separate parts and then piece these parts back together. Another disadvantage is that the model does not easily account for the accuracy and coverage of data in the sense that it does not give a direct reliability estimate of the resulting flow on a single link. The main disadvantage, however, is the fact that the input for the model (i.e. the OD matrix) is not available. Rather, the data consists of discrete traffic counts, covering a fraction of the network. Even if the counts were available for each network link, finding an appropriate OD matrix is a strongly underspecified problem. A possible solution, see (Yang et al. '94) or (Abrahamsson '88), would be to start with an OD matrix estimate and choose amongst the matrices one that is close to the initial estimate. Given the coverage of data, it is likely that this would still be an underspecified problem.

Flow simulation model

The flow simulation model calculates network flow by recursively distributing flow derived from AADF counts throughout the network, respecting the network topology and the characteristics of the individual links.

The advantages of this approach are its computational and conceptual simplicity. Modifications to the algorithm to accommodate different sources/types of information and different requirements can be easily implemented since the logic behind the model is very simple. Moreover, it allows for reliability assessment, in fact it is even possible to calculate probability distributions for each link (however it is not clear that this is reasonable).

The disadvantage of this approach is that in most cases it likely does not achieve an equilibrium solution. Further the model is not guaranteed to give a balanced solution (e.g. one satisfying flow conservation) and the output must be postprocessed to achieve flow conservation. Another

problem is that areas of the network with a limited amount of AADF counts will not have any flow at all. This problem (which is present also in the OD approach) can be mitigated by specifying a lower bound for the flow along links based on their construction characteristics or residential survey data.

Conclusion

We have chosen the distribution model, mainly because of the unavailability of an OD matrix, the simplicity of the model and the availability of a reliability assessment. Moreover, the OD approach is methodologically somewhat inappropriate, since the output of the model will be used to build an OD matrix, while the OD approach expects a matrix as its input. Given the coverage and accuracy of available data it is expected that the OD approach would not be better than a pure guess. As a summary, the following are the reasons why we have chosen the distribution model.

- simplicity
- methodological soundness
- reliability assessment available

Algorithm Description

The computation is split into two independent phases. In the first phase the simulation model is applied to construct an initial flow and a reliability estimate based on the traffic counts, network topology and link classification. In the second phase this initial estimate is balanced to satisfy flow conservation. The balancing process is carried in such a way so that it minimizes the changes necessary to achieve balancing.

Simulation phase

During the simulation phase the main question is how to distribute traffic flow between the outgoing links of a node. A first possible approach is to split the traffic proportionately according to the link classification. We have opted for another approach. We assume that the classification divides the network into mostly flow-independent subnetworks, each of which separately satisfies flow conservation. We then distribute the traffic in each of the subnetworks separately and split the traffic evenly among the outgoing edges in each subnetwork. This means that a traffic count in a subnetwork will not directly contribute to the traffic in other subnetworks[2]. Given the coverage of data, this is not very economical. To mitigate this problem, we assume that at links connecting different subnetworks a flow exchange between the network happens and this flow exchange does not affect flow conservation (i.e. the flow which subnetwork A sends into the subnetwork B is equal to the flow subnetwork B sends into the network A). This flow exchange will give us a lower bound on the flow along the connecting links. The simulation phase is itself further split into four stages.

[2]The situation is a little more complicated in case where the link connecting the subnetworks is a one-way road. See implementation details.

1. In the first stage, we take each traffic count and distribute the respective flow along the subnetwork to which the traffic count belongs. The flow is distributed in the direction of traffic.
2. The second stage is a mirror counterpart of the first stage, where we distribute the flow in the reverse direction.
3. In the third stage we average the results of the first two stages.
4. In the final stage we iterate along nodes belonging to at least two different subnetworks and ensure minimal flow exchange between the subnetworks by recursively assigning appropriate lower bounds on links.

Flow balancing phase

The second phase balances the results of the first phase to achieve flow conservation while departing from the initial estimate as little as possible. This problem can be formally written as follows:

Find flow

$$f : E \to \mathbb{R}_0^+$$

such that for each node N, the inflow is equal to the outflow, i.e.

$$\sum_{e \in In(N)} f(e) = \sum_{e \in Out(N)} f(e)$$

the AADF counts are respected[3], i.e.

$$f(e) = count(e), \quad e \in \text{AADF}$$

and

$$\sum_{e \in E} c_e (i(e) - f(e))^2 + g(e)$$

is minimal, where *i(e)* is the initial estimate, c_e is a positive constant and *g(e)* is some positive, convex function.

This is equivalent to a *minimum convex cost network flow* problem. Varying the constants c_e, we can control the amount of balancing allowed on links based on the reliability of the initial estimate at the link. The *g(e)* part is used to "enforce" capacity constraints. Making *g(e)* grow significantly outside the capacity constraints we can effectively ensure that the constraints are satisfied if

[3]Actually, in some circumstances an AADF count *can* be changed in the simulation phase. In the balancing phase it does not change. See implementation notes for details.

consistently possible. The constraint *f(e) = count(e)* for links with AADF counts is guaranteed by introducing *source/sink* nodes splitting the edges where an AADF count is available.

The minimum cost network flow problem is well studied. We have used an ε-relaxation method from (Bertsekas et al. '97), see also (P. Beraldi et al. '01). To speed up the computation we have allowed both for up and down iterations[4] The main advantage of this method is its amenability to parallelization[5].

Reliability estimate

The proposed approach for calculating the reliability of the estimated flow is based on the intuition that the farther we are from a point where a measurement was made, the less reliable the estimate is. The final result of the computation will give us traffic flow on each link of the network a so-called *arc-flow.* However, in the first phase when we distribute the flows starting from AADF count sites, we actually know the *path flows*[6]. In other words, given a path through the road network, we know how many cars had actually travelled across this path. Actually, we do not know the traffic flow along *any given* paths, but the paths for which we do know the flow will cover the whole graph. This is all we need for our reliability estimate. We proceed as follows. Given a link *e* in the graph we calculate the reliability *r(e)* of the estimated traffic flow *f(e)*. We first look at all the paths which contain *e* and for which we know the flow. For each such path *p*, we determine the proportion z_p of traffic it contributes to the total traffic *f(e)* through *e*. We then calculated a *modified distance* d_P from *e* along the path *p* to a AADF count site:

$$d_p = \prod_{n \in p} c_n$$

where c_n is a positive constant < 1 which quantifies the reliability drop when passing through node *n* (along path *p*). The reliability *r(e)* of will then be the weighted sum:

$$\sum_{p \in P} z_p \cdot d_p$$

The proposed model is rather simplistic and could be elaborated upon. For example, the constant c could vary from road to road. Or we could assign to each measurement point a probability distribution (e.g. gaussian) and then inductively derive the probability distribution for the other edges. However, such a computation would still be rather arbitrary while potentially giving the user a false impression of what it actually is. Moreover, a precise computation of the distributions would

[4] See (De Leone et al. '95) where the authors show empirical evidence that the down iterations significantly decrease the time needed for calculation. This has been confirmed in our own tests.

[5] Parallelization is not implemented in the first version of the algorithm. Empirical experiments suggest that the algorithm runs reasonably fast.

[6] As mentioned in the Network topology section, we can actually always reconstruct the path flow even after the balancing phase. This would be more reasonable however due to time constraints it will not be done in the first version of the algorithm.

be challenging due to having to calculate maximums of several distributions[7].

Updating the estimate

When additional AADF counts are available (or updates to the original AADF counts are made) the process above will be modified to allow for faster computation. The simulation phase (which is relatively fast) will be run with the new data. On the links that the estimate does not (significantly) differ from the results of the simulation phase with the original data, the results of the balancing phase with the original data will be used instead of the new simulation estimate. This "merged" estimate will then be run through the balancing phase which, depending on the amount of change, will run much faster since most of the network will already be balanced.

Implementation notes

Input data consistency

The original design assumed no errors in the AADF counts as a simplification. However, this turned out to be too simplistic. There are two types of errors occurring in the AADF counts. One error comes from the inaccuracy of the counting process. Since the accuracy of the simulation process is much lower, this type of error is not important. A second source of errors is much more relevant however. The actual AADF counts are given in a table where each count is identified by geographic coordinates and the name of the road counted. In many situations this information does not identify the network link uniquely (or there are errors in this information). Since the process of assigning each count to a link must be automatic (due to the large volume of data) errors in the assignment are inevitable. A wrong assignment may, however, lead to blatantly inconsistent situations. Based on empirical tests we have estimated that this type of error affects around 1% of all the AADF measurements. Due to time constraints we have chosen to omit the problematic[8] measurements altogether. In a subsequent version we may try to manually correct the problematic measurements (the number of the links is in the order of hundreds).

Results granularity

The output of the model will not be precise traffic intensity values on each link. Rather each link will be classified into one of several **intensity intervals or classes**. This is reasonable since the actual numbers will be rather arbitrary due to the nature of the simulation process. The number of intervals and their exact size and choice will be described in an updated version of this document. The same approach has been taken with the reliability estimate where the actual numbers would

[7]This could, of course, be overcome. However due to time constraints and the questionable value it would bring we have decided not to implement it.

[8]E.g. links which have two or more conflicting measurements assigned to them, function class 5 links with measurements exceeding 8000 are assumed to be problematic.

be hard to interpret for users. The reliability of the estimate will be classified into around 10 classes.

Simulation phase

Sometimes the forward/backward runs generate contradicting flows for links where an AADF measurement is available. In such cases we have opted to modify the AADF measurement to a weighted sum of the forward/backward flow where the weight is given by the reliability of the respective flow. In practice the AADF measurement change will not be large, since the flow generated by this measurement will have much higher reliability than the other one.

Another technical detail concerns the case when two subnetworks are connected by one-way links. In these cases, flow along the connecting link will actually decrease flow in the source subnetwork. Although it is likely that the decrease will be balanced by an increase due to a different link connecting the networks in the other direction, we have no way of pairing these links together. To avoid inconsistencies in the case of one-way links we have to break the rule of not distributing traffic from one subnetwork to the other.

Complicated topology

Empirical results show that the model does not always work optimally in places where the topology is locally very complex. This is due to the fact that the balancing phase to minimise necessary changes prefers some connections over others when there is no empirical reason for this. Since these places are relatively rare (we have identified around 5 such situations) it is possible to adjust them manually. Moreover, the inaccuracies are limited to the actual place and do not affect surrounding traffic.

References

A. OUOROU, P. MAHEY AND J.-PH. VIAL, *A Survey of Algorithms for Convex Multicommodity Flow Problems*, Management Science, Vol. 46, pp. 126-147, 2000

BIN RAN, DAVID E. BOYCE AND LARRY J. LEBLANC, *A New Class of Instantaneous Dynamic User-optimal Traffic Assignment Models*, Operations Research, Vol. 41, pp. 192-202, January 1993

Boyce, D.E. and LeBlanc, L.J. and Chon, K.S., ***Network Equilibrium Models of Urban Location and Travel Choices: A Retrospective Survey***, Journal of Regional Science, Vol. 28, pp. 159-183, 1988

DP Bertsekas, LC Polymenakos, and P. Tseng, ***An epsilon-relaxation method for convex network optimization problems***, SIAM Journal on Optimization, Vol. 7, pp. 853-870, 1997

GARY A. DAVIS AND NANCY L. NIHAN, *Large Population Approximations of a General Stochastic Traffic Assignment Model*, Operations Research, Vol. 41, pp. 169-178, January 1993

H. Yang, Y. Iida and T. Sasaki, ***The Equilibrium-Based Origin-Destination Matrix Estimation Problem***, Transportation Research B, Vol. 28B, pp. 23-33, 1994

HAMDY A. TAHA, Operations Research: An Introduction, 8th edition, Prentice Hall 2006

J. MATOUŠEK, Lineární programování a lineární algebra pro informatiky, ITI 2006

JOHN R. BIRGE AND JAMES K. HO, *Optimal Flows in Stochastic Dynamic Networks with Congestion*, Operations Research, Vol. 41, pp. 203-216, January 1993

M.S. SODHI, *What about the 'O' in O.R.?*, OR/MS Today, Vol. , pp. 12, December 2007

P. Beraldi, F. Guerriero and F. Musmanno, ***Parallel Algorithms for Solving the Convex Minimum Cost Flow Problem***, Computational Optimization and Applications, Vol. 18, pp. 175–190, 2001

R. CHANDRASEKARAN, Network flow references, http://www.utdallas.edu/~chandra/documents/6381/books-2.pdf

R. De Leone, R. R. Meyer AND A. Zakarian, ***An e-Relaxation Algorithm for Convex Network Flow Problems***, Computer Sciences Department Technical Report, Vol. , pp. 430--466, 1995

RAWINDRA K. AHUJA, THOMAS L. MAGNANTI AND JAMES B. ORLIN, Network Flows: Theory, Algorithms, and Applications, , Prentice Hall 1993

SOREN S. NIELSEN AND STAVROS A. ZENIOS, *A Massively Parallel Algorithm for Nonlinear Stochastic Network Problems*, Operations Research, Vol. 41, pp. 319-337, 1993

T. Abrahamsson, ***Estimation of Origin-Destination Matrices Using Traffic Counts: A Literature Survey***, IIASA Interim Report IR-98-021/May, Vol., pp. , 1998

WIKIPEDIA AUTHORS, Operations research http://en.wikipedia.org/w/index.php?title=Operations_researc