HAL
open science

Link Dependent Origin-Destination Matrix Estimation : Nonsmooth Convex Optimisation with Bluetooth-Inferred Trajectories

Gabriel Michau

To cite this version:

Gabriel Michau. Link Dependent Origin-Destination Matrix Estimation : Nonsmooth Convex Optimisation with Bluetooth-Inferred Trajectories. Data Analysis, Statistics and Probability [physics.data-an]. Université de Lyon; Queensland University of Technology. Brisbane, Australie, 2016. English. NNT : 2016LYSEN017 . tel-01417805

HAL Id: tel-01417805

https://theses.hal.science/tel-01417805

Submitted on 16 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Doctorat de l'Université de Lyon
opérée par
l'École Normale Supérieure de Lyon

Numéro National de Thèse: 2016LYSEN017
École Doctorale N° 52
École Doctorale de Physique et d'Astrophysique de Lyon

Discipline: Physique,
Spécialité: Traitement du Signal

Doctor of Philosophy
opéré par :
the Queensland University of Technology

PhD Student: n8922039
Smart Transport Research Centre,
Civil Engineering and
The Built Environment School,
Science and Engineering Faculty,
Queensland University of Technology

Soutenue publiquement le *21/07/2016* par:

Gabriel MICHAU

Link Dependent Origin-Destination Matrix Estimation: Nonsmooth Convex Optimisation with Bluetooth-Inferred Trajectories

Estimation de Matrices Origine-Destination-Lien: Optimisation convexe et non lisse avec inférence de trajectoires Bluetooth

Devant le jury composé de :

Dr. Patrice ABRY	Directeur de Recherche CNRS, ENS de Lyon	Directeur
Dr. Ashish BHASKAR	Senior Lecturer in Civil Engineering, QUT	Co-encadrant
Dr. Pierre BORGNAT	Chargé de Recherche CNRS, ENS de Lyon	Co-encadrant
Pr. Edward CHUNG	Professor of Intelligent Transport, QUT	Directeur
Dr. Nour-Eddin EL-FAOUZI	Directeur de Recherche, LICIT, IFSTTAR	Rapporteur
Pr. Eric MOULINES	Professeur des Universités, École Polytechnique ParisTech	Rapporteur
Dr. Latifa OUKHELLOU	Directrice de Recherche, IFSTTAR	Présidente
Pr. Cédric RICHARD	Professeur des Universités, Univ. Nice Sophia-Antipolis	Examinateur

Acknowledgement

THE present thesis is the end-result of a three-year journey. It started in Brisbane, at the Smart Transport Research Centre, when Edward Chung suggested I could investigate how Bluetooth data could provide traffic information. What started as a short-term volunteer work soon appeared to be the opportunity for more advanced research on how the new technologies could potentially disrupt the traditional approaches to traffic engineering. Two months later, it turned into a PhD. The itinerary I followed brought every day new research questions and some of them are answered here. Among these challenges, early developments have naturally encouraged a strong collaboration with the signal processing field for solving challenging estimation problems and this collaboration materialised as a Joint-PhD with the Physics Laboratory at the *École Normale Supérieure de Lyon*, within the SiSyphe team. This journey, at the border between two fields: Transport Engineering and Signal Processing has been the occasion to collaborate with many amazing personalities that I would like to thank.

In particular, I would like to thank my five thesis supervisors for their support: At the *Queensland University of Technology*, I would like to thank Edward Chung for his advices when I freshly arrived in Brisbane, his suggestion for me to become his PhD student, his support (both financially and in the every day life, in particular with administrative requirements) and the flexibility he left me during those three years, including the choice of doing a co-tutelle. I thank Alfredo Nantes, my associate supervisor for the first two years of my PhD. His daily availability, his pertinent inputs on my projects, and his presence as a friend were important to me. I thank warmly Ashish Baskhar, who followed the development of my PhD since its early days, two years before becoming my new associate supervisor when Alfredo Nantes could not keep that role anymore. I valued his advices and his regular encouragements, in particular concerning my first scientific contributions. I thank him as well for presenting my first poster at TRB.

At the *ENS de Lyon*, I would like to thank Pierre Borgnat, my associate supervisor, for his enlightened supervision, for having encouraged this joint PhD, for his help with my various publications and for his very valuable inputs. Last, I thank deeply Patrice Abry in his role as principal supervisor. His strong efforts for making this joint PhD possible, his very strong involvement in all the steps of this thesis, and in particular his meticulous reading of my written publications and of this manuscript taught me a lot. This manuscript along with my journal papers, owe a lot to Patrice's patience and advices.

I thank my supervisors for having stayed involved in my work despite a difficult geographical situation and for the regular meetings that helped me to stay on tracks.

Many thanks to Nour-Eddin El-Faouzi and to Eric Moulines for having accepted to review my thesis and to Latifa Oukhellou and Cédric Richard for being part of my jury.

I thank deeply Nelly Pustelnick for her major inputs and for her support on the convex optimisation part of this work. She helped, in particular, with most of the algorithms developed here.

I acknowledge the Office of International Affairs at ENS de Lyon and the International Partnerships and Mobility Service at QUT for making the administrative existence of the joint-PhD possible.

I would like to thank all the people I had the chance to meet during this PhD, non-exhaustively: The team of the office 151 and more generally the SiSyphe team, the other PhD students (Christophe, Nicolas, Céleste, Alexandre, Jordan among others) and for the good atmosphere and the frequent lunches. I spent some very good time at the Physics Laboratory with its wonderful members. I thank my colleagues at QUT (Marc Miska, Kai Becker, Martin Peron, Aleksei, Takahiro among others) for the good time we had.

Listing all the people I would like to thank outside this academic environment would probably take too long but I am not forgetting my family, my old and my new friends.

Last, I thank Morgane for having taken me to Brisbane, for her daily support and for all we lived together.

Statement of Original Authorship

THE work contained in this joint thesis undertaken between QUT and ENS de Lyon has not been previously submitted to meet requirements for an award at these or any other higher education institution.

To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: *Gabriel* MICHAU

Date: July 18, 2016

Table of Contents

Acknowledgement **1**

Statement of Original Authorship **3**

Table of Contents **5**

List of Figures **10**

List of Tables **12**

List of Algorithms **13**

Notations **15**
1 Abbreviations . . . 15
2 Indexing . . . 16
3 Operators and functions . . . 16
4 Variables . . . 16

1 Introduction **19**
1 Context . . . 19
1.1 Traffic Estimation . . . 19
1.2 Signal Processing on Graphs . . . 20
1.3 Aims and Objectives . . . 21
2 Research Questions and Contributions . . . 21
3 Outline and Publications . . . 22
3.1 Outline . . . 22
3.2 Journal Papers . . . 22
3.3 International Conferences . . . 22
3.4 National Conferences and Workshops . . . 23

2 Traffic Estimation in Transport **25**
1 OD Matrix: Definition and Estimation . . . 27
1.1 Context . . . 27
1.2 Generalities on OD Matrix . . . 28
1.2.1 Static OD Matrix . . . 29
1.2.2 Dynamic OD Matrices . . . 30
1.2.3 OD Matrix Estimation: Principles . . . 30

1.3 OD Matrix Estimation: Review of the literature . . . 33
1.3.1 The Gravity Model . . . 33
1.3.2 Statistical Approaches . . . 34
1.3.3 The Assignment . . . 37
1.3.4 Dynamic OD Matrices Estimation . . . 38
1.3.5 The Role of the New Technologies . . . 38
1.4 Research Gaps Identified on OD Matrix Estimation . . . 40
2 Bluetooth as a New Source of Data for Trajectories . . . 42
3 Conclusions . . . 43

3 Data **45**
1 Road Network . . . 47
1.1 Nature - Principle . . . 47
1.2 Network Interpretation . . . 47
1.3 Network Simplification . . . 48
2 Traffic Counts . . . 49
2.1 Nature - Principle . . . 49
2.2 Formatting the Data . . . 51
3 Bluetooth . . . 51
3.1 Nature of the Data . . . 51
3.2 Characterising the Bluetooth Data . . . 55
3.2.1 The Inquiry Cycle . . . 55
3.2.2 Overlapping Detection Zones . . . 57
3.2.3 Penetration Rate of the Bluetooth . . . 59
3.2.4 Distribution and Dynamics . . . 60
3.2.5 MAC identifier . . . 61
3.2.6 Shared MAC address . . . 62
3.3 Missed Detections . . . 63
3.3.1 Evidence for Missed Detections . . . 63
3.3.2 Independent Hypothesis . . . 64
3.3.3 Origins for Missed Detections . . . 65
4 Taxi Data . . . 67
4.1 Nature . . . 67
4.2 Matching with Bluetooth Data . . . 68

4 Trajectories from Bluetooth Data **69**
1 Retrieving Trips form Bluetooth Detections . . . 71
1.1 From Monthly Data to Trips . . . 71
1.2 Application to the Brisbane Dataset . . . 73
2 Retrieving Trajectories from Bluetooth Trips . . . 74
2.1 Intuitions for Reconstruction . . . 74
2.2 A Spatially Constrained Shortest Path Algorithm . . . 75
3 Application to Brisbane Case Study . . . 76
3.1 Simulated Trajectories and detections . . . 77
3.1.1 Simulated Trajectories . . . 77
3.1.2 Simulated Detections . . . 78
3.1.3 Results of the Simulated Case Study . . . 78
3.2 Real Trajectories: Taxi Dataset . . . 80
3.3 Conclusion of the case studies . . . 82
4 Trajectories for Further Analysis of the Bluetooth Data . . . 82

4.1 Speed Distribution . . . 82
4.2 Discriminating Mode of Travel . . . 82
4.3 Missed Detections: Binomial and Gaussian Mixture Models . . . 84
4.3.1 Bluetooth equipped Vehicles Detection Probability Distribution . . 84
4.3.2 Expectation Maximisation (EM) Algorithm for Mixture of two Distribution . . . 86
4.3.3 Binomial Mixture Model on Retrieved Trajectories . . . 88
4.3.4 Gaussian Mixture Model on Retrieved Trajectories . . . 89
4.3.5 BMM and GMM on Corridors: Bias and Correction of the Mixture Coefficient . . . 90
4.4 Combining Trajectories with other Datasets . . . 92

5 Link dependent Origin Destination Matrices 93
1 From OD Matrix to LOD Matrix Estimation: Summary of the Problem . . . 95
2 Road Network and LOD matrix . . . 96
2.1 Problem Statement . . . 96
2.2 Road Network as a Graph . . . 96
2.3 Model, Measures and Estimates . . . 98
3 Functional Optimisation Formulation . . . 98
3.1 Objective Function . . . 98
3.1.1 Traffic Count Data Fidelity f_{TC} . . . 99
3.1.2 Poisson Bluetooth Sampling Data Fidelity f_P . . . 99
3.1.3 Consistency Constraint f_C . . . 99
3.1.4 Kirchhoff's Law f_K . . . 100
3.1.5 Total Variation f_{TV} . . . 100
3.2 Algorithm . . . 101
4 Simulated Case Study . . . 103
4.1 Experimental setup . . . 103
4.1.1 Simulation context . . . 103
4.1.2 Algorithmic parameter setup . . . 104
4.1.3 Performance evaluation . . . 105
4.2 Performances . . . 105
4.2.1 Impact of the Regularisation Parameters . . . 105
4.2.2 Impact of each Objective Function . . . 107
4.3 Lower Time Granularity . . . 109
5 Alternative Formulations of the Problem . . . 109
5.1 Optimisation on Bluetooth Penetration Factors . . . 109
5.2 A simple Forward-Backward approach . . . 110
6 Conclusion . . . 110

6 Brisbane Case Study 113
1 From Toy Models to Large Real Case Studies . . . 115
1.1 Adapting the Datasets . . . 115
1.2 Objective Function for the Real Case Study . . . 116
1.2.1 Notations and Definitions . . . 116
1.2.2 Traffic Count Data Fidelity f_{TC} . . . 116
1.2.3 Poisson Bluetooth Sampling Data Fidelity f_P . . . 116
1.2.4 Consistency Constraint f_C . . . 117
1.2.5 Kirchhoff's Law f_K . . . 117
1.2.6 Total Variation f_{TV} . . . 117

1.3 Algorithm . . . 118
2 LOD Matrix in Brisbane . . . 118
2.1 Size of the Problem . . . 118
2.2 Estimation and Assessment of Brisbane LODM . . . 120
2.3 Example of Traffic Map from LOD matrix . . . 121
3 Conclusion . . . 123

7 Conclusion and Perspectives 125

A Algorithm for Network Simplification 129

B Expectation Maximisation Estimation 131
1 Generalities on the EM algorithm . . . 131
2 Two Binomial Mixture Model . . . 134
3 Two Gaussian Mixture Model . . . 135

C Optimisation on Bluetooth Penetration Factors 137
1 Objective Function . . . 138
1.1 Consistency with Traffic Counts . . . 138
1.2 Consistency with Bluetooth Sampling . . . 138
1.3 Definition Domain . . . 138
1.4 Kirchhoff's Law . . . 139
2 Algorithm . . . 139
3 Simulated Case Study . . . 140

D A simple Forward-Backward approach 141
1 Objective Function . . . 142
1.1 Consistency with Traffic Counts . . . 142
1.2 Definition Domain . . . 142
1.3 Kirchhoff's Law . . . 142
1.4 Poisson Bluetooth Sampling Data Fidelity . . . 143
2 Algorithm . . . 143
3 Case Study and Results . . . 144

E Flow Conservation and Kirchhoff's law 147
1 No Cycle Hypothesis . . . 147
1.1 Outgoing Flows at Destination and Incoming Flows at Origins . . . 147
1.2 No Cycle Hypothesis : From LOD to OD Matrix . . . 147
2 Proposition: Local Kirchhoff's law induces other relationships . . . 148
2.1 Local Kirchhoff's law induces Equation (5.3) . . . 148
2.2 Local Kirchhoff's law induces the Global Kirchhoff's law . . . 149

Bibliography 151

List of Publications 167

Résumé Long du Manuscrit 169
1 Introduction . . . 169
1.1 Ingénierie du Trafic . . . 169
1.2 Traitement du Signal sur Réseaux . . . 170
1.3 Objectifs . . . 171

1.4 Contributions 171
2 Résumé des Chapitres 172
2.1 Chapitre 2 172
2.2 Chapitre 3 172
2.3 Chapitre 4 172
2.4 Chapitre 5 173
2.5 Chapitre 6 173
3 Conclusion 174

Abstract - Résumé **179**

List of Figures

2.1 OD Matrix - Definition . . . 28
2.2 OD matrix Estimation: Surveys and Models . . . 31
2.3 Traditional OD matrix Estimation Process . . . 33
2.4 Traffic Description Tools Example . . . 41
2.5 Traditional and Proposed Framework for LOD matrix Estimation . . . 42

3.1 Network Simplification: Brisbane . . . 50
3.2 Map of Brisbane and Bluetooth Scanners . . . 54
3.3 Example of Bluetooth Detection Sequences . . . 54
3.4 Histogram of Bluetooth Timestamps . . . 56
3.5 Histogram of the Duration Field . . . 56
3.6 Time Intervals for Detections of the Same MAC ID at the Same BMS . . . 57
3.7 Example of Overlapping Detection Areas . . . 58
3.8 Bluetooth Penetration Rate . . . 59
3.9 Brisbane Traffic Dynamic from Bluetooth Data . . . 60
3.10 Time and Speed distribution: Analysis of the Bluetooth Data . . . 61
3.11 Identification of the Main Bluetooth Manufacturers . . . 62
3.12 Example of Shared MAC ID . . . 63
3.13 Five Corridors in Brisbane and Detection Probability . . . 64
3.14 Probability of Detection under Independence Assumption . . . 65
3.15 Characteristics of Shared MAC IDs . . . 66
3.16 Example of Taxi Trajectories . . . 67

4.1 Framework for Trajectory Recovery . . . 71
4.2 Inter-Detection Time Intervals Distribution . . . 72
4.3 Number of Resulting Trips against Threshold Time . . . 73
4.4 Illustration of the CSP Algorithm . . . 77
4.5 Inner Brisbane Road Network for the Simulated Case Study . . . 78
4.6 Results of the Trajectory Recovery Method on the Simulated Case Study . . . 79
4.7 Results of the Trajectory Recovery Method on the Taxi Case Study . . . 81
4.8 Brisbane Traffic Speeds from Bluetooth Trips . . . 83
4.9 Brisbane Traffic Speeds: 3 Sets of Trajectories . . . 84
4.10 Bluetooth Device Detection Probability Distribution . . . 85
4.11 Average Bluetooth Device Detection Probability . . . 86
4.12 Binomial Mixture Models of the Detection Probability . . . 89
4.13 Gaussian Mixture Models of the Detection Probability . . . 90

4.14 Mixture Models on Waterworks Road 92

5.1 Topology and Traffic Description Tools 97
5.2 Simulated Road Network and Traffic 104
5.3 RMSE and EMD Value Evolution 106
5.4 Distribution of LOD Matrix Values for Four Estimates 107

6.1 Traffic Counts in Brisbane 119
6.2 Bluetooth Assignment Example 119
6.3 Distribution of OD traffic flows 121
6.4 Traffic Counts in Brisbane 122
6.5 OD Flows for Bradfield Highway Bridge 122
6.6 From CBD to Moorooka: Road usage 123

List of Tables

2.1 Terms to Refer to OD Matrix 29

3.1 BCC's BLUETOOTH Table Example 52
3.2 BCC's AREA Table Example 52
3.3 BCC's LOCATION Table Example 53
3.4 BCC's DEVICE Table Example 53

4.1 Notation Summary 75
4.2 Results of the Trajectory Recovery Method on the Simulated Case Study 80
4.3 Results of the Trajectory Recovery Method on the Taxi Case Study 81
4.4 Non-Motorised Mode: Trajectory Lengths versus Threshold Speed 83
4.5 Mixture Models on Waterworks Road 91

5.1 LOD Matrix Estimation: Results for Optimal Estimates 106
5.2 Results for Partial Objective Function (1 or 2 Terms) 108
5.3 Results for Partial Objective Function (3 or 4 Terms) 109
5.4 Best Achieved Results with $N = 10\,000$ 109

6.1 LOD Matrix Estimation: Results for Brisbane Case Study 120

D.1 LOD Matrix Estimation: Results for the 3 Approaches ($N = 100\,000$) 144
D.2 LOD Matrix Estimation: Results for the 3 Approaches $N = 10\,000$ 145

List of Algorithms

4.1 Bluetooth Data to Trip Sequencing Algorithm 73
4.2 EM Algorithm for the Binomial Mixture Model 87
4.3 EM Algorithm for the Gaussian Mixture Model 88
5.1 Proximal Primal Dual Algorithm for LOD Matrix Estimation 103
A.1 Procedure for Simplifying the Road Network 129
C.1 Forward Backward Algorithm for LOD Bluetooth Penetration Factor Estimation . . . 140
D.1 Forward Backward Algorithm for LOD Matrix Estimation 144

Notations

1 Abbreviations

Abbr.	Meaning
Entities	
BCC	Brisbane City Council
CNRS	Centre National de la Recherche Scientifique *(National Center for Scientific Research)*
ENS de Lyon	École Normale Supérieure de Lyon
QUT	Queensland University of Technology
STRC	Smart Transport Research Centre
Concepts	
AVI	Automated Vehicle Identification
BMM	Binomial Mixture Model
BMS	Bluetooth Media access control address Scanner
CBD	Central Business District, the commercial center of a city
CSP	Constrained Shortest Path
EM	Expectation Maximisation
EMD	Earth Mover's Distance
GMM	Gaussian Mixture Model
GPS	Global Positioning System
ITS	Intelligent Transport Systems
kph	Kilometres per hour
LOD	Link dependent Origin-Destination
MAC (address)	Media Access Control address
MAC ID	Media Access Control IDentification address
NNLLh	Normalised Negative Log-Likelihood
OD	Origin-Destination
PDF	Probability Density Function

PhD	Doctor of Philosophy
RMSE	Root Mean Square Error
TV	Total Variation

2 Indexing

Within formulas and equations, when referring to variables defined along with a graph:

- Subscript indices are preferred for dimensions over the nodes of the graph (e.g., scanners, intersections, ...).
- Similarly, superscript indices are used for dimensions over the links of a graph (e.g., roads).

Index	Meaning
i	Preferred notation for indexing nodes (*cf.* V) when referring to origins
j	Preferred notation for indexing nodes (*cf.* V) when referring to destinations
k, m, n & p	Alternative notation for indexing nodes (*cf.* V), without particular meaning
l	Preferred notation for indexing links (*cf.* L)
e	Alternative notation for indexing links

3 Operators and functions

- $\underline{X}$, $\underline{\underline{X}}$ and $\underline{\underline{\underline{X}}}$ respectively refer to vectors, matrices and tensors.
- The Hadamard (element-wise) product of $\underline{\underline{\underline{C}}}$ and $\underline{\underline{\underline{X}}}$ is denoted $\underline{\underline{\underline{C}}} \circ \underline{\underline{\underline{X}}}$.
- The element-wise division of $\underline{\underline{\underline{C}}}$ and $\underline{\underline{\underline{X}}}$ is denoted $\underline{\underline{\underline{C}}}./\underline{\underline{\underline{X}}}$.
- The symbol $\bullet$ is used to denote the dimension that does not contribute to a sum: e.g., the sum over first and third dimensions is written $\sum_{i\bullet l} \underline{\underline{\underline{X}}}$.
- We denote by $\|\cdot\|_1$ the element-wise first norm for matrices: e.g., $\|\underline{\underline{X}}\|_1 = \sum_{ij} |X_{ij}|$.
- The notation $\tilde{}$ (as in $\tilde{X}$) is used preferentially for measured variables.
- The notation * (as in $X^\star$) is used preferentially to refer to the original variable (or true variable).
- The notation $\widehat{}$ (as in $\widehat{X}$) is used preferentially for estimates.

4 Variables

Notation	Meaning
Road Network and Graph	

$\mathscr{G} = (V, L)$	Oriented graph representing the road network
$V = \{v_k\}_{k \in \lvert V \rvert}$	Set of nodes v_k (traffic intersections are nodes)
N_V	Number of elements in V ($N_V = \lvert V \rvert$)
L	Set of directed edges. An edge is a direct itinerary linking two nodes.
N_L	Number of elements in L ($N_L = \lvert L \rvert$)
$l(v_k, v_m)$	Edge in L linking $v_k \in V$ to $v_m \in V$
$w_{l(v_k,v_m)}$	Length of $l(v_k, v_m)$
$\underline{\underline{I}} = \left(I_k^l\right)_{k \in V, l \in L}$	*Incidence* matrix. $I_k^l = \begin{cases} 1 & \text{if the edge } l \text{ is arriving to the node } k, \\ 0 & \text{otherwise.} \end{cases}$
$\underline{\underline{E}} = \left(E_k^l\right)_{k \in V, l \in L}$	*Excidence* matrix. $E_k^l = \begin{cases} 1 & \text{if the edge } l \text{ is starting from the node } k, \\ 0 & \text{otherwise.} \end{cases}$
$\underline{\underline{I}} = \left(I_k^l\right)_{k \in V, l \in L}$	*Incidence* matrix. $I_k^l = \begin{cases} 1 & \text{if the edge } l \text{ is arriving to the node } k, \\ 0 & \text{otherwise.} \end{cases}$
$\underline{\underline{E}} = \left(E_k^l\right)_{k \in V, l \in L}$	*Excidence* matrix. $E_k^l = \begin{cases} 1 & \text{if the edge } l \text{ is starting from the node } k, \\ 0 & \text{otherwise.} \end{cases}$
r	Scanning radius of the Bluetooth detectors
$S = \{s_k\}_{k \in \lvert S \rvert}$	Set of scanners on the network
$\mathscr{M}_r^V$	The mapping from the space of scanners S to a space of nodes V
$\mathscr{M}_r^V(s) = V_{r,s}$	Set of nodes in V within r to scanner s
$\{\mathscr{M}_r^V\}_{s \in S}$	Set of nodes in V within r of any scanner in S
Notations Specific to Traffic	
N	Number of individual users on the road network
$\underline{\underline{T}} = (T_{ij})_{(i,j) \in V^2}$	Origin Destination Matrix
$\underline{q} = \left(q^l\right)_{l \in L}$	Traffic volumes on roads (e.g., measured by magnetic loops)
$\underline{D} = (D_j)_{j \in V}$	Destinating flows to nodes in V
$\underline{O} = (O_i)_{i \in V}$	Originating flows from nodes in V
$\underline{\underline{\underline{Q}}} = \left(Q_{ij}^l\right)_{(i,j) \in V^2, l \in L}$	Link dependent Origin Destination Matrix
$\underline{\underline{\underline{\tilde{B}}}} = \left(B_{ij}^l\right)_{(i,j) \in V^2, l \in L}$	Bluetooth Link dependent Origin Destination Matrix
Notations specific to the Spatially Constrained Shortest Path (Chapter 4)	
d_i	Sequence of n detections for i-th car
$= (s_i, t_i, \delta_i)_{i \in [1,n]}$	(s_i: scanner, t_i: timestamp, δ_i: duration)
Δ	Time threshold for Bluetooth detections sequencing
$p(v_k, v_m)$	Shortest path between nodes v_k and v_m
$d(v_k, v_m)$	Length of the shortest path from v_k to v_m
Π_u	Sequence of observed detections for user u
$\Pi(p)$	Sequence of scanners within r of any node in p
r_{sim}	Detection Radius used in the simulated case study
Notations specific to the Mixture Models (Chapter 4)	

$X = (G, H)$	Set of Bluetooth Detectors Sequences for devices
$G = (g_i)_{(i \in [1:N])}$	Number of actually observed detections in the sequence
$H = (h_i)_{(i \in [1:N])}$	Number of detectors along the trajectory of device i
$Nmax$	Maximum number of detector in H
$\Pi = (\pi_1, \pi_2)$	Mixture coefficient for two-distribution mixtures
$\Theta = (\theta_1, \theta_2)$	Binomial Distribution parameters
$\Theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$	Gaussian Distribution means (μ) and variances (σ)
Notations specific to Chapter 5	
$\underline{\underline{\underline{\alpha}}} = \left(\alpha_{ij}^{l}\right)_{(i,j) \in V^2, l \in L}$	Link origin destination Bluetooth Penetration Factor
$\underline{\underline{\eta_o}} = \left(\eta_{0ij}\right)_{(i,j) \in V^2}$	OD Bluetooth penetration rate
$\underline{\tilde{\eta}}$	Link penetration rate, estimated from traffic counts and Bluetooth
$\underline{\varepsilon}$	Noise on traffic count measures

CHAPTER 1

Introduction

1 Context

1.1 Traffic Estimation

THIS research emerged in 2013, at the *Smart Transport Research Centre* in Brisbane, where since few months, an extensive dataset collected by Bluetooth scanners spread over most of Brisbane City was available. While according the City Council's engineers, the primary objective of these detectors was the measurement of travel times, it soon appeared that such an extensive dataset would be very valuable for other purposes in transportation studies. More generally around the world, in transport engineering as in many other fields, new technologies have made possible the collection of huge datasets, giving its name to the current period as the *Big Data era*.

For the transport field and until recent years, traffic studies were traditionally based on data stemming from surveys and magnetic loops. Nowadays, new technologies are in use for data collection: GPS (through specific data collection devices or through mobile applications), Video (with or without plate recognition), electronic tagging, and LiDAR, among others. They allow for direct and automated vehicle identification, while other technologies such as mobile phone call detail recording, Bluetooth and WiFi device detection aim for the detection of in-car devices. All these technologies allow for the identification of the detected vehicles (or objects giving an identity to the car in which they are). They are thus usually referred to as Automated Vehicle Identification (AVI) systems.

The scales at which these technologies are spreading, are constantly increasing. To give a sense of this evolution, the Bluetooth detector network in Brisbane started with a pilot project of one detector in 2007, had around 600 detectors in 2014 and consists of more than 900 Bluetooth detectors today.

Often, the primary objective of these new technologies is to provide precise and complementary ways of monitoring the traffic. Yet, the new technologies, especially thanks to their identification abilities, can also bring valuable data for other transport applications in traffic engineering.

Traffic engineering objectives can be divided in two classes: First, the questions related to *traffic condition* aim to understand the usage of a road network by means of several indicators as speed, travel time, density, volumes among others and to infer the relationships linking these quantities. Second, the problems in the *transport demand* category aim to understand the factors driving the mobility, but also to quantify and to qualify this mobility. The two major questions in transport demand are, first, the estimation of origin-destination matrix (OD matrix), that is, of a table taking census of

the origins, the destinations and the volumes of the traffic flows, and second, the analysis of route related problems, e.g., route choice modelling and path estimation. The process that connects OD matrix and routes, constrained by *transport supply*, or road infrastructures, is called the *assignment*. It is only when these two problems of OD matrix estimation and route estimation are combined through an assignment procedure, that transport demand can eventually be compared with traffic condition. This highlight the difficulty of transport demand problems: the estimation of transport demand usually requires assumptions and models that are calibrated after comparison with traffic data. This comparison relies on an assignment model, also calibrated on those same traffic data. These problems being strongly interdependent, the reliability of the OD matrix estimation depends on the reliability of the other models involved.

The new technologies, by providing new datasets with identified vehicles, permit to revisit these problems and to question the separation between traffic condition estimation and traffic demand estimation. The new data can jointly provide information on transport demand, with origin, destination and trajectory of detected users, while measuring some traffic condition indicators, e.g., speed and travel time. Notably, the trajectories, if used as primary dataset for traffic demand estimation, are the opportunity to reformulate both the OD matrix estimation and the assignment problems into a single problem. This approach is introduced in this manuscript and is referred to as link dependent origin destination matrix (LOD matrix) estimation. This problem relies on the possibility to have trajectories for some cars. However, with the exception of the GPS technology, most technologies are based on detector network spread along the roads, thus providing point measures only. A second issue is then how to reconstruct trajectories from point measures.

1.2 Signal Processing on Graphs

The traffic demand estimation problem fits within the generic problem of estimating unknown quantities from partial observations or from the observation of aggregated values (e.g., traffic volumes on roads), which is commonly referred to, in the Signal Processing field, as an *inverse problem*. Moreover, this problem is constrained by an underlying infrastructure, the road network. Thus, traffic demand estimation is a good candidate for the subcategory of *inverse problems on graphs*. Indeed, one can easily interpret a road network as a directed graph were intersections would be represented as nodes and direct itineraries from one intersection to another would be the links (or edges), whose direction corresponds to the allowed traffic direction.

The OD matrix estimation problem is about estimating volumes of traffic between each pair of nodes of the graph, either from surveys, which only provide partial observations, or from automatically collected data, that consisted only, up to a recent point, of traffic counts on links. The number of unknown quantities and the number parameters involved in the underlying models (e.g., the assignment) being more numerous than the number of observations, this problem is ill-posed. It thus needs regularisation through extra information, for example on the distributions of the data.

Such analogies between real world questions and problems on graphs are not uncommon, e.g., the internet traffic analysis leads to the same question of origin-destination traffic estimation based on observation collected on the physical internet network and has been treated, in the past litterature, as an inverse problem on graph as in [1]–[4]. Another example, in biology, is the observation of gene expression from which gene regulation relationships (analogue to origin-destination relationships) need to be infered (see for example [5], [6]). The most common application of inverse problems are found in image restoration, e.g., where the value of missing or blurred pixels is to be estimated by inference from neighbour pixels [7]. By direct analogy, missing pixels would be road users for which travel information is missing while neighbour pixels would be users for which the trajectory could

have been retrieved, and thus the reconstructed image is the link dependent origin destination matrix for the whole set of users.

Recent advances in Signal Processing led to the developement of efficient algorithm for solving multiobjective convex functions, as an extension of the traditional gradient descent algorithm [8], [9], with the capacity to handle convex but non necessarily differentiable functions. For these functions, a workaround is the use of proximal operator [10], [11] which generalises the notion of derivative for convex, proper and semi-continuous functions. These recent methods give therefore a new freedom in term of objective function design as they handle a much larger variety of functions than traditional ones and are gaining a lot of interest for minimisation problems.

1.3 Aims and Objectives

This research proposes to take benefit of new technologies for the estimation of traffic demand. It aims first, to develop a method to retrieve vehicles trajectories from point-measures with vehicle identification capabilities. Second, it proposes a new formulation of the traffic demand estimation problem: it extends the concept of the classical origin-destination matrix to the one of link dependent origin destination matrix. It addresses the regularisation problem by proposing two types of regularisation function: functions quantifying the fidelity between the estimates and the measures and functions modelling inherent properties of traffic over networks. This allows for the LOD matrix estimation problem to be formulated as a nonsmooth convex optimisation problem and to propose, for its resolution, an efficient algorithm based on the recent advances in Signal Processing [12], [13].

2 Research Questions and Contributions

This research aims to show how new technologies, such as Bluetooth, permit to formulate both problems of OD matrix estimation and assignment as a LOD matrix estimation problem:

1. This research aims to keep a strong bond with the real case study of the city of Brisbane. Using the Bluetooth technology as a traffic data collection system, this research directly raises the challenges of dealing with such data. First, it aims to characterise the quality and statistical properties of the data.

2. A second contribution is to propose a method for efficiently extracting travel information from the Bluetooth data. This method retrieves the Bluetooth OD matrix and the trajectories of the Bluetooth equipped users. Using information inferred from trajectories, the Bluetooth data properties are further characterised.

3. It proposes, then, a reformulation of the OD matrix and assignment estimation problems into a single problem: the estimation of link dependent origin destination matrix.

4. It formulates the LOD matrix estimation process as an optimisation problem. To do so, it proposes an objective function built upon important properties of the problem. These properties can represent the fidelity between measures and estimates or stem from properties of the traffic on a graph (e.g., car conservation).

5. It develops a framework for solving the LOD matrix estimation problem using adapted algorithms from advanced signal processing tools.

6. It realises the proof of concept of the framework by applying it on simulated toy-model road network.

7. It illustrates LOD matrix estimation, by applying the trajectory retrieval procedure and by estimating the LOD matrix, as proposed in this work, to the real case study of Brisbane.

3 Outline and Publications

3.1 Outline

To that end, the manuscript is organised as follow:

Chapter 2 restates the basics and the history of the transport demand estimation, for both the OD matrix estimation problem and the assignment.

Chapter 3 presents the datasets at hand and their characterisation and the low-level processing required for their use.

Chapter 4 proposes a methods to extract the travel information of Bluetooth equipped users by first, recovering the OD matrix and, second, by recovering the trajectories.

Chapter 5 presents the formulation of the link dependent origin destination estimation problem as an optimisation problem and proposes a method to solve it.

Chapter 6 illustrates the benefits of all the previous contributions for traffic analysis by applying the results of the previous chapters to the Brisbane dataset.

3.2 Journal Papers

G. Michau, A. Nantes, A. Bhaskar, E. Chung, P. Borgnat, and P. Abry, "Bluetooth data in urban context: Retrieving vehicles trajectories", *Submitted in IEEE Transaction on Intelligent Transport Systems*, 2016

> This article presents some important characteristics of the Bluetooth data in urban context as in Chapter 3, Section 3 and proposes a method for Bluetooth OD matrix estimation and trajectories retrieval. The results submitted in this article are mostly the matter of Chapter 4.

G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, A. Bhaskar, and E. Chung, "A primal-dual algorithm for link dependent origin destination matrix estimation", *Submitted in IEEE Transactions on Signal and Information Processing Over Networks*, 2016

> This publication formalises the LOD matrix estimation process and presents a proximal primal-dual algorithm for estimating LOD matrix. The results submitted in this article are presented in Chapter 5.

3.3 International Conferences

G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving dynamic origin-destination matrices from Bluetooth data", in *Transportation Research Board, 93rd Annual Meeting*, Washington DC, Jan. 12–16, 2014. [Online]. Available: http://eprints.qut.edu.au/66511/

This conference paper presents preliminary results on the retrieval of travel information from Bluetooth data. Those preliminary works were developed to give rise to the results presented in Chapter 3, Section 3 and in Chapter 4.

G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 19–24, 2015, pp. 5480–5484. DOI: 10.1109/ICASSP.2015.7179019

This contribution consists in a preliminary version of a framework for solving the LOD matrix estimation problem. The contributions of this paper are detailed in Appendix C.

3.4 National Conferences and Workshops

G. Michau, A. Nantes, and E. Chung, "Towards the retrieval of accurate OD matrices from Bluetooth data: Lessons learned from 2 years of data", in *36th Australasian Transport Research Forum (ATRF)*, QUT, Brisbane, Australia, Oct. 4, 2013. [Online]. Available: http://eprints.qut.edu.au/62727/

This contribution presents the first results obtained from Bluetooth data analysis. Those results led, later on, to those presented in Chapter 3, Section 3 and in Chapter 4.

G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving trip information from a discrete detectors network: The case of Brisbane Bluetooth detectors", in *CAITR*, Sydney, Feb. 17–18, 2014. [Online]. Available: http://eprints.qut.edu.au/83110/

This contribution, build on the previous contribution on Bluetooth data analysis, raised the supplementary question of handling the road network underlying the traffic. This led to the contributions presented in Chapter 3, Section 1 in addition to those presented in Chapter 3, Section 3 and in Chapter 4.

G. Michau, P. Borgnat, N. Pustelnik, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", presented at the XXV GRETSI, Lyon, France, Sep. 8, 2015. [Online]. Available: http://eprints.qut.edu.au/86449/

This contribution consists in a second preliminary version of a framework for solving the LOD matrix estimation problem. The contributions of this communication are detailed in Appendix D.

G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Graph Signal Processing Workshop*, Philadelphia, May 25–27, 2016

The results of Chapter 5 and Chapter 6 will be presented at this workshop.

G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Complex Networks*, Marseilles, France, Jul. 11–13, 2016

The results of Chapter 5 and Chapter 6 will also be presented at this workshop.

CHAPTER 2

Traffic Estimation in Transport

Contents

1 OD Matrix: Definition and Estimation **27**
1.1 Context 27
1.2 Generalities on OD Matrix 28
1.3 OD Matrix Estimation: Review of the literature 33
1.4 Research Gaps Identified on OD Matrix Estimation 40
2 Bluetooth as a New Source of Data for Trajectories **42**
3 Conclusions **43**

1 OD Matrix: Definition and Estimation

1.1 Context

WHEN one looks at the definition of traffic, it appears that traffic is "*The passing to and fro of persons, or of vehicles or vessels, along a road, railway, canal, or other route of transport.*"[1]. Thus, traffic in urban context indicates the set of users along with their modes and their trajectories. When transport engineers try to gain an understanding of the usage of a network for further applications such as, traffic light plan optimisation, infrastructure need forecast, traffic condition analysis or prediction, providing them with a full set of timestamped trajectories would be answering most of their needs.

This information however has, up-to-date, never been available. Even if one could imagine some extensive surveys, designed such that the users could provide very detailed information on their trajectories, those surveys would probably have little reliability on the timestamps and would have a prohibitive cost. Moreover, survey capture stated behaviour [23]–[27], as opposed to observed behaviour captured by automated field data collection and might therefore be biased by the subjective perception of the user of its own journey. In particular, it was demonstrated in 2013 that an important fraction of the Off-peaks travels were missing from varied HTS (Household Travel Surveys) conducted throughout Australia [28]. Last, surveys have a the limited lifetime and need to be updated frequently.

Faced to the impossibility of a full description of the traffic with trajectories, the traffic description has traditionally been broken down to a two level descriptions: traffic demands and traffic conditions. Traffic demand aims to describe and quantify the movements of the users. That is: how many users, where do they start, where do they go and by means of which roads. The classical tools for its representation are the origin-destination (OD) matrix and route choice models. Its estimation can be done through observations (e.g., surveys) or stems from some models, which will also require observations for their calibration.

However, apart from surveys, most traffic observations are designed for traffic condition measurements. Traffic condition indicators are travel time, average speed and road density, for which, link counts, registration plate detection and other detection technology can bring valuable information.

Here, a paradox arises: direct observation of traffic demand is, at this time, not possible, or for part of the traffic only and thus statistical inferences or models are needed for its estimation. In order to calibrate those methods however, one has to compare the results with observations on traffic conditions and to do so, trajectories that allow the transcription of the demand in term of traffic flows, are needed. In other words, the OD matrix is estimated because trajectories can not be observed, yet its estimation process requires it to be converted into trajectories.

Very recently, new technologies have given the opportunity to access more detailed information. At first, due to their cost, those technologies were scattered on the networks and thus would only prove useful for travel time and volume estimation, thus getting redundant with traditional way of observing traffic (yet with higher precision or lower price). Nowadays, some technologies, such as GPS, Bluetooth, WiFi, Cellulars have become popular and cheap so that one could hope for the retrieval of detailed timestamped trajectories with satisfying resolution. Yet this is for a small fraction of the users only. At first, many works have used this information as an additional way to calibrate OD matrix estimation process. In this work, we propose here to rethink the objective of the OD matrix estimation which is, ultimately, just a tool for representing the traffic demand used for lack of a more

[1] Oxford English Dictionary, http://www.oed.com/

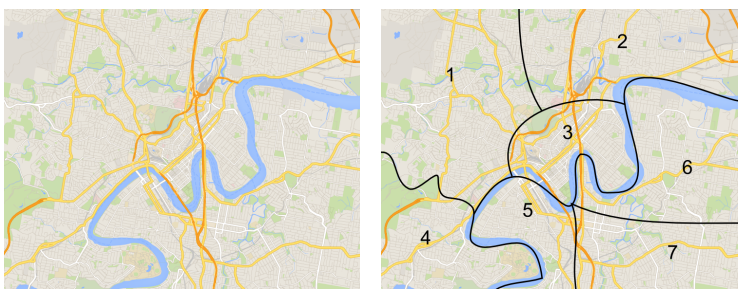
detailed representation (yet already a difficult and challenging problem when only traffic counts are available). Thus, very similarly to many works that have tried to generalise prior OD information to that of the whole set of users on a network, here, we propose a method to generalise the trajectory information retrieved from new technologies to the whole set of users. Therefore, for the reason that the OD matrix is not adapted for such information as all link information is lost when aggregating the data in OD flows, we propose the use of a new tool, the link dependent OD matrix which directly represents and provides, for each OD, the assignment on the links and, for each link, OD information on its users.

On the ground that this work proposes a new framework, static travel demand will be the primary focus, and the dynamic aspect left for future discussions.

1.2 Generalities on OD Matrix

Origin destination matrix is one of the most important tools used by transport engineers for its being a powerful indicator of the travel demand. This concept can be traced back at least to the late fifties, e.g., Reilly (1959) [29] modelled OD tables for representing volumes of goods traded between cities. If the context seems different, the aim was actually the same, estimating the movement of agents (people, cars, goods) over a particular region.

OD matrices are defined for an area of interest which is usually partitioned into smaller geographic zones and for a particular period of time. They are two-dimensional matrices, with rows and columns denoting the origin and destinations, respectively. The elements of these matrices are the census of the traffic volume, from origin zones, indexed by the rows, to destination ones, indexed by the columns. Figure 2.1 illustrates the concept of OD matrix.



(a) (b)

$$
\begin{array}{c|ccccccc|c}
\text{Zones} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \Sigma \\
\hline
1 & T_{1,1} & & & \cdots & & & T_{1,7} & \Sigma_j T_{1,j} \\
2 & & & & & & & & \\
3 & & & & & & & & \\
4 & \vdots & & & \ddots & & & \vdots & \vdots \\
5 & & & & & & & & \\
6 & & & & & & & & \\
7 & T_{7,1} & & & \cdots & & & T_{7,7} & \Sigma_j T_{7,j} \\
\hline
\Sigma & \Sigma_i T_{i,1} & & & \cdots & & & \Sigma_i T_{i,7} &
\end{array}
$$

(c)

Figure 2.1: (a) Case study area (b) Divbision of the area (c) Definition of the OD matrix. Each element T_{ij} represents the volume of traffic from zone i to zone j. The sum over each column corresponds to the traffic leaving the zones. Similarly, the sum over the rows takes census of the traffic having the zone for destination.

OD matrices can have many definitions and many names. Almost every combination of the terms inventoried in the Table 2.1 can refer to similar concepts. Subtle differences of meaning can appear, depending on the aim of such tables, to describe either a travel demand (the theoretical demand without any network consideration or the demand constraint by the network), the actual or the predicted use of the network. However, OD matrices often refer to all of them, depending on the context.

As a consequence, depending on the objective of the research, OD matrices, if always tacking census of flows between two locations, may describe different characteristics. Although the following

Table 2.1: Terms to refer to OD matrix

Origin-Destination OD Traffic Trip Journey Travel Vehicular	matrix table demand distribution split pattern flow volume

description does not aim to give an exhaustive description of all the existing definition, it appears to be important to present the main ones and what their implications are.

1.2.1 Static OD Matrix

Static OD matrices are how OD matrices were first described. As a traffic description tool, OD matrix are obviously defined over a given period of time. However, static OD matrix implies that the studied temporal period is considerably longer than the average length of a trip.(e.g. day, month, ...) so that boundary effects are mitigated. Yet, few differences might arise from how the boundaries are handled: The matrix can either take census of the trips that started in the considered interval or that existed in this interval (even if started of finished outside the interval boundaries). Another source of difference arise from the definition of origins and destinations:

- *Point-to-point OD matrix*: If the geographic area related to the matrix is described through a finite number of points (e.g. intersections, detectors, ...) then the OD matrix has a very clear meaning: It takes the census of the traffic between two points of the networks. It is very useful for data gathered by detectors or field surveys as it is straightforward to process the matrix. However, the major drawback is that a lot of information, as for example from traffic happening on alternative, non equipped or surveyed roads, will not be part of the traffic described by the matrix. The matrix gives therefore a very meaningful and clear information on the chosen points of the network, nonetheless, some information might be missing.

- *Zone-to-zone OD matrix*: To the opposite, if the area is cut in several smaller geographic zones, then the matrix aims to take a census of all the traffic going from one zone to another, considering all the possible paths. These matrices are thus far more efficient to get a comprehensive overview of the travel dynamics but are more difficult to gather and usually result of models. Zone-to-zone matrices are harder to assign as the traffic needs first to be distributed over the possible starting points and ending points for each zone.

As shown above, the definition of the OD matrix is important as it can lead to very different meanings and results. A confusion between this two kind of matrices might happen as several ways exist to go from one type to the other. Zone-to-zone matrices can be transformed in point-to-point OD matrices by considering, for example, that the centroid of each zone is representative of the zone. To the opposite, point-to-point matrices can be interpreted as zone-to-zone OD matrices: The points are associated to the zone they are in, based either on a *a priori* zoning or on a *a posteriori* zoning (e.g. by doing the Voronoi partition of the set of points). Although useful for comparison purposes and

results analysis, such transformation will not mitigate the drawbacks of the initial OD matrix and therefore, it is highly important to be aware of the characteristics of the OD matrix one is handling.

1.2.2 Dynamic OD Matrices

On top of the above considerations, adding time dependencies can lead to several definition of dynamic OD matrix. In the literature, researches usually differentiate the pseudo-static approaches from online estimation. The pseudo-static approach consists in estimating successive static OD matrices (but defined over a small time interval) possibly with dependencies in between successive estimation. Those approaches are adapted when past traffic evolution is the matter of interest. For online estimation, the OD matrix of interest is the one at present time. The online estimation relies on the estimation at previous time and on inferences from past measures.

In both cases, once the time interval is of the order of magnitude of the duration of a trip, there are several ways to handle the boundaries:

- First, one can consider trips starting within the time interval only. Then the matrices embody the changes of the network. This definition of the dynamic OD matrix is widely accepted. However, the trips longer than the time intervals are not necessarily accounted for. This makes this matrix not adapted for network condition analysis as one would need the assignment of all the matrices.
- Thus, another way to handle the boundaries is to consider all the trips ongoing at the time of estimation. It could be seen as a picture of the network. This description is convenient to describe the network state at a given time step but is problematic for short time period analysis as trip longer than the time interval will appear in several successive matrices.
- Sometimes, dynamic OD matrices will split each user movements into successive trips in between adjacent zones. Thus, ideally, the longer trip will appears in successive OD matrices in different OD pairs. However, handling the boundaries is still as problematic as with the other definition. Moreover, as those matrices do not really take the census from origins to destinations (as trips are split), we prefer to denote them under the term of *link matrices*.

Unless otherwise specified, we will use in this thesis the first definition of a dynamic matrix and we will focus on point-to-point matrices, adapted to our Bluetooth detectors network.

1.2.3 OD Matrix Estimation: Principles

Once extensive surveys are ruled out, automated field data collection becomes the critical element in the quest of estimating OD matrices. This has been a popular topic since the seventies [30] as a consequence of the generalisation, in occidental cities, of the access to link counts (mostly by pneumatic tube and magnetic/inductive loops). This problem of estimating OD matrices from traffic counts is still today an important area of research due to, first, that the system is ill-posed, hence, having an infinite number of solutions and, second, that assessing the efficiency of the estimation requires a inaccessible ground truth, that is, the ability to actually observe the OD matrix.

To give a sense of why this is an ill-posed problem, let us consider a N_V nodes network. Then the number of possible origin (resp. destination) is proportional to N_V (with a factor smaller than one if nodes are grouped within zones). Thus the number of possible OD pairs is proportional to N_V^2. On the other hand, for each intersection, one can expect to have in average 6 to 8 roads connected (e.g., 3-4

in each direction) and thus the number of roads, which correspond to the maximum number of traffic counts available, would be around $3N_V$ to $4N_V$. The traditional problem is therefore about estimating N_V^2 ODs from $\sim 4N_V$ measures.

The Four-Step Model

OD matrices estimation is traditionally interpreted through the four-step model [31], well known from transport engineers. This model originally aimed to create traffic flows on networks from socio-economic data of the geographic zoning. It relies on the four following steps:

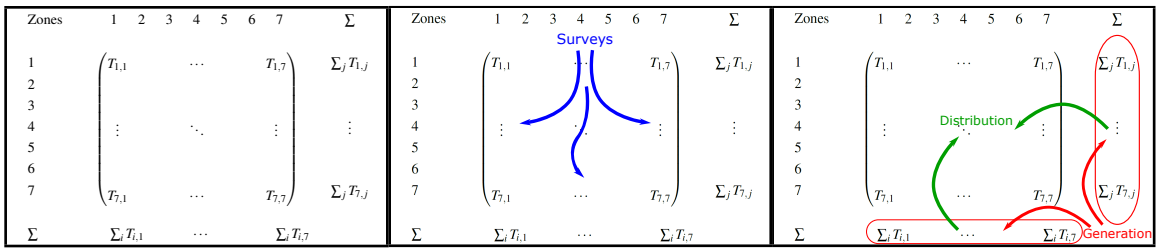
1. **The Trip Generation** aims to associate, to each zone, a power of attraction (or a potential of being a destination) and power of production (or a potential of being an origin). This step aims therefore to estimate how many users will leave the zone and how many aims to reach it. A simple example is to choose a power of production proportional to the population living inside the zone and a power of attraction proportional to employment (and/or leisure, shops,...).

2. **The Trip Distribution** aims to calculate an OD matrix $\underline{\underline{T}}$ consistent with the criteria generated at the previous step. For example, one of the most famous model is the gravity model (see Chapter 2, Section 1.3.1).

3. **The Modal Split** aims to divide the user of the network between the varied transport modes available. If this step is of interest for multi modal networks, for this research, we will focus on cars only and this step will not be more developed.

4. **The Trip Assignment** aims to assign a trajectory to each user, based on the origin and destination inferred in step 2 and according to the mode (steps 3). Thus a decision process is modelled, depending on varied criteria as length of the path, cost, traffic density, travel time, etc...

The reader can refer to [32], [33] for more information on this framework. In the light of this model, the OD matrix estimation process can be illustrated as in Figure 2.2.



(a) (b) (c)

Figure 2.2: (a) OD matrix (b) Surveys sample some of the element of the matrix and the challenge is to generalise (c) Model estimates how many users leave and aims each zone based on parameters as population density, office area, commercial surface area,... After this generation process, a second steps aims to distribute the users amongst OD pair consistently with the generation process. The challenge is thus the calibration of the model.

If surveys are used, they directly sample elements of the OD matrix. The OD matrix then needs to be generalised. This replaces the first two steps of the four-step model.

If in theory the OD matrix is estimated either after the generalisation process for surveys or at the distribution step of the four-step model, the question of its reliability is raised as OD information to which to compare the estimation is hard to obtain. Thus, traditionally, the matrix is evaluated by comparing the results of the assignment step (step 4) with observed traffic counts (measured by loop detectors giving the volume of vehicles for each link of the network).

Given that surveys or steps 1 and 2 of the four-step model give an estimation $\underline{\underline{\widehat{T}}}$ of the true OD matrix $\underline{\underline{T}}^{\star}$, then steps 4 give the proportion of the traffic T_{ij}, originating in zone i and having for destination j that goes through each link l, denoted Q^l_{ij}:

$$(\forall(i,j,l) \in V \times V \times L) \quad \widehat{Q}^l_{ij} = p^l_{ij}(\widehat{T}_{ij}) \tag{2.1}$$

where $p^l_{i,j}$ can, in simpler models be the proportion of the traffic T_{ij} going through link l. In more complex studies, it can be a function of many parameters for more realistic traffic modelling.

Then, an estimated volume of the traffic passing through each link, $\underline{\widehat{q}}$, is given by the sum over all origins and destinations:

$$(\forall l \in L) \sum_{ij} p^l_{ij} \cdot \widehat{T}_{ij} = \widehat{q}^l \tag{2.2}$$

The main challenge, on which most researches focused on, is to minimise the difference between estimated traffic counts ($\underline{\widehat{q}}$) and observed traffic counts ($\underline{\tilde{q}}$), for example by minimising the mean square error function $\|(\underline{\tilde{q}} - \underline{\widehat{q}})\|^2$. In practice, this is done through a process iterating over the four steps of the model, adjusting the parameters, until $\|(\underline{\tilde{q}} - \underline{\widehat{q}})\|^2$ becomes small enough.

Statistical Approach

Given that the goal is to minimise the difference between the estimated and the measured traffic counts, the OD matrix estimation problem can directly be formulated as an inverse problem expressed as:

$$\begin{aligned} (\underline{\underline{\widehat{T}}}, \underline{\widehat{q}}) \in \operatorname{Argmin}_{\underline{\underline{T}},\underline{q}} & \left\{ \mathscr{D}(\underline{\tilde{q}}, \underline{q}) \right\} \\ \text{s.t.} \quad & \underline{q} = F(\underline{\underline{T}}) \\ & \underline{\underline{T}} = M(\underline{\underline{\tilde{T}}}) \end{aligned} \tag{2.3}$$

where $\mathscr{D}$ is a distance function, F is the function corresponding to the assignment and aggregation of flows and $M(\underline{\underline{\tilde{T}}})$ is a model on $\underline{\underline{T}}$ from a priori information $\underline{\underline{\tilde{T}}}$ (gravity model, survey, prior OD knowledge).

To solve Problem (2.3), there are three main levers: adjusting $M(\underline{\underline{\tilde{T}}})$, F and $\mathscr{D}$. Adjusting $M(\underline{\underline{\tilde{T}}})$, for models, corresponds to different generation and distribution models. Similarly, for surveys, it consists in modifying the generalisation step. Adjusting F is to focus on the assignment method. Thus, the cost function of each path can be changed from a simple *all or nothing* assignment, considering only one possible path between each OD pair to more complex assignment considering, for example, the Wardrop's equilibrium of the networks. Last, the choice of $\mathscr{D}$, ensuring consistency between the estimates and the measures will also impact the results.

These steps are summarised with Figure 2.3.

In any case, this minimisation problem is highly under-determined (as by acting on these many levers infinity of solutions can be found). Thus the reliability of the results is questionable. Also and

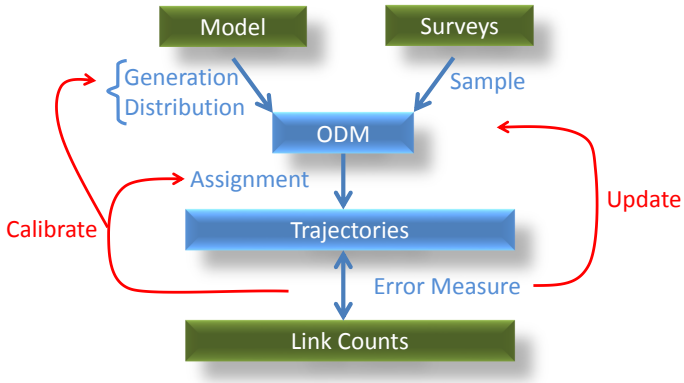


Figure 2.3: Traditional OD matrix estimation process. In green, external inputs (observed data or assumptions for modelling), in blue rectangles estimated variables, and with blue arrows the processes names. Red arrows represent possible use of inverse problem (feedback loops) so that the process ultimately fit observed link counts.

the methods developed are not always transferable from the network they were defined on to another because of the modelling involved in adjusting M and F.

Problem (2.3) can also be rewritten as the following generic problem where the condition $\underline{\underline{T}} = M(\underline{\underline{\tilde{T}}})$ has been relaxed:

$$\begin{aligned} (\underline{\underline{\widehat{T}}}, \underline{\widehat{q}}) \in \underset{\underline{\underline{T}}, \underline{q}}{\text{Argmin}} \; & \left\{ \gamma_1 \mathscr{D}_1(\underline{\underline{\tilde{T}}}, \underline{\underline{T}}) + \gamma_2 \mathscr{D}_2(\underline{\tilde{q}}, \underline{q}) \right\} \\ \text{s.t.} \quad & \underline{q} = F(\underline{\underline{T}}) \end{aligned} \tag{2.4}$$

where $\mathscr{D}_1$ and $\mathscr{D}_2$ are two distance functions and γ_1, γ_2 two weights representative of the relative belief in $\underline{\underline{\tilde{T}}}$ and in $\underline{\tilde{q}}$ respectively.

Similarly to Problem (2.3), three levers impact the estimates: $\mathscr{D}_1$, $\mathscr{D}_2$ and F.

1.3 OD Matrix Estimation: Review of the literature

Whether one aims to solve Problem (2.3) or its relaxed expression Problem (2.4) for solving the OD matrix estimation, the many levers are source for a large literature with variations on both the distribution, the distance functions and the assignment.

1.3.1 The Gravity Model

The gravity model can be traced back at least to the late fifties [29], [34] were the economic trade between cities were modelled by a function proportional to the population of both cities and inversely proportional to the square of their distance. Low (1972) [35] first developed a model of traffic demand based on the gravity model [31], [35]. It assumes that the traffic between origin i and destination j is proportional to the population of zone i and to the employment in zone j but inversely proportional to the distance (raised to the β-th power):

$$T_{ij} = \alpha \frac{Pop_i \cdot Emp_j}{d_{ij}^{\beta}}, \tag{2.5}$$

where Pop_i is the population of zone i, Emp_j the employment in zone j, and d_{ij} the distance between zone i and j.

Several models were then derived from this first one ([31]):

- Using surveys to perform the assignment for a fraction of the volumes (*external volumes*) and calibrating the model on the *internal volumes* (counted flows less *external volumes*) [35].
- By complicating the power of attraction and emission equations (including percentage of population in *one family houses* [36]).
- Calibrating the gravity model using the traffic counts with a least squares method [30].
- Testing several socio-economic indicators [37].
- Calibrating different functions for the trip generation depending on trip purposes [38].

The interested reader could find a deeper survey on the use of the gravity model in transport in [39]. The gravity model has been involved with Bayesian modelling and inference in [40], without involving traffic counts and is still used today for OD matrices estimations (e.g., in [41]), yet more from an historic point of view rather than for further development of the model.

Major concern regarding the gravity model is the *a priori* assumption of a trip distribution function. This assumption makes the model easier to handle by using the formula of gravity laws whose apparatus are well understood but is also its own limitation as no clear evidence exists regarding its applicability to any travel pattern.

These works, by trying to fit the parameters of a model to the observed traffic counts are also referred as *parameters calibration techniques* [42] by opposition to statistical approaches as presented below.

1.3.2 Statistical Approaches

Thus, another approach to solve Problem (2.4), referred as *matrix estimation methods* [42] is to focus on the adjustment of a matrix based on statistical analysis. For these methods, the aim is not to find the right values for the parameters of a model anymore, but to solve an inverse problem, based on the assumption of a prior knowledge of the OD matrix ($\underline{\underline{\tilde{T}}}$ in Problem (2.4)) and/or of the distribution of the elements in $\underline{\underline{T}}$. An overview of the major works identified in the literature is presented in the following.

Entropy Maximisation

As OD estimation problem has many characteristics in common with some Statistical Physics problems (large numbers of components, apparent disorganised complexity), it was realised that using an entropy maximisation (or information minimisation) analogy was suited to constrain the distribution model (e.g., the gravity model [31], [43]).

The idea behind entropy maximisation, as presented in [43], [44], and updated in [45] was to shift from a Newton-like description of the distribution (gravity model) to Boltzman description of physics systems composed of a high number of particles interacting weakly. This analogy is so that

each OD pair is a *box* and a state of the system is when every individual is assigned to a box. With this description, the number of possible states for a given matrix $\underline{\underline{T}}$, W is:

$$W = \frac{(\sum_{ij} T_{ij})!}{\prod_{ij} T_{ij}!} \tag{2.6}$$

where $\sum_{ij} T_{ij}$ is the total population.

Then, for this given matrix $\underline{\underline{T}}$, its probability to occur can be related to the number of possible states that stem from the matrix. Thus, maximising the "entropy" W will give the most probable state of the system.

Maximising Equation (2.6) is equivalent, to maximise the logarithm of W and thus can be expressed under the generic formulation of Problem (2.4) with

$$\mathscr{D}_1(\underline{\underline{T}}, \underline{\underline{\tilde{T}}}) = \sum_{ij} T_{ij}(\log T_{ij} - 1) \tag{2.7}$$

In [43], [44], $\mathscr{D}_2$ is not specified and the problem is still ill-posed, so extra constraints are added:

$$\sum_j T_{ij} = O_i, \tag{2.8}$$

$$\sum_i T_{ij} = D_j, \tag{2.9}$$

$$\sum_{ij} T_{ij} \cdot c_{ij} = C, \tag{2.10}$$

where

$$\begin{cases} O_i \text{ is the number of individuals having } i \text{ for origin,} \\ D_j \text{ is the number of individuals having } j \text{ for destination,} \\ c_{ij} \text{ is the cost from } i \text{ to } j \text{ (distance, time, monetary etc...),} \\ C \text{ is the total transport expenditure (the energy of the system).} \end{cases}$$

Interestingly, the solution of this maximisation problem can be written as follow:

$$T_{i,j} = A_i \cdot B_j \cdot O_i \cdot D_j \cdot e^{-\beta c_{i,j}} \tag{2.11}$$

which is very similar to the gravity model where $c_{i,j}$ would have been replaced by $\log(c_{i,j})$, which would tend to mitigate the weight of longer trips.

In [46], the authors involve link counts by solving

$$\begin{aligned} \underline{\underline{\widehat{T}}} &\in \underset{\underline{\underline{T}}}{\text{Argmin}} \left\{ \mathscr{D}_1(\underline{\underline{\tilde{T}}}, \underline{\underline{T}}) \right\} \\ \text{s.t.} \quad \underline{\tilde{q}} &= F(\underline{\underline{T}}) \end{aligned} \tag{2.12}$$

or, using the Lagrangean form:

$$\underline{\underline{\widehat{T}}} \in \underset{\underline{\underline{T}}}{\text{Argmin}} \left\{ \mathscr{D}_1(\underline{\underline{\tilde{T}}}, \underline{\underline{T}}) + \lambda \sum \left(\underline{\tilde{q}} - F(\underline{\underline{T}}) \right) \right\}, \tag{2.13}$$

with F a proportional assignment (*cf.* Section 1.3.3).

Fisk (1988) [47] combined the entropy maximisation with a more complex assignment procedure (proportional assignment with proportions varying with link congestion), and constrained with observed link counts.

Similarly, in [48], the author proposes to take the *information minimisation* paradigm with Brillouin's measure:

$$\mathscr{I} = -\log \frac{\prod_{ij} \left(\frac{\tilde{T}_{ij}}{\sum_{ij} \tilde{T}_{ij}} \right)^{T_{ij}}}{\prod_{ij} T_{ij}!} \tag{2.14}$$

which itself can be interpreted as Problem (2.4) with

$$\mathscr{D}_1(\underline{\underline{T}}, \underline{\underline{\tilde{T}}}) = \sum_{ij} T_{ij} (\log \frac{T_{ij}}{\tilde{T}_{ij}} - 1) \tag{2.15}$$

Interestingly, this model is very similar to that of *Entropy Maximisation* where values in $\underline{\underline{T}}$ are weighted according to the expected flow in $\underline{\underline{\tilde{T}}}$.

Lam and Lo (1991) [49] presented a comparison between Entropy maximisation models and Information minimisation, highlighting the importance the the prior OD within the estimation process.

Maximum Likelihood

The maximum likelihood estimation process is adapted when the prior OD matrix is assumed to be a sampling of the one to be estimated. If one assumes that the OD matrix estimates are random variables with some *a priori* distribution, the idea is to maximise the probability of the estimates to occur. The *a priori* distribution can be a Gaussian, similarly to what has been done in [50] for estimating flows (but not origin-destination volumes) or in [51]. However, a much more popular assumption is the one of the Poisson distribution (e.g., [52], [53]). In this case, if $\underline{\underline{\eta}}$ is the sampling factor matrix for each OD pair we have:

$$Prob[Poisson(\eta_{ij} T_{ij}) = \tilde{T}_{ij}] = \frac{(\eta_{ij} T_{ij})^{\tilde{T}_{ij}} e^{-\eta_{ij} T_{ij}}}{\tilde{T}_{ij}!} \tag{2.16}$$

Therefore, the joint probability of observing the sample matrix $\underline{\underline{\tilde{T}}}$ is

$$Prob[\underline{\underline{\tilde{T}}}] = \prod_{ij} \frac{(\eta_{ij} T_{ij})^{\tilde{T}_{ij}} e^{-\eta_{ij} T_{ij}}}{\tilde{T}_{ij}!} \tag{2.17}$$

Then, by applying the maximum likelihood estimation technique, this problem correspond to Problem (2.4) with:

$$\mathscr{D}_1(\underline{\underline{T}}, \underline{\underline{\tilde{T}}}) = \sum_{ij} \left(\eta_{ij} T_{ij} - \tilde{T}_{ij} \ \log T_{ij} \right) \tag{2.18}$$

Based on the maximum likelihood method, the use of an expectation maximisation (EM) algorithm to estimate the OD has been proposed in [54] (*cf.* Appendix B for more information on the EM algorithm). In [55] the authors used this maximum likelihood combined with an assignment based on random link choice proportions.

A variant to the maximum likelihood method is the Bayesian inference, where, similarly, a distribution linking the estimates and the observed variables is assumed with the variance representing the confidence in the prior belief. For more details, the reader can refer to [55]–[60]

Least Squares

The least squares method refers to cases when one of the two functions, $\mathscr{D}_1$ or $\mathscr{D}_2$, or both, are similar to:

$$\mathscr{D}_k(\tilde{x},x) = \frac{1}{2}\|\tilde{x}-x\|^2. \tag{2.19}$$

This expression can be reformulated in order to take into account variances of observations or even covariances, referred as *Generalised Least Squares* or Aitken estimator (*cf.* [61]–[67] and reference therein).

The problem is then the one of inverting the following equation:

$$\begin{bmatrix} \underline{\underline{\tilde{T}}} \\ \tilde{q} \end{bmatrix} = \begin{bmatrix} \mathbb{1} \\ A \end{bmatrix} \cdot \underline{\underline{T}} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{2.20}$$

where A is the assignment matrix, ε_1 (resp. ε_2) is a random vector with dispersion matrix Σ_1 (resp. Σ_2). Then Problem (2.4) formulated with the Generalised Least Squares can be expressed as follows:

$$\widehat{\underline{\underline{T}}} = \underset{\underline{\underline{T}}}{\text{Argmin}} \begin{bmatrix} \underline{\underline{\tilde{T}}} - \underline{\underline{T}} \\ \tilde{q} - A\underline{\underline{T}} \end{bmatrix} \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2^{-1} \end{bmatrix} \begin{bmatrix} \underline{\underline{\tilde{T}}} - \underline{\underline{T}} \\ \tilde{q} - A\underline{\underline{T}} \end{bmatrix} \tag{2.21}$$

1.3.3 The Assignment

For both Problem (2.3) and (2.4), the function that have not been much discussed yet is the assignment function F. We will however only give a quick overview of the various existing assignment procedure, keeping in mind that our objective is the estimation of the LOD matrix that does not require this assignment step.

One of the assumption in the oldest works is the *proportional assignment* assumption. That is that the demand ($\underline{\underline{T}}$) is proportional to the link flows ($\underline{q}$). This kind of assignment is used in [55], [56], [63], [68] amongst others. This assignment has the strong advantage of being simple, however it is obviously not adequate in congested situation. Smock (1962) [69] proposed an iterative algorithm, based on a cost-flow relationship ($C_l(q_l)$), to converge toward a Wardrop's equilibrium [31]. This model was improved by Holm, Jensen, Nielsen, Christensen, Johnsen, and Ronby (1976) [70] by also calibrating a gravity like distribution model at the end of each iteration. Iterative solution for assignment proportion computation was presented by Yousefikia, Mamdoohi, and Noruzoliaee (2016) [71].

Yang, Sasaki, Iida, and Asakura (1992) [72] implemented the *user equilibrium* as in the work of Cascetta [62] in a bi-level program where the *lower level problem* is a deterministic user equilibrium assignment and the *upper level problem* the estimation of the trip table (for example based on the Generalised Least Squares estimation). For more details on User Equilibrium approaches, the interested reader can refer to [73]. The reader can also refer to [74] for more details on bi-level programming. Sherali, Sivanandan, and Hobeika (1994) [42] introduced a linear programming model estimating the flow for different paths instead of the usual *link* approach in order solve the deterministic user equilibrium assignment. Variants of the deterministic user equilibrium involve stochastic user equilibrium as in [75], [76]. Codina and Barceló (2004) [66] use convex optimisation to solve the elastic demand traffic assignment, using the notion of subgradient for cases where $\gamma_1\mathscr{D}_1(\underline{\underline{T}},\underline{\underline{\tilde{T}}}) + \gamma_2\mathscr{D}_2(\underline{\tilde{q}},\underline{q})$ would not be differentiable.

To directly account for the effects of congestion, *Combined Distribution and Assignment* (CDA) models were developed with the aim of estimating the trip tables through a single objective function in which congestion is considered. To the opposite of user equilibrium where the assignment is implicitly defined, (going from OD flows to links flows without intermediate trajectories), CDA, as in [77], uses an assignment function.

Fisk and Boyce (1983) [78] proposed a CDA based on the work of Erlander, Nguyen, and Stewart (1979) [77] but including count data in the process of model calibration. Then, the same authors demonstrated that, for congested networks, the network equilibrium approach and the CDA approach give very similar results when the traffic counts correspond to an user equilibrium pattern [79] .

1.3.4 Dynamic OD Matrices Estimation

The dynamic OD matrix estimation can be seen as a problem very similar to Problem (2.4) but with an extra dimension involved (time). The OD matrix becomes thus a 3-dimensional matrix where timesteps are indexed over the third dimension and, accordingly, where the measures are aggregated for each of those timesteps. The assignment F can be time dependent or not (e.g., [80] ([33])).

Additional time-based constraints can then be added, for example, to enforce some correlation in time of the successive estimates [81], to constraint their deviation, etc.

Cascetta and Marquis (1993) [82] proposed a time dependent proportional assignment associated with an assignment probability: a function that describes for each user, for each timestep, the probability of the user to choose a given path. Several works were built upon this idea [64], [83], [84].

Similarly, in [85], [86], Hazelton proposes that the variations of proportional assignment are linked to the OD matrix by a Poisson distribution.

The deterministic user equilibrium models have also been extended to involve time dependencies as in [87], [88].

The dynamic estimation problem has also been treated for in-real time or online applications, that is, for cases where observations are available up to the present time only. The Kalman filtering theory has been used in the early seventies ([89] but only for one link, using data of adjacent links) and then further developped as in [90]–[93]. Iterative algorithms have also been used for incomlete observations on link counts [94].

More recently, some works had a stronger focus on congested situations [88], [95]. Frederix, Viti, and Tampère (2011) [95] proposed a method based on hierarchical decomposition, that is, that uses diverse procedures for estimating OD matrix on sub-networks, clustered according to their characteristics (importance, capacity, usage,...).

More references on this approach of the problem can be found in [33], [96] and references therein.

1.3.5 The Role of the New Technologies

The dream of being able to identify vehicles for traffic studies goes back to the seventies [97]. It was discussed of the opportunities offered by partial licence plate numbers recordings on freeway. In the late nineties, Van Der Zijpp (1997) [98] discussed the potential of AVI (Automated Vehicle Identification) systems for OD matrix estimation. His work was based on registration plate identification but did not have a practical case study. Very similar methods have been further developped in [99]–[101].

In the later one, the method is applied to a real case study: a small linear section of the Han-Shin expressway.

Among the literature, contributions based on hypothetical AVI systems can be distinguished from others.

Dixon and Rilett (2002) [102] developed and tested constrained estimators based on generalised least squares and Kalman filtering to estimate dynamic OD matrix from AVI data (partial trajectory information through hypothetical tagging), traffic counts and a prior knowledge of origin volumes. The authors demonstrated that including AVI does improve the estimation of OD matrix and that the Kalman filtering approach was the most effective method compared to that of the generalised least squares estimator.

Zhou and Mahmassani (2006) [103] proposed a nonlinear ordinary least-squares model combining AVI counts with other available information into a multiobjective optimisation framework. This work highlights the importance of market penetration to obtain reliable information from AVI counts. The method was tested on a relatively simple network of 31 nodes and composed of freeway and arterial links only.

A method to retrieve static OD matrix from traffic counts and discrete trajectories, based on Maximum Likelihood estimation method, assuming a Poisson Distribution for traffic flows and using the routing data as a target matrix ($\underline{\tilde{T}}$) is discussed in [68]. The studied network has 21 nodes and the time windows is of 4 hours. The author recognised that their model would not be resilient to large variations of the AVI penetration rate.

More theoretical works have been realised to formalise the relationships between link volumes, flows per path and OD matrix under the assumption that the trajectories do not involve cycles by Teknomo and Fernandez (2014) [104], [105]. The authors did not consider however the problem where only a set of trajectories is known.

Feng, Sun, and Chen (2015) [106] proposed a method based on the particle filter to recover trajectories from AVI data. In this model, five factors are taken into account: the path consistency factor (AVI nodes belong to the trajectory and in the same order), the travel time consistency factor (comparison between estimated travel time on a trajectory and travel time computed from AVI detections), the measurability criterion factor (to account for error of detection), the gravity flow model factor (trajectories are expected to take links between adjacent zones that have a high volume of traffic), and the path-link flow matching factor (to check the consistency between trajectories and flows). The case study is simulated, based on the Olympic Park traffic network in Beijing. This network contains 127 nodes, 151 links, and 42 traffic zones. The simulation model was used to generate and export 100 link traffic volumes and travel times for a comparison with field measurements. Static OD were then estimated from this trajectory reconstruction. The authors identified that 50% of AVI coverage is a threshold below which the error of the static OD rises dramatically.

The technologies that have been used within OD estimation frameworks are mostly:

Registration Plate: Path-flow estimation involoving identification of the users at several links of the network has been studied by Castillo, Menéndez, and Jiménez in [59], [107] with the aim of recovering the trip table from partial plate scanning system. In this case however, the whole set of users is observed but only for a small fraction of the network. Using Bayesian network and Wardrop minimum variation model, the authors applied their model to estimate the OD matrix on the Nguyen–Dupuis network (18 OD-pairs).

Electronic Tagging: Kwon and Varaiya (2005) [108] used the method of moments to estimate unbiased OD matrix from sample trajectories collected through an electronic tagging system. OD-pairs

consisted in entry and exit points on a freeway. The method can however only be applied to graphs where each OD pair is exactly uniquely *traversable* (there is one and only one path). Tiratanapakhom Tawin (2013) [109], and, Kim, Kurauchi, Uno, Hagihara, and Daito (2014) [110] used the same AVI system to respectively measure and help to understand variability in individual travel behaviour and to retrieve OD information on freeway.

GPS: A method to estimate time-dependant OD matrix from both FCD (floating car data) systems (in this case, Taxi equipped with communicating GPS) and remote traffic microwave sensors (RTMS, road-side radar providing with speed, occupation, density of users nearby), using time-varying splitting rates (proportion of the total demands for a particular OD pair for the entire modelling period that will fall into a particular time interval) has been presented in [111]. The method is applied to a freeway (part of the Western 3rd Ring-Road in Beijing, China). Opportunbities of GPS enabled phone for traffic data are further discussed in [112]. GPS has also been used for calibrating route choice models (hence the assignement) in [113]. GPS datasets, however are generally limited in size and managed by private operators. Moreover, data also need to be pre-processed for correction (especially in urban environment) and appropriately map matched [114].

Mobile Phone: Mobile phone data for traffic indicators estimation have been used since 2000 [115], [116], mostly for travel time estimation. The potential to estimate OD matrices using mobile phone Call Detail Records (CDR) has been discussed by in early 2003 in [117] and led to many research: [118]–[124]. CDR has been combined with traffic counts data to estimate OD matrix: e.g., in Dhaka, Bangladesh [125] with 67 nodes and 215 links covering an area of about 300 km2, a population of about 10.7 million, and with 13 traffic counts detectors (video) for one month of data. Mobile phone data have also been used at a city scale in Paris [124] but without any attempt to generalise the matrix to the whole set of users.

Bluetooth: Researches have been conducted into Bluetooth-based data collection for improving the estimation of OD matrices. Precursory works focused on free-ways or corridors as in [126] (Kalman filtering based method applied to a freeway with 11 entries and 12 exits), [127] (two cases studies in Brisbane: one with two OD pairs and one with 29 detectors), [128] (18 detectors - The study I-526 corridor is located in Charleston County, South Carolina) and [129] (based on a single case study in Jacksonville with 14 detection devices spread along one corridor). Some works focused as well on urban context: [130] (with 48 detectors) and [131] (a case study in Ankara for an open system composed of 10 intersection and 4 major roads equipped with 4 Bluetooth devices), but for relatively small networks. Finally, Bluetooth had also been used for public transport studies as in [132] and [133].

WiFi: The WiFi technology is very similar to the Bluetooth technology (MAC address scanners) and have also been used for traffic information inferences [126]. Yet, for vehicle traffic analysis, it has been proven to provide much smaller datasets [134], [135] compared to Bluetooth.

Last, many researches focused on developing methods for building optimal detectors networks (for more reference: [136]–[139]).

1.4 Research Gaps Identified on OD Matrix Estimation

In the light of these elements, the identified research gaps on OD matrix estimation addressed here are:

- Urban networks: Many works focused on method adapted for the study of traffic on freeway. Urban networks are composed of nodes with higher connectivity and several paths can connect

the same OD pair (to the contrary to freeways). Thus, estimating OD matrix in urban context is much more challenging.

- Large Network: Among the works that focused on urban networks, the networks are usually small (few tens of nodes). Brisbane Bluetooth detectors are over 580 and the number is constantly increasing. Thus a network based on this set of detectors can already be considered as a large network compared to the literature.

- Real data case study: Many researches have been applied to simulated cases only. Here, Bluetooth data and traffic counts in Brisbane are available for a real case study.

- Complex objective function: While the use of both traffic counts with AVI information has already been explored (e.g., [125]) the objective functions used to fit the OD matrix with the measured data remain simple (error with traffic counts minimisation). Here, we will propose to extend the objective function, and involve more complex properties of the problem, using tools from graph theory and signal processing (Total variation, link-path relationship, optimisation algorithm ...).

- Non stationary expansion factors: Most of the previous researches assume that the penetration factors (inverse of the penetration rate, to deduce total volume from AVI measurements), do not vary along time or space (making thus the underlying assumption of a perfect random distribution of AVI system in the population). This assumption has not been supported by precise studies and might not be verify in the general case (e.g., the Bluetooth penetration rate could depend on wealth). Therefore, the method proposed here permit for variations of the penetration rate.

- Trajectory based estimation of the OD matrix: User trajectories have never been used as the primary source of traffic information in the estimation process. Yet probe trajectories are more and more available thanks to new technologies. Therefore we propose here to base the estimation process on probe trajectories and thus to extend the concept of OD matrix to the one of link dependent OD matrix (LOD matrix). The LOD matrix is similar to the intermediate variable $\underline{\underline{\underline{Q}}}$ of the Four-Step model presented in Equation (2.1). It is directly sampled by probe trajectories and its comparison with traffic counts is straightforward, making its direct estimation possible, that is, without the need of an assignment step. Figure 2.4 provides an illustration for a small road network and the corresponding traffic variables $\underline{q}$, $\underline{\underline{T}}$ and $\underline{\underline{\underline{Q}}}$.

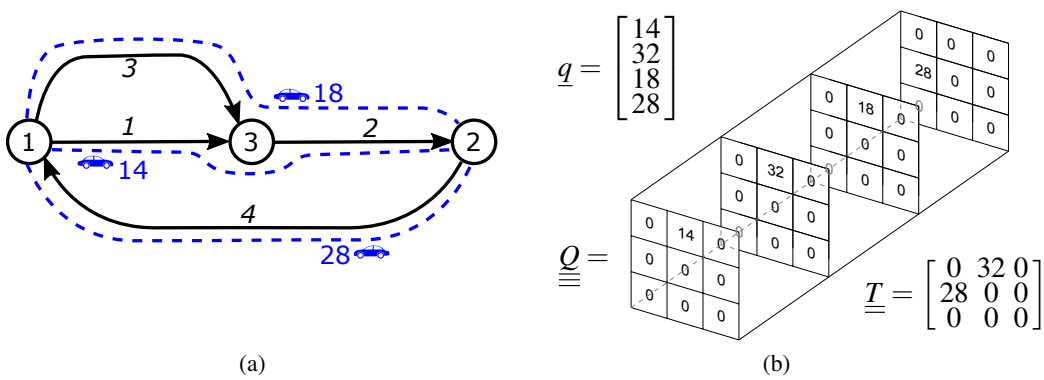
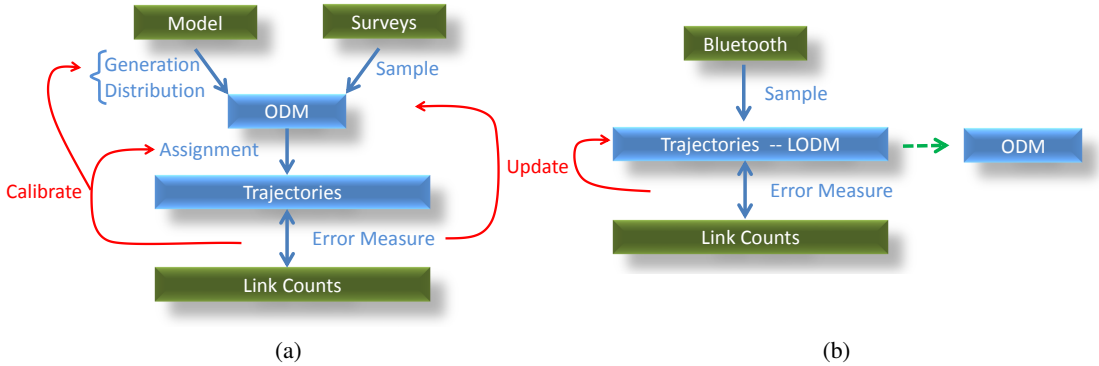


Figure 2.4: Example of a simple network (a) with the associated tools describing the traffic (b).



(a) (b)

Figure 2.5: Traditional and proposed framework for LOD matrix estimation. The new framework is more constrained: only one step is the results of a model or an inverse problem and only one possible feedback exists.

Therefore, we propose here a new estimation framework as illustrated in Figure 2.5b. The estimation of LOD matrix, as it is developed in this work, relies on probe trajectories. Most AVI systems, however, provide point-measures from which trajectories need to be inferred. For example, in Brisbane, the road users are only detected in the neighbourhood of the Bluetooth detectors. Consequently, an additional contribution of this work is to understand how the Bluetooth technology can be used as a complementary dataset for traffic related problems, with a particular focus on the problem of retrieving trajectories from point detectors scattered over a network.

2 Bluetooth as a New Source of Data for Trajectories

The potential of Bluetooth for data collection in a automotive environment has been explored in early 2000: Nusser and Pelz (2000) [140] proposed the use of Bluetooth for in-car services. The proof of concept by Murphy, Welsh, and Frantz (2002) [141] verified that Bluetooth devices could be discovered in moving vehicles (see also [142]) and the Bluetooth technology also started to be used for tracking individuals [143]. Around 2010, the Bluetooth technology has been the matter of many researches for ITS (Intelligent Transport Systems) applications, mostly for travel time estimation [126], [127], [137], [144]–[156].

The trade-off reached by this technology is that it has less accuracy than, e.g., GPS, but it achieves a much higher sampling rate amongst users (around 25% in Brisbane). Further, it is much cheaper than video or plate scanning systems and can therefore be more extensively deployed, covering larger portions of the network. Compared to loop detectors, it enables the individual tracking of vehicles, yet for a subset of the users only. For those reasons, Bluetooth technology has already proven to be of interest for many transport applications: travel time estimation and OD matrix estimation, the estimation of the congestion at intersections [157], analysis of behavioural patterns of visitor at major mass events [158], or traffic management [159]. A more comprehensive review of the uses of Bluetooth can be found in [160].

Several works also bring important information on Bluetooth characteristics and behaviour. They are useful in our quest to understand the limitation of this technology. Amongst other works, the most important ones are:

- The description of the detector inquiry cycle process is described in [161].

- The possibility to clone Bluetooth devices parameters for fleet's specific needs in presented in [162]. Consequently, the question of the MAC address uniqueness is raised in [163].
- Bluetooth tracking without discoverability has been investigated in [164].
- The influence of vertical sensor placement on data collection efficiency has been demonstrated in [137].
- Variables impacting the detection rate have been explored in [165]. It has been shown in particular that when the number of detectable devices increases, interference may affect the effectiveness of the detection.
- Moreover, the influence of the antenna on the signal strength and on detection probability has been highlighted in [166], confirming the assumption that not all scanners and devices are equally powerful and that some have stronger signals than others.
- More informations on Bluetooth in general can be found in https://www.bluetooth.com/, a website focused on Bluetooth technology.

Concerning the retrieval of trajectories, one must differentiate between the study of motorways and urban corridors, where trajectories are mostly defined by entry and exit points of the corridors, (the road segments used in-between being obvious), from the study of trajectories in urban context where, for one given origin-destination pair, several paths can coexist. On motorway vehicle trajectories have been estimated using loops [167], fusion of loops and Bluetooth [156]. In urban context, numerous works have focused on recovering a vehicle trajectory by discriminating it from an *a priori* set of feasible path. This discrimination has been done using plate scanning systems [107], Bluetooth technology [149] by matching each possible path with a set of detectors (inclusion set) and a set of detectors outside the path (exclusion set), or with the reconstruction of the route Origin-Destination matrix (origin, destination and the detectors in between) [129]. More complex models have been proposed for the generic case of automated vehicle identification systems. In [106], a particle filter model based on five layers is developed but not applied to any technology in particular.

3 Conclusions

The present work aims to advance the general question of how new technologies can be involved in the frameworks designed to estimate, represent and give a better understanding of the traffic. In particular, we raise the question of retrieving and using probe trajectories for the travel demand estimation, by means of an extended version of the OD matrix, that inherently contains links to OD relationships. In addition, the context of this research is the availability of a unique dataset in Brisbane with more than 580 Bluetooth scanners. For our proposed framework to be realistic, we believed it had to be developed parallel to the analysis of the Brisbane Bluetooth dataset. This technology having been installed for traffic data collection quite recently (at least at such scale), and mostly for travel time estimation, assessing the adequacy of such data with other applications is one of the contributions of this work. Here, the Bluetooth data are used as a additional source for providing insights on the traffic, in particular, for the retrieval of probe trajectories.

These two questions, of assessing the Bluetooth data from such a huge network and of a method to recover Bluetooth trajectories are therefore two additional research gaps that have been identified and addressed in this PhD.

CHAPTER 3

Data

Contents

1 Road Network **47**
- 1.1 Nature - Principle 47
- 1.2 Network Interpretation 47
- 1.3 Network Simplification 48

2 Traffic Counts **49**
- 2.1 Nature - Principle 49
- 2.2 Formatting the Data 51

3 Bluetooth **51**
- 3.1 Nature of the Data 51
- 3.2 Characterising the Bluetooth Data 55
- 3.3 Missed Detections 63

4 Taxi Data **67**
- 4.1 Nature 67
- 4.2 Matching with Bluetooth Data 68

Build upon works published in:

- G. Michau, A. Nantes, and E. Chung, "Towards the retrieval of accurate OD matrices from Bluetooth data: Lessons learned from 2 years of data", in *36th Australasian Transport Research Forum (ATRF)*, QUT, Brisbane, Australia, Oct. 4, 2013. [Online]. Available: http://eprints.qut.edu.au/62727/
- **Section 3** in G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving dynamic origin-destination matrices from Bluetooth data", in *Transportation Research Board, 93rd Annual Meeting*, Washington DC, Jan. 12–16, 2014. [Online]. Available: http://eprints.qut.edu.au/66511/
- **Section 3** in G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving trip information from a discrete detectors network: The case of Brisbane Bluetooth detectors", in *CAITR*, Sydney, Feb. 17–18, 2014. [Online]. Available: http://eprints.qut.edu.au/83110/
- **Section 2** in G. Michau, A. Nantes, A. Bhaskar, E. Chung, P. Borgnat, and P. Abry, "Bluetooth data in urban context: Retrieving vehicles trajectories", *Submitted in IEEE Transaction on Intelligent Transport Systems*, 2016

ONE of the challenges of this work is to ensure that the processes and methods proposed here are not disconnected from the context and the reason of this work, that is, the real case study in Brisbane. Hence, a major part of this work is to understand the data at hand, their strengths, their biases and their limits, an essential step for gaining insight of what information one can hope to retrieve.

Therefore, the aim of this chapter is to gain this understanding of the datasets available and this will justify most of the assumptions made for more theoretical contributions in Chapter 4 and Chapter 5. In the following, the four main datasets used in this work will be presented, more or less briefly according to their criticality. First, the road network is crucial for defining the concept of trajectory, and thus, it is needed both for reconstructing trajectories, and for LOD matrix estimation. It is presented in Section 1. Second, the traffic counts dataset, that is combined with trajectories for LOD matrix estimation, needs to be located on the network and is presented in Section 2. Third, the Bluetooth dataset is presented in Section 3 and more deeply characterised. In fact, this technology is recent and there is not yet much expertise available on its use. Consequently, a clear understanding of this data is of utmost importance for further uses, e.g., here, the retrieval of user trajectories. Last, the taxi dataset, that is used for validating the trajectory recovery process, is presented in Section 4.

1 Road Network

1.1 Nature - Principle

Whether we are dealing with Bluetooth scanners, taxi data, traffic counts, and for many other datasets, most data are geo-referenced with longitude and latitude coordinates, independently of any road network information. Therefore, once more complex analysis of the road network usage and dynamic is sought (e.g., the concept of trajectory or the concept of shortest path), the road network has to be combined as an additional layer to the problem. A good format to represent a road network and to combine it with other datasets is the GIS (Geographic Information System) representation. Indeed, each road being represented by a geometrical object (most commonly a poly-line), the road network can easily be interpreted as a graph $\mathscr{G} = (V, L)$.

In this thesis, we decided to use the road network from OpenStreetMap (OSM)[1] for it is freely available and has a good reputation for reliability. The layer, once downloaded for the Brisbane area is composed of around 62 000 roads described with 432 000 coordinates (longitude, latitude) or 370 000 road segments. Each road is also characterised with a set of attributes describing its essential properties. In our case, the attributes of interest are: the type of road (motorway, primary road, secondary road, residential street, foot-way, cycleway, bus-way...), whether it is in a tunnel, on a bridge and what are the allowed direction of travel (one way or both ways). Within this representation, nodes are the extremities of the road segments and therefore they do not have other attribute than their coordinates.

1.2 Network Interpretation

In order to obtain a graph $\mathscr{G} = (V, L)$ out of the GIS representation, the above description, consisting in a collection of road segments, had to be converted into a collection of nodes V linked by edges L. This thesis has a particular focus on traffic inferred from observations and most of the observations are only gathered at major intersections (where Bluetooth scanners are installed) and on major roads

[1] https://www.openstreetmap.org/#map=15/-27.4728/153.0268

(where traffic counts are performed). Therefore, in a first step, the road segments are first filtered by their *type* attribute. The roads that common motorised vehicles can not take, like footway, cycleway, and service roads, amongst others, are unnecessary in the description of the network. Moreover, residential streets are also filtered on the ground that traffic information is not collected on those streets.

The resulting GIS layer, from which the graph is derived, is now made of 14 000 remaining roads (96 000 road segments, 110 000 coordinates). Of this collection of road segments, a list of unique coordinates is extracted and denoted as V, the set of nodes of the graph. Yet, due to precision errors on the latitude and longitude coordinates, a connectivity check is first needed. It is performed with threshold of one meter, that is: any two road segment extremities closer than one meter from each other are considered as identical and lead to only one node in the final graph. Once the set of unique nodes V is obtained, each road segment is described in L with, the index of the node it is leaving from, the index of the node it is going to, and its length. If the road segment attribute indicates a both ways road, another similar link with inverted nodes is added to the set of edges L. Some attribute information is kept at the node level: a flag indicates if the node is part of a road in a tunnel or on a bridge.

Such process applied to the Brisbane case study creates a graph with $|V| = N_V = 78\,000$ nodes and $|L| = N_L = 121\,000$ links. The graph is directed (to the opposite to the OSM GIS representation where the direction was an attribute). It explains why the final number of links is higher than the initial number of road segments.

1.3 Network Simplification

The size of the graph obtained in the previous section can be problematic for efficient computations (e.g., shortest path algorithm has a complexity proportional to square of the number of nodes ($\mathscr{O}(N_V^2)$)). Moreover, the number of traffic detectors in a city like Brisbane is of few thousands and therefore a description of the graph with hundreds of thousands objects is surely too much detailed. For example, many nodes exist only as a way to detail the road geometry (those nodes were intermediate points on the polylines[2]). In this section, a method for diminishing the number of nodes of the graph while keeping a precise knowledge of the real infrastructure is presented.

Let us now denote by $\mathscr{G}^\star = (V^\star, L^\star)$ the original graph as obtained above. In this case, each element in $V^\star$ is of size 2 (for each node: its two coordinates) and each element in $L^\star$ is of size 3 (for each edge: the node index in $V^\star$ it is starting from, the one it is going to and the length of the road segment). In addition, let us define S the set of Bluetooth scanners of size $M \times 2$ containing, for each of the M Bluetooth scanners, their two coordinates. Finally, let us define $\mathscr{M}_r^V$ the mapping from the space of scanners S to a space of nodes V with parameters r, such that, $\mathscr{M}_r^{V^\star}(s)$ is the set of vertices in $V^\star$ within r of the scanner s. $\{\mathscr{M}_r^{V^\star}\}_{s \in S}$ is the set of nodes in $V^\star$ within r of any scanner in S.

Then, we propose the following process:

- Let us initiate our simplified graph $\mathscr{G} = (V, L)$ as a copy of $\mathscr{G}^\star = (V^\star, L^\star)$. Let us also define L_{Map} of size $|L^\star|$, a variable keeping track of the link modifications, initialised with zero values. L_{Map} gets a '-1' value when an edge does not appear in the resulting graph anymore. If an edge has been combined with another link, its value in L_{Map} is the index in $L^\star$ of this other link.

- While any value changed in L_{Map} in the previous iteration do the following:

[2] A polyline is a connected sequence of line segments created as a single object also called polygonal chain

1. For every node $v \in V/\{\mathscr{M}_r^V\}_{s\in S}$:
   a) If the node corresponds to a dead end, that is it has either only one link (whichever direction it is) or only two links shared with another same node, then, those links are flagged for removal (*i.e.,* flag those links with '-1' in L_{Map}).
   b) If the node is along a one way road, that is, it has one incident link and one exiting link, then, the incident link becomes the concatenation of both links and the exiting link is flagged for removal (*i.e.,* flag the exiting link with the index of the incident link in L_{Map})).
   c) If the node is along a two ways road, that is, it has, twice, two links shared with another same node, then, both incident links are prolonged with the corresponding exiting links and the exiting links are flagged for removal (*i.e.,* flag with corresponding index in L_{Map})).
   d) If the node is along a one way road for one direction of travel and is a dead end for the other direction of travel, that is it has one incident link, one exiting link and another link corresponding to the other direction of travel for one of those two links, then this last link is flagged for removal (*i.e.,* flagged with '-1') and the node is then treated as in case (b).
2. For every link $l \in L$:
   a) If its origin is identical to its destination, then, it is flagged for removal ('-1' flag).
   b) If it has same origin and destination than another link, then, the link with the length the closest to the beeline distance from origin to destination is kept while the other one is flagged for removal (*i.e.,* flagged with the index of the other one).

- Obsolete links are removed from the list L: $L \leftarrow L(L_{Map} = 0)$.
- Accordingly, nodes that do not appear in L are removed from V.
- L and V are re-indexed accordingly to the new cardinality.

Note that this process ignores nodes in $\{\mathscr{M}_r^{V^\star}\}_{s\in S}$. This is justified as it keeps unchanged the road infrastructure around the Bluetooth detectors, where traffic is observed. From now on, one will refer to the original graph as $\mathscr{G}^\star = (V^\star, L^\star)$ and to the simplified one as $\mathscr{G} = (V, L)$. More formally, this procedure can be written as in Algorithm A.1 in Appendix A.

This procedure, for the Brisbane OSM network, with parameter r set to 150m, requires 13 iterations of the `while` loop, and the size of the graph decreases from $N_{V^\star} = 78\,000$ and $N_{L^\star} = 121\,000$ to $N_V = 8\,900$ and $N_L = 18\,300$ with 2 300 nodes in $\{\mathscr{M}_r\}_{s\in S}$. Meanwhile, the array L_{Map} enables to keep track of the modification of the graph and therefore of the initial infrastructure. Hence, it will allow us, later on, to be able to adapt information (like Taxi GPS tracks), matched on the initial graph to this new simplified graph. Figure 3.1 represents the Brisbane network before and after simplification.

2 Traffic Counts

2.1 Nature - Principle

Traffic counts are among traffic data that were first automatically collected. They consist in measuring the number of vehicles on a set of roads: only volumes are measure and vehicles are not identified. Temporary measurements of traffic counts are usually performed with pneumatic tube: the air inside

(a) (b)

Figure 3.1: (a) Network as downloaded from OpenStreetMap composed of 122000 segments and 77500 extremities. (b) Simplified networks with 18300 links and 8900 nodes. Red dots are the Bluetooth scanners.

the rubber tube is compressed by the wheels of heavy enough vehicles. The overpressure spikes are used to counts the number of vehicle axles. The number of vehicles can be derived by dividing the number of spikes by the average number of axles per vehicle (when known or assumed). Alternatively, algorithms have been designed to discriminate spikes caused by the same vehicle from spikes caused by two or more vehicles. This technology has been developed in the thirties [168] and is still popular for it can be easily installed, moved and removed.

In the sixties, magnetic (or inductive) loop detectors have been widely introduced in occidental cities [169]. This technology aims to measure Eddy currents induced by a ferro-magnetic masses moving above the loop and can be used for several applications, that are, mostly, traffic counts and waiting cars at traffic lights detection. Since then, other technologies have been developed for traffic volumes measurements: video, infra-red, LiDar, among other, but none are as widely deployed as the magnetic loops.

Brisbane City Council has kindly provided us with the SCATS® data stemming from some 915 intersections equipped with some 8000 magnetic loops. The dataset consists in two kind of files:

1. Intersection Layout: For each intersection, the plan, as a *pdf* file of the intersection with the position of the loops (also called approaches).

2. Counts per region: Those *txt* files contain, for a defined period of time (in this case from the 27th to the 31st of October 2014), for each 5 minutes interval, and, for each intersection in the chosen region, the traffic measured by each of the loop detectors.

2.2 Formatting the Data

Formatting the traffic counts so that the data could be combined with the Bluetooth data mostly consisted in manually interpreting each of the intersection plans in order to locate on the GIS representation of the road network, most of the 8 000 magnetic loops for which both the intersection plan and the count values were available. This has been done using the QGis software[3] so that the resulting file (as a Shapefile) could be interpreted with data analysis software.

In a second step, the counts have been analysed to identify faulty loops. Two kind of errors have been identified: First, some loops for which the counts are always 0, and second, some intersections for which the counts on every loop, at every time interval, are 2048 even for non-existing approaches. Those loops and values have been filtered from the traffic counts data. Last, the counts have been adapted to the simplified network and this will be presented in Chapter 6.

3 Bluetooth

3.1 Nature of the Data

Bluetooth is the global standard protocol IEEE 802.15.1 [170] for exchanging wireless information between mobile devices. It uses the 2.4 GHz short-range radio frequency bandwidth. Bluetooth is nowadays a standard technology included in many kinds of devices: smart-phones, smart-watches, hands-free kits, computers, tablets, in-car computers, etc.

The Bluetooth data, as available in Brisbane and several other cities worldwide (e.g., Calgary, Ankara, Adelaide, Sydney, Aarhus, Houston, Windsor, in Michigan...), are gathered by a network of time synchronised scanners, detecting and encrypting data about the discoverable Bluetooth devices in their surrounding. Those scanners are placed close to the roads of interest to maximise the detection of the devices belonging to users of the roads while minimising other detections: the low range frequency ensure that further devices, e.g., inside buildings, are not detected.

The Bluetooth devices are identified by their MAC IDs which stems for *Media Access Contol* Identification address. They are physical and unique identifiers stored within the electronic devices and are required by network connection protocols. The MAC ID is a code consisting in the combination of 6 alphanumeric pairs (Hexadecimal), ensuring a large collection of possible addresses: The first 3 pairs are representative of the manufacturer and are allocated by the Institute of Electrical and Electronics Engineers (IEEE) while the last three pairs are set by the manufacturer. The uniqueness is required by the connection protocol as devices with identical MAC IDs would not be able to pair together.

The Bluetooth MAC scanners (or BMS) detect discoverable devices within their communication range to which we will refer in the following as *scanning* (or *detection*) *area* (or *zone*). Conceptually, by matching the unique MAC IDs with the BMS one can get a sequence of logged timestamps corresponding to the movement of the associated user.

Concerning the Brisbane dataset, Bluetooth data are gathered by the Brisbane City Council (BCC), since 2007, with a pilot project of one scanner. The dataset available at the time of the writing of this manuscript ranges from Mars 2007 to October 2014. In October 2014, 580 Bluetooth detectors were installed. Early 2016 this number reached above 900 scanners. The data are stored within SQL tables as presented in Table 3.1, Table 3.2, Table 3.3 and Table 3.4.

[3] http://qgis.org/

Table 3.1: BCC's BLUETOOTH Table Example

id	deviceid	areaid	entered	duration
1	3085965	10773	8/10/2014 9:00	39
2	4252289	10622	8/10/2014 9:01	1
3	4452098	10282	8/10/2014 8:59	39
4	1003409	210386	8/10/2014 8:56	1
5	2428042	10291	8/10/2014 9:00	39
6	2316529	10622	8/10/2014 9:01	1
⋮	⋮	⋮	⋮	⋮

BLUETOOTH tables (Table 3.1) are created for each month of data and collect necessary information concerning scanning events:

- The **id** field indexes the detection events in the table.
- The **deviceid** field consists of identifiers each encrypting a single MAC ID. MAC IDs are numbered by order of appearance. Each time a new ID is detected by one of the BMS it is encrypted with the first available key. It exists therefore a table doing the correspondence between the **deviceid** and the MAC ID, but it is classified by BCC and one cannot access this table without special authorisation. This method is debated for its lack of security and for its lack of efficiency: each time an ID is scanned, it has to be compared to the full table for a match with its encryption key. Moreover, the size of this encryption table increases with each new MAC ID detected. It is estimated to be composed of several millions entries in early 2016.
- The **areaid** field stores the identification number of the scanner at which the event occurred. It also corresponds to the identification number of the intersection covered by the scanner.
- The **entered** field contains the timestamps of the detection events.
- The **duration** field corresponds to the time in second during which the MAC address has been detected by the same scanner.

Table 3.2: BCC's AREA Table Example

ID	AreaNum	LocationID	Region	Suburb	st1	...
10285	B0285	960	7	ASHGROVE	WATERWORKS RD	...
10539	B0539	399	9	MACGREGO	MAINS RD	...
10150	B0150	145	7	MITCHELT	OSBORNE RD	...
10320	B0320	541	6	NORTHGAT	TOOMBUL RD	...
19009	B9009	748	17	BOWENHIL	INNER CITY B	...
10417	B0417	285	5	EST BRIS	LYTTON RD	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

The AREA table (Table 3.2) is a table containing most of the intersections in Brisbane and thus where Bluetooth devices or magnetic loops are (or might be) installed. It is made of:

- The **ID** field with the indices of the intersections, to which refers the **areaid** field of the BLUETOOTH table.
- The **AreaNum** field with the indices of the intersections as initially indexed by BCC.
- The **LocationID** field that gives a correspondence with the index at which the coordinates of the intersection can be found in the LOCATION table.
- The other fields, **RegionID**, **Suburb**, **st1**, and fields thereafter give more information on the intersection.

Table 3.3: BCC's LOCATION Table Example

ID	**X**	**Y**	**AMGX**	**AMGY**
1	153.0206	-27.4679	502033	6961733
2	153.023736	-27.468108	502342	6961710
3	153.025435	-27.466789	502510	6961855
4	153.026305	-27.466121	502596	6961929
5	153.027165	-27.465462	502681	6962002
6	153.028824	-27.464153	502845	6962147
⋮	⋮	⋮	⋮	⋮

The LOCATION table (Table 3.3) is a table gathering the coordinates of every intersection in Brisbane both as latitude-longitude (X,Y) in the Coordinate Reference System EPSG:4326, WSG 84 and as (AGMX, AGMY) in the Australian Map Grid 1984 Coordinate Reference System [171]. A manual verification of this table led to the identification of around 50 intersections with wrong coordinates.

Table 3.4: BCC's DEVICE Table Example

id	1	2	3	4	5	...
OUI	00:08:E0	00:10:48	00:1D:F6	00:24:91	00:13:6C	...

The DEVICE table (Table 3.4) is a table gathering the 3 first pairs of digit of the MAC address, that is, the part of the MAC address identifying the manufacturer and making a correspondence with the **deviceid** field of the BLUETOOTH table.

From these tables, queries can be performed in order to locate scanners on a map as illustrated by Figure 3.2 (joint query between the BLUETOOTH table (using the command `DISTINCT areaid` to get a list of scanners actually in use), the AREA table (`ID, LocationID`) and the LOCATION table (`ID, (X,Y)`). Then, one can also represent in chronological order a sequence of detection as in Figure 3.3 or in Video 1[4] to get a first sense of vehicle movements in Brisbane.

Although the data available range from 2007 to October 2014, in this manuscript we focus mostly on two particular periods of time. The first period of interest is around the Thursday 24th of July 2014, the latest date at which taxi data were also available to us (*cf.* Section 4). The second period is the week from Saturday the 25th to Friday the 31st of October 2014, period for which SCATS®traffic counts have also been provided by the Brisbane City Council.

[4] http://perso.ens-lyon.fr/gabriel.michau/Videos/5_cars.mp4

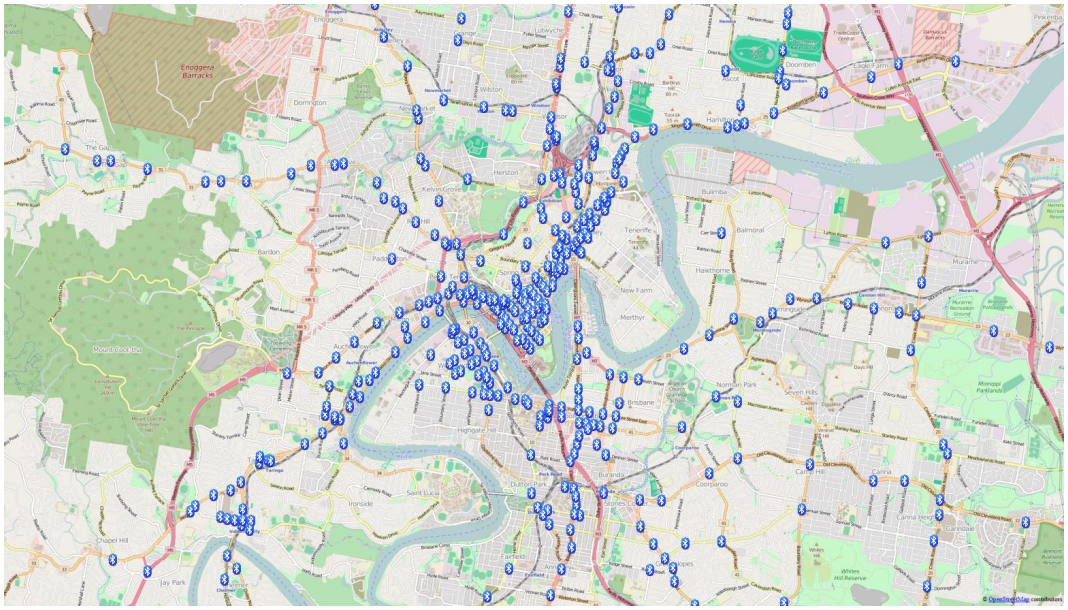


Figure 3.2: Map of Brisbane with the Bluetooth logos representing BMS, located according to the LOCATION table. (Base Layer: OpenStreetMap).

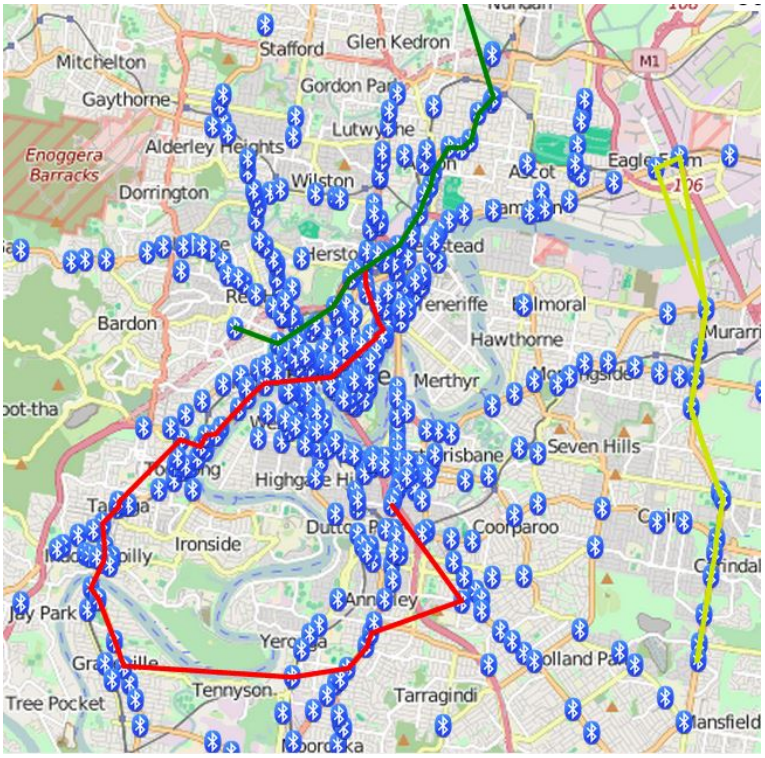


Figure 3.3: Representation of 3 sequences of detections. Detections are chronologically ordered and each successive pair of detections are linked together. (Base Layer: OpenStreetMap).

The aim of the works presented in this manuscript is to take full advantage of the Bluetooth data within the study of traffic information retrieval. Therefore, it does not focus on the impact of the scale of the Bluetooth technology network. Its study would be interesting and possible as the Brisbane Bluetooth scanners were continuously installed over almost 10 years, starting from 1 scanner in 2007 to over 900 in early 2016. This could be explored as part of future works. In any case, this justifies our focus on narrow periods of time, which corresponds to the latest period at which the needed datasets could be gathered, hence ensuring that a maximum number of Bluetooth scanners were in use.

3.2 Characterising the Bluetooth Data

At first sight, the Bluetooth technology, by its capacity to identify Bluetooth devices with timestamped detections, seems to offer innumerable opportunities for information retrieval, ranging from travel time estimation, trajectories analysis, OD matrix sampling to discreet choice route model calibration, among others. We will show in the following, however, that even if opportunities offered by the Bluetooth technology are not be looked down at, the BMS networks, at least as the one installed in Brisbane, have major inherent noises and biases. We will show next, in Section 3.2.1 that the inquiry cycle leads to an important uncertainty on timestamps. In Section 3.2.2, we will discuss the radius of the detection zones, source of spatial uncertainty, which can lead to erroneous interpretation of travel patterns. Moreover, in Section 3.2.6 we will demonstrate that some Bluetooth devices share a same MAC ID and can therefore not be properly identified. Last, in Section 3.3, the problem of detectors not detecting Bluetooth devices within range is analysed and quantified.

3.2.1 The Inquiry Cycle

A Bluetooth device has two major states: standby and connection [161], and seven substates, one of which is the *inquiry* state during which the master Bluetooth device (here, the BMS) prospects pseudo-randomly the Bluetooth frequency band (slots of 79.1 MHz in the 2.4GHz band), with a frequency change every 1.28s. Therefore, inquiry cycles of 10.24 seconds are recommended by Bluetooth standards for a high probability of Bluetooth impairing. For more details about the Bluetooth inquiry cycle, the reader can refer to [161], [172].

The specifications of the inquiry cycle have their importance as it impacts directly the timestamps and the duration times recorded by the system. This is illustrated in Figure 3.4 and in Figure 3.5. In Figure 3.4, the occurrences of detection over two minutes for the whole set of BMS are plotted as a histogram with one bin per second. It illustrates that either scanners are synchronised to the second and inquiry cycles can start every $\sim 8-9$ seconds and last at least $\sim 17-18$ seconds (e.g., the scanner 179003 represented with blue crosses finishes its cycles on seconds 6, 23, 49, 66, 84, 101 and 118 while the scanner 10301 represented with black circles finishes its cycle on seconds 14, 40, 58, 92 and 109) or that the data are stored in the database every $\sim 8-9$ seconds. Moreover, the cycles never last less than ~ 17 seconds but sometimes can last some extra half standard cycles ($\sim 8-9$) (e.g., the second cycle of scanner 179003 or the first cycle of scanner 10301 in Figure 3.4).

Similarly, in Figure 3.5 is the distribution, as a histogram of the field **duration** over one week of data (25th to 31st of October 2014) with a resolution of 1 second. In this figure, some values can be identified, being up to 10^5 times more frequent than others. Those values correspond to multiple of the inquiry cycle duration. The minimum value of the **duration** field is 1 seconds. Thus the **duration** can be broadly computed as:

$$\text{duration} = n_c \times (\Delta_c + 1) \tag{3.1}$$

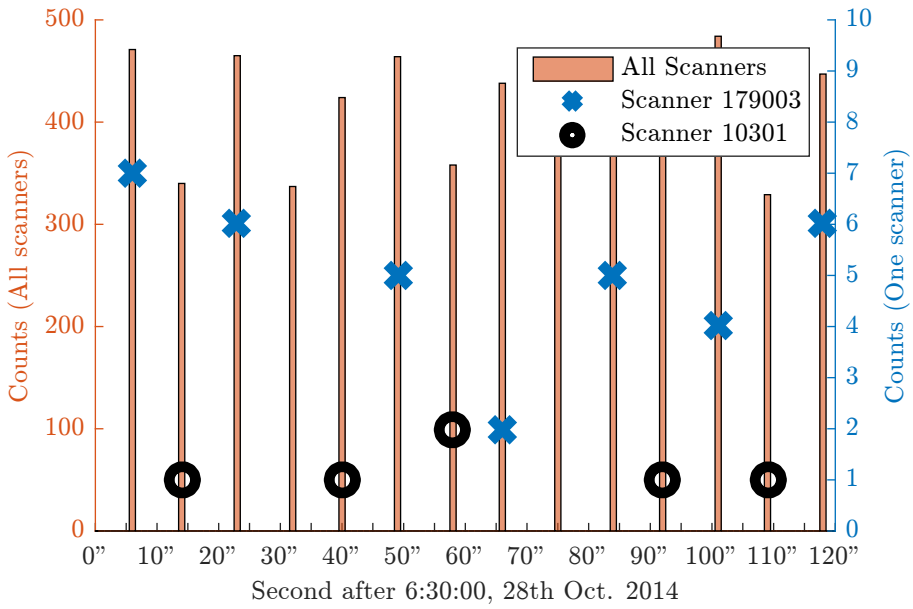


Figure 3.4: Plot as a histogram (one bin per second) of the timestamp of every detection between 6:30:00 a.m. and 6:32:00 a.m. on the 28th of October 2014: red bars are representative of all the BMS in the networks, blue cross represents one particular scanner (**areaid**=179003) and black circles another (**areaid**=10301). (∼5700, 35 and 6 recorded detections for all scanners, scanner 179003 and scanner 10301 respectively.)

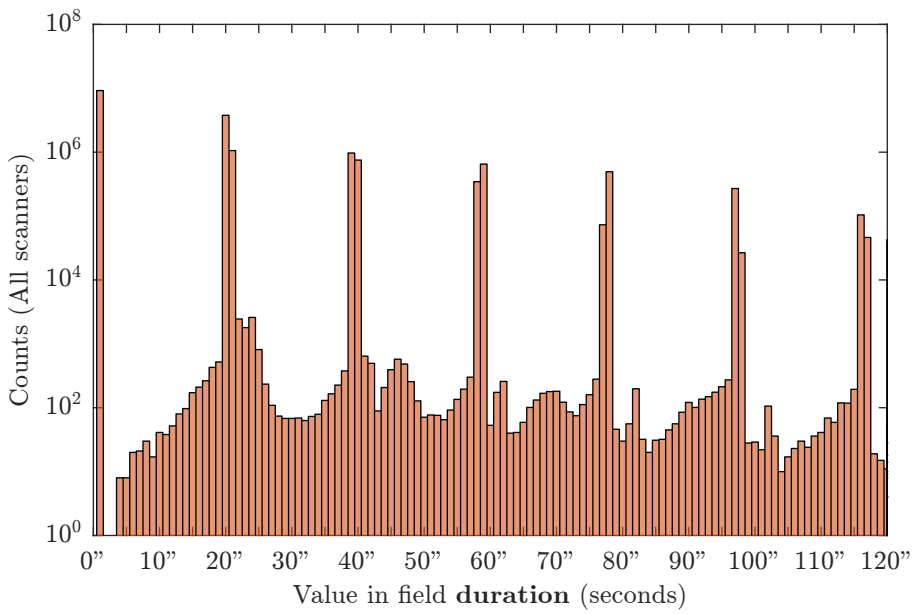


Figure 3.5: Plot as a histogram (one bin per second) of the values in the **duration** field. Note that the Y-axis is logarithmic. (Data: 25-31 October 2014, 18 750 092 recorded detections)

where n_c is the number of cycles during which the device was in the surrounding of the BMS and Δ_c the time required for an inquiry cycle ($\sim 17-18$ seconds) to which the minimum of one second is added. The peaks indeed appear every ~ 19 seconds.

This effect has a huge impact on the reliability of the data as the time precision of the detection is of at most $\sim$17 seconds. This uncertainty on the detection time is substantial in view of the inter-BMS distances: the average of the nearest neighbour distances for this network is 363m (median is at 211m), that is 26 seconds (resp. 15 seconds) at a speed of 50 kph, to be compared to the 17 seconds of a cycle.

Last, the question of what information is really provided by the **duration** field is discussed: If one can expect that it corresponds to the time during which a device has been detected at every successive cycle by the same BMS, it appears in fact that detections by the same detector, in the dataset, never occur within less than 78 seconds, that is around 9 half cycles (*cf.* Figure 3.6). Thus this suggests that if a device is detected by the same scanner with less than 4 cycles in between the two successive detections, it is considered by BCC's system as the same detection and the field duration is updated accordingly. This threshold is a choice of the Transport Engineers who designed Brisbane's system and is only a way to reduce the amount of entries in the Bluetooth table. Again, this information has to be taken into account when one try to infer information on the traffic states from the **duration** field [157]: the **duration** value also has a precision of only 17 seconds and devices detected by the same scanner within 78 seconds will only get a single entry in the detection table, independently of what the device did in this time laps.

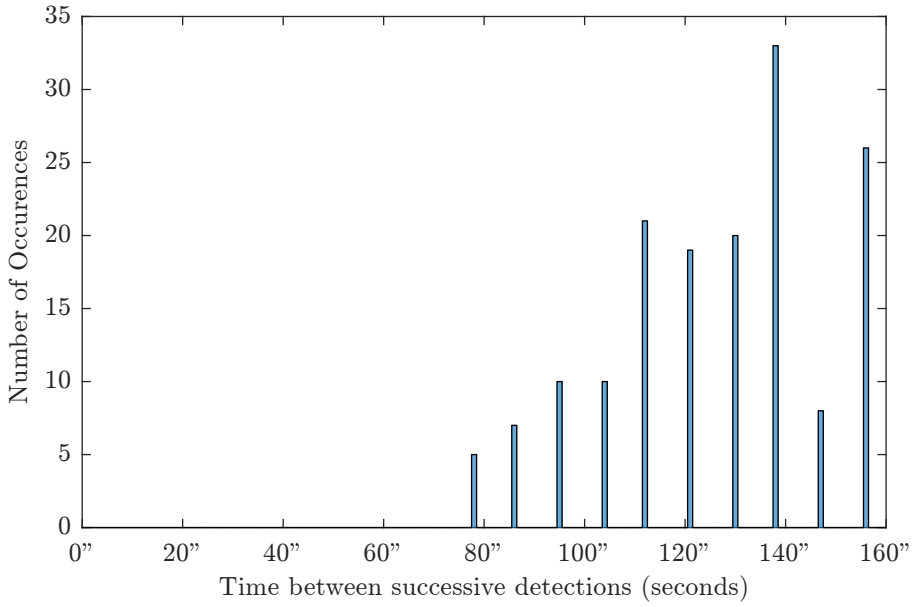


Figure 3.6: Distribution of time intervals between successive detections at the same scanner (one bin per second) for every BMS of the network. (Data: 25-31 October 2014).

3.2.2 Overlapping Detection Zones

In addition to time uncertainty, BMS also have an inherent spatial uncertainty: BMS such as these installed in Brisbane scan the MAC ID over a certain area. With a 5 dBi omni directional antenna, a

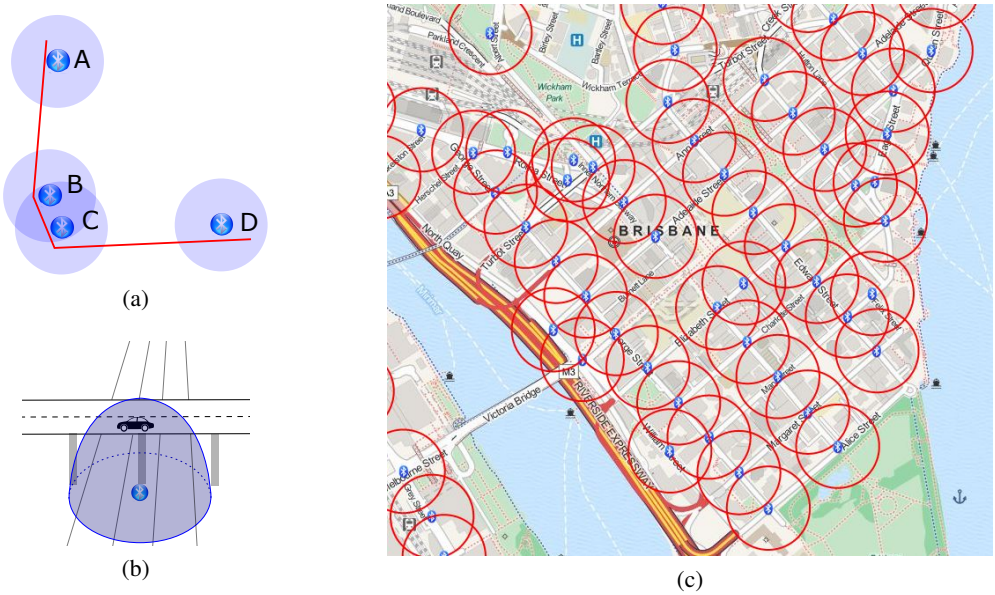


Figure 3.7: (a) A car following the itinerary ABCD might be detected as ACBD. (b) A BMS might detect vehicles on another corridor than the target one. If this other corridor is not monitored by other BMS, the BMS appears wrongly as origin or destination. (c) Map of Brisbane CBD with BMS and 100m radius circles.

BMS scans the MAC IDs over a zone of 100m radius [135]. The exact location of the device within the scanning area is therefore unknown and this has two other consequences:

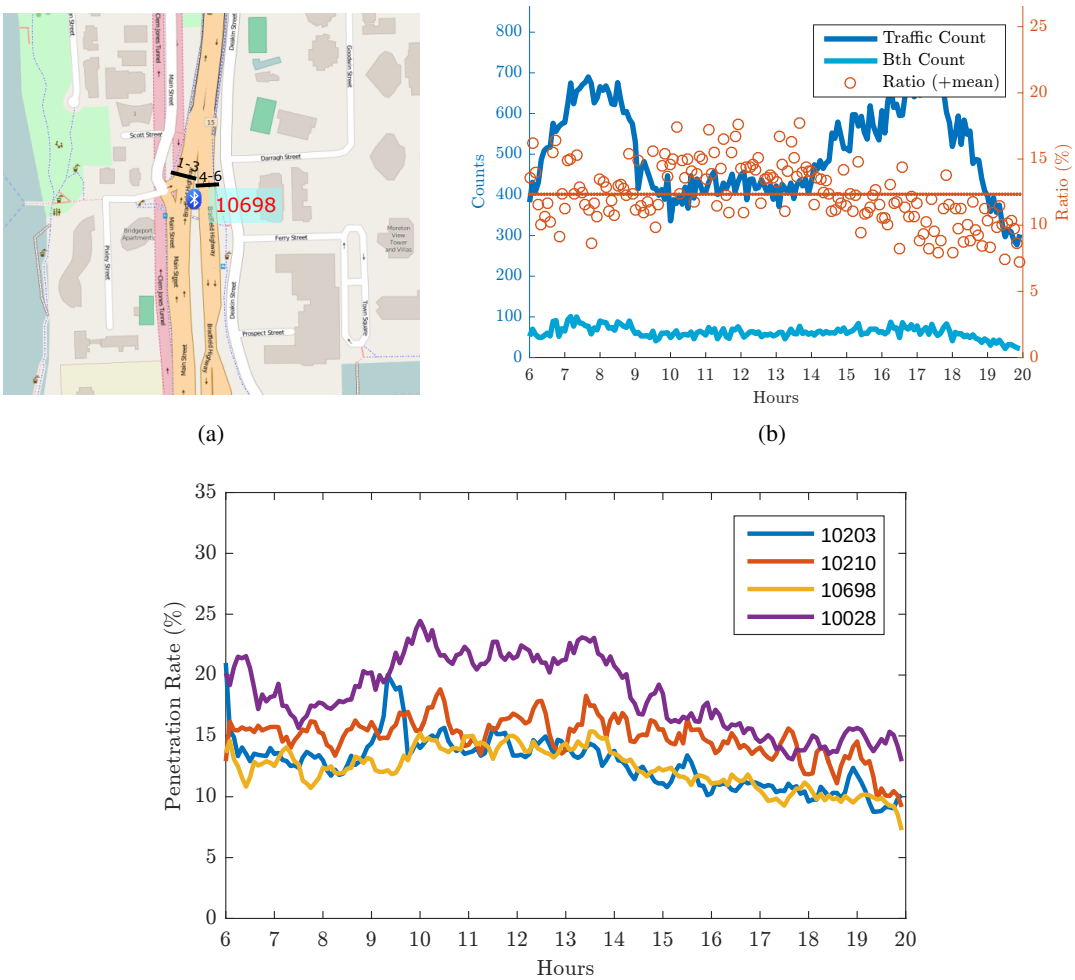
Firstly, sensors located in close proximity one to another can have overlapping detection zones. Accordingly, a downstream scanner might detect a device at the same time or before the upstream one does, yielding later to erroneous interpretation of the travel patterns. Such travel patterns are illustrated in Figure 3.7a. In our Brisbane dataset, 3.5% of the pairs of successive detections (with identical MAC ID) happen to have exactly the same timestamps. This value reaches 12%, 18%, and 25% if one looks for scanners less than respectively 200m, 100m and 16m apart, 16m corresponding to the smallest distance in-between scanners. Those values reach receptively 21%, 55%, 66% and 68% if one looks at pairs of successive detection within one scanning cycle (17 seconds). In this first analysis of the data, the detections are ordered chronologically and not according to the real movement of the devices. Therefore, it is not possible to quantify the amount of successive detections for which the downstream scanner detected the device before the upstream one. It is however a fact to be taken into account and which can sometimes be detected by monitoring, for each device, the main direction of travel. To give a sense of the overlapping detection areas issue, Figure 3.7c represents, for the CBD of Brisbane, the BMS with 100m radius circles around them.

Secondly, the detection area might span multiple corridors. Thus, the traffic that is detected by a sensor may not necessarily belong to the target corridor. Figure 3.7b illustrates this phenomenon: the detected car is driving in a corridor (e.g., a bridge) that is different from the target corridor (the road underneath). This can become a real bias if this other corridor is not covered by Bluetooth sensors. A concrete example in Brisbane is the one of the Pacific Motorway which is not manage by BCC (but by Transport and Main Roads) and for which data are therefore not available. When one is interested by origin and destination information inferred from first and last detections of Bluetooth devices, similarly to what will be presented in Chapter 4, Section 1, it appears that BMS close to the motorway are always important origins and destinations. Actually, devices entering the motorway

will often leave the urban area covered by BMS with the motorway and not be detected later on. This effect can lead to erroneous origin-destination patterns, and consequently, such biased sensors have to be identified.

3.2.3 Penetration Rate of the Bluetooth

The penetration rate is defined as the fraction of users equipped with a given technology. The penetration rate of the Bluetooth technology can be estimated by comparing the number of Bluetooth detections to the traffic counts at intersections. For example, in Figure 3.8 counts are compared at intersection 10698, on Tuesday the 28th of October 2014 (Figure 3.8a represents the intersection and Figure 3.8b compares the measured volumes). This intersection has one Bluetooth scanner and 6 magnetic loops. Figure 3.8c summarise, for the same day, measured penetration rates for three intersections scattered in Brisbane.



(a) (b) (c)

Figure 3.8: Comparison of Bluetooth counts with magnetics loops. (a) Intersection 10698 the Bluetooth scanner and magnetic loops (black line, numbers correspond to loops). (b) Traffic Counts, Bluetooth Counts, Ratio and its mean at intersection 10698. (c) Bluetooth to total flow ratio at four intersections. (Data: from 6:00 a.m. to 8:00 p.m. on the Tuesday 28th of October 2014.)

The penetration rate varies between 10 and 25 %, a value consistent with [159] while other works proposed higher rates [173] (15 to 35%). Those value are likely to be underestimated however, as Bluetooth scanners often miss passing cars (*cf.* Section 3.3) and because magnetic loops might tend to over count the traffic [174]. These effects are hard to quantify; assuming a 15% rate of missed detections and a 5% error on magnetic loops, the real penetration rates are some 21% higher. Thus, penetration rates should be in the range 12% to 30%.

3.2.4 Distribution and Dynamics

Basic analysis of the BLUETOOTH table can already provide information on the dynamic of the traffic in Brisbane City. Figure 3.9 illustrates the evolution of global traffic flows in Brisbane for one week (Figure 3.9a), and for one weekday (Figure 3.9b). One can easily identify, on weekdays, the sharp morning peak, ranging from 6:00 to 9:00 a.m., the broader evening peak, from 3:00 to 7:00 p.m., the minimum traffic flows at 2:00 a.m., and a gentle midday peak, from 1:00 to 2:00 p.m.. This dynamic can also be represented by means of a video as in Video 2[5].

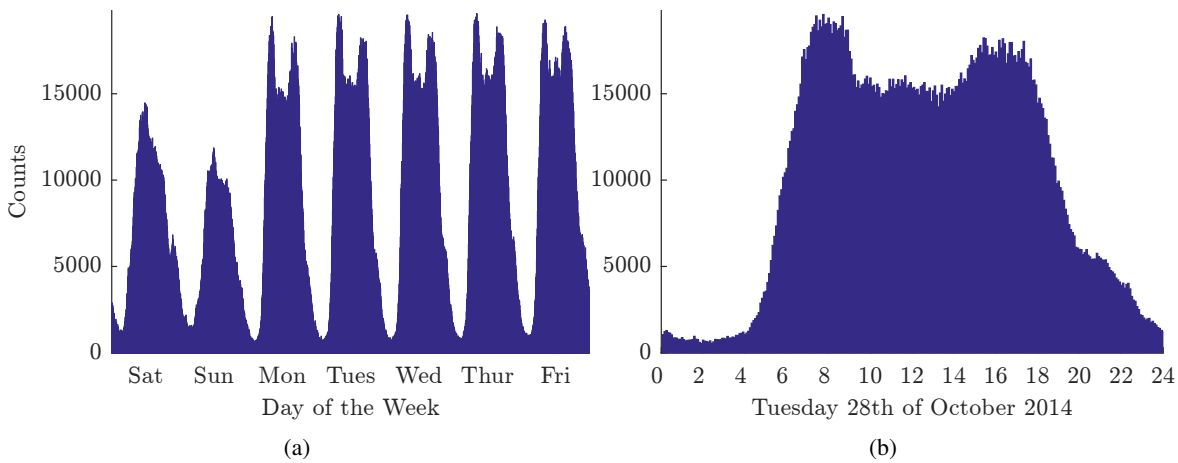


Figure 3.9: Bluetooth number of detections per 5 minutes, (a) over one week, (b) over one day of the week (Tuesday). (Data: 25th to 31st of October 2014).

Then, by looking at the time differences between pairs of successive detections (for the same Bluetooth device) as represented in Figure 3.10a, it appears that 40% of the detections happen within 10 minutes of each other, 65% of the detections within half an hour. By dividing the distance between the two scanners (here as the Euclidean distance), by this time difference, an indicative speed can be computed: Figure 3.10b represents a histogram of those speeds. Two peaks can be identified, one centred on 0-1 km per hour, with $\sim$ 15% of the successive detections corresponding to a speed slower than 5 kph, and another peak at 35 kph. Those values are consistent with other measures on the Brisbane network [175].

The identification of such values is important as they can help to discriminate, within a sequence of detections for a given Bluetooth device, the trip(s) it is made of. Such sequencing is the object of Chapter 4, Section 1.

[5] http://perso.ens-lyon.fr/gabriel.michau/Videos/Average-oct-3-1.mp4

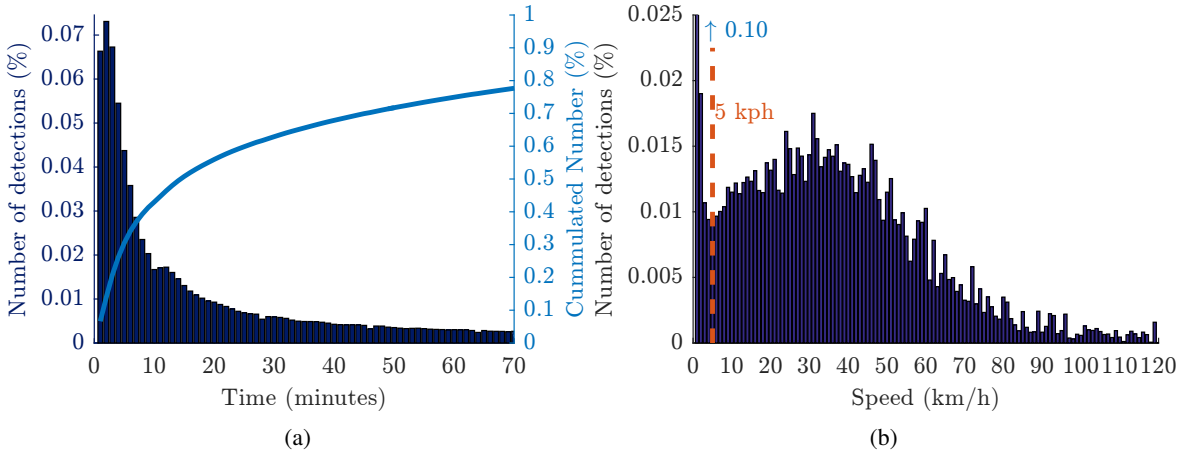


Figure 3.10: (a) Distribution of time differences between successive detections. (b) Distribution of inter-detection speeds calculated as the Euclidean distance between scanners divided by the timestamps difference. Speed less than 1kph represents 10% of the distribution. (Data: 25th to 31st of October 2014).

3.2.5 MAC identifier

Another question concerning the Bluetooth data is whether it is possible to get a hint of which kind of Bluetooth devices are involved. This question is of importance as different devices might have different behaviours or characteristics. For example, while phones can automatically switch to hibernation after few minutes of non-use, other devices like GPS navigation devices or Bluetooth hand-free-kits are more prone to be connected during the full length of the trip and could therefore be considered as more reliable.

It is not possible, from the available data and for privacy reasons, to get a precise identification of the detected devices. However, as presented in Section 3.1, the first three pairs of hexadecimal digits from the Bluetooth MAC Address are representative of the manufacturers, most of which are registered with the IEEE Standards Association through the Registration Authority program. Thus, a table matching the first half of the MAC address to their corresponding manufacturer (if registered) is available[6] and can be used to match the detected Bluetooth devices to their manufacturer. Such a match involves the **DEVICE** table, which provides a correspondence between the **DeviceID** of the device (in the **Bluetooth** table) and the first three pairs of digits of the MAC address.

By doing so, the share of each manufacturer among the detected devices can be derived and Figure 3.11a illustrates the share of the most represented manufacturers. Similarly, Figure 3.11b illustrates the share of each manufacturer among the whole set of detections.

The main representative manufacturers are: Parrot, Nokia, Bury, ATO Technology, etc., all well known for manufacturing hand-free-kits and headsets. Then, other well represented manufacturers are: Novero specialised in in-car connectivity and communications, ALPS, a major electronic multinational involved in Automotive, Home, and Mobile connectivity.

Headsets and in-car Bluetooth chips seem thus to generate most of Bluetooth activity recorded by the system. Some manufacturer only specialised in GPS navigation systems (TomTom, Garmin)

[6] http://standards-oui.ieee.org/oui.txt

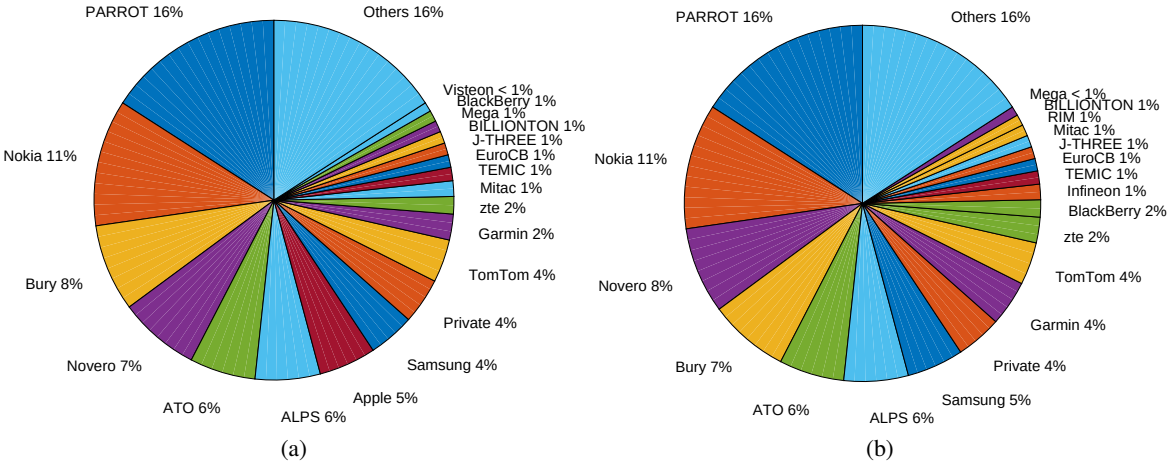


Figure 3.11: Share of each registered manufacturer, (a) among detected MAC IDs, and (b), weighted by the number of detections. Only shares bigger than 1% are represented. Others are gathered within the 'Others' slice. (Data: 25th to 31st of October 2014).

have a more important share in term of detections than in term of devices while, to the opposite, manufacturers only specialised in Mobile Phone, have a smaller share in the detection set than in the ID's one. For example Apple represents 5% of the detected devices but only a negligible fraction of the whole set of detections. Indeed, iPhone are known to go easily on sleep mode for battery economy and the data they produce might therefore not be representative of the full trip achieved by its user.

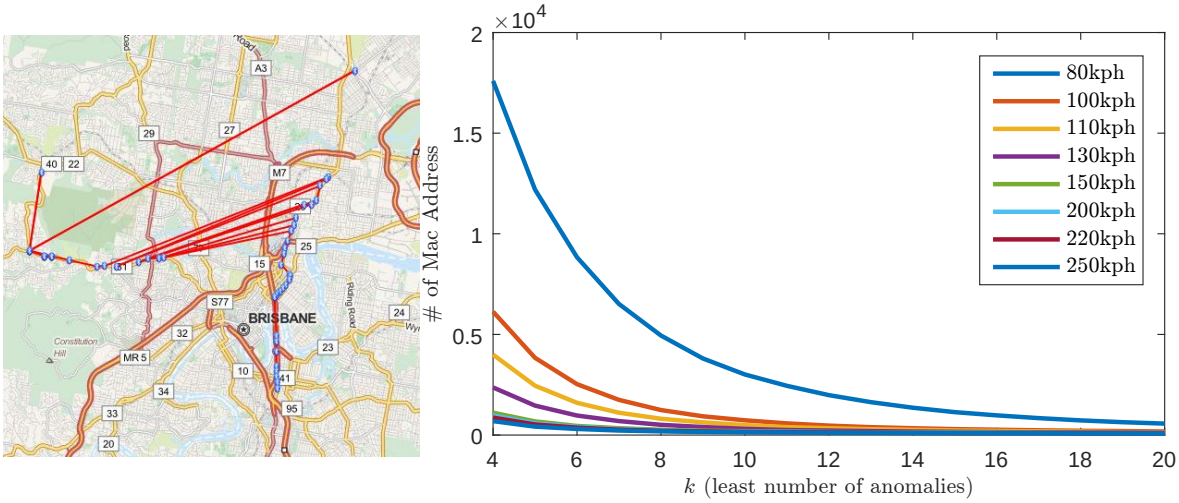
In a nutshell, those results are mostly positive as they prove that cars-related devices represents most of the detections while providing data of better quality. In addition, they are also more frequent among the detected Bluetooth devices.

3.2.6 Shared MAC address

Although MAC IDs are expected to be unique according to common standards, it appears from the dataset that some IDs are shared among vehicles. A possible explanation stems from the possibility to clone Bluetooth device parameters for fleet's specific needs [162]. As a matter of fact, the devices that shared their MAC are also frequent users of the network. This suggests that some MAC IDs may be shared among professionals (e.g., taxi drivers, postmen, etc.).

These shared MAC IDs can nevertheless be detected, as they will be likely to appear at two different places of the network, over a short period of time. The detections of a suspicious ID are illustrated in Figure 3.12a. The aerial speed is computed between the successive detections (as the Euclidean distance divided by the travel time) and very high speeds are indicative of *suspicious* IDs. From the dataset, it is observed that any threshold speed over 130 kph would detect a similar number of MAC IDs with 14 occurrences of speed higher than the threshold (*cf.* Figure 3.12b, 14 occurrences of speed above the threshold is the point of convergence of the four curves: 130, 150, 200 and 220 kph). Those two values, of 14 successive detections with a speed above the threshold of 130 kph, have thus been used for filtering suspicious MAC IDs. This corresponds to, in average, 2 parallel detections per day at speed over 130kph while the maximum speed on the urban network is 80kph.

This filters 0.2% of MAC IDs and 1.4% of detections.



(a) (b)

Figure 3.12: (a) Suspicious MAC address, 28th of October 2014 14:45 to 14:52. (b) Stat (Data: 25-231 October 2014, Base Layer: MapQuest OpenStreetMap).

3.3 Missed Detections

3.3.1 Evidence for Missed Detections

A major problem of the Bluetooth technology for traffic analysis stems from the fact that BMS and Bluetooth devices do sometimes not pair while being at reaching distance. Considering that, as seen in Section 3.2.1, the inquiry cycle lasts 17 seconds, with a detection radius of 100m, a car driving over 42 kph could go through the detection area in less time than required for an inquiry cycle. Other reasons might cause missed detections, some of which will be detailed in the following.

As a first step however, it is important to quantify this phenomenon within the Brisbane dataset. The main difficulty for characterising missed detections stems from the following paradox: if a car has not been detected, how can we know that it has been in the neighbourhood of a scanner? To this question we bring two separate answers:

- First, by looking at five corridors, cars detected at first and last scanners within a reasonable time window (so that it is likely for those car to have driven the corridor) are selected. Assuming the car went through the detection areas of all intermediate scanners on the corridor, BMS which did not detect the cars can be considered as missed detections. This method is further developed in the following.
- The other method we propose is developed in Chapter 4, Section 4.3, where, after having developed a method to recover trajectories from the Bluetooth data, the scanners on the recovered trajectories are compared with the actually observed detections.

For the corridor based approach, the focus is on five major corridors: Coronation Drive, Waterworks Road, Logan Road, Milton Road and Ipswich Road (see Figure 3.13). In a first analysis, each corridor is treated as follow: If h_{corr} is the number of detectors along the corridor, and g_i the number of observed detections for the i-th car on that corridor, then we compute the probability of detection

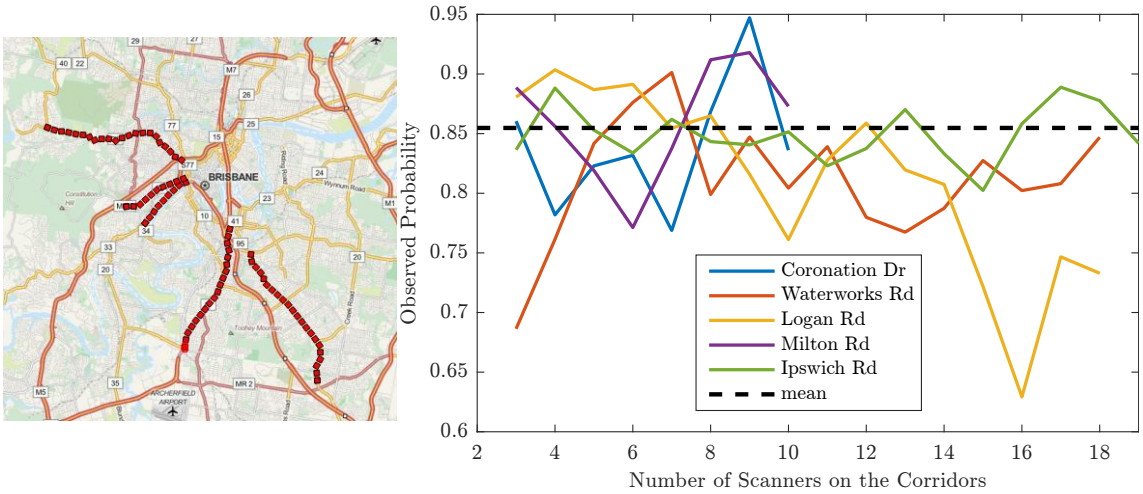
p as:

$$p = \frac{1}{N}\sum_{i}^{N}\frac{g_i - 2}{h_{corr} - 2}, \tag{3.2}$$

where N is the number of car detected on that corridor.

The first and last detections are imposed in order to be sure that the car was indeed on the corridor. It is thus the detection probability conditioned by the first and last detections that we are interested in. As a consequence, the first and last detections are removed from the statistics (hence $(g_i - 2)$ and $(h_{corr} - 2)$).

To virtually increase the number of road segments on which to perform the statistics, each corridor can be subdivided into smaller corridors: For example a corridor with h_{corr} detectors can be subdivided into 2 corridors with $(h_{corr} - 1)$ detectors, 3 corridors with $(h_{corr} - 2)$ detectors, up to $(h_{corr} - 2)$ corridors with 3 detectors. The value of p as per Equation (3.2) for the five corridors, and for all the sub-corridors is illustrated in Figure 3.13b.



(a) (b)

Figure 3.13: (a) Map of the five chosen corridors. (b) Observed probability of detection at intermediate detectors on the five corridors for increasing number of scanners taken into account (from 3 scanners to the maximal possible number). (Data: October 2014, Base Layer: MapQuest OpenStreetMap).

From this study, a first estimation of the detection rate on the network is 85% (*cf.* Figure 3.13b). This is however an upper bound as by selecting cars detected at least twice on the corridor, one ignores cars that were not detected at all or only once.

3.3.2 Independent Hypothesis

When an average probability of detection is estimated by methods similar to the one from the previous section, a hidden assumption is that this rate of detection is unique and therefore independent from other variables as, for example the Bluetooth device or the Bluetooth scanner.

To test for independence, a proof by contradiction is proposed. The independence in both the detectors and the Bluetooth devices is assumed and we show that the data do not obey to inferred

statistical relationships.

Let us assume that detections are occurring randomly and independently at a rate p. Then, for one corridor with h_{corr} detectors, the binomial relationship ensures that if v_k is the fraction of vehicles detected at the first, last and $(k-2)$ other detectors, then:

$$\forall k \in [2, h_{corr}] \qquad \begin{aligned} v_k &= p^2 \binom{h_{corr}-2}{k-2} \cdot p^{k-2} \cdot (1-p)^{(h_{corr}-2-(k-2))} \\ &= \binom{h_{corr}-2}{k-2} \cdot p^{k} \cdot (1-p)^{(h_{corr}-k)} \end{aligned} \tag{3.3}$$

From which it can be inferred that:

$$\forall k \in [2, h_{corr}-1] \quad \frac{v_{k+1}}{v_k} = \frac{p}{1-p} \cdot \frac{h_{corr}-k}{k-1} \tag{3.4}$$

$$\forall k \in [2, h_{corr}-1] \quad p = \frac{\frac{v_{k+1}}{v_k}}{\frac{v_{k+1}}{v_k} + \frac{h_{corr}-k}{k-1}} \tag{3.5}$$

In consequence, if the random independent hypothesis holds, p, the probability of a detection to occur should be independent of any variable, in particular of k. We show however that is not the case in Figure 3.14. In this figure, we computed p as per equation (3.5) for the five corridors and for increasing values of k. This figures clearly highlights an increasing trend for p as k increases. Thus the more a device is detected the more its probability to be detected next is high. In other words, all devices do not have the same detection probability. There might also be other sources of dependencies as for example the weather, the density or the detectors. In the next section, possible explanations for missed detections, found in the literature, are presented.

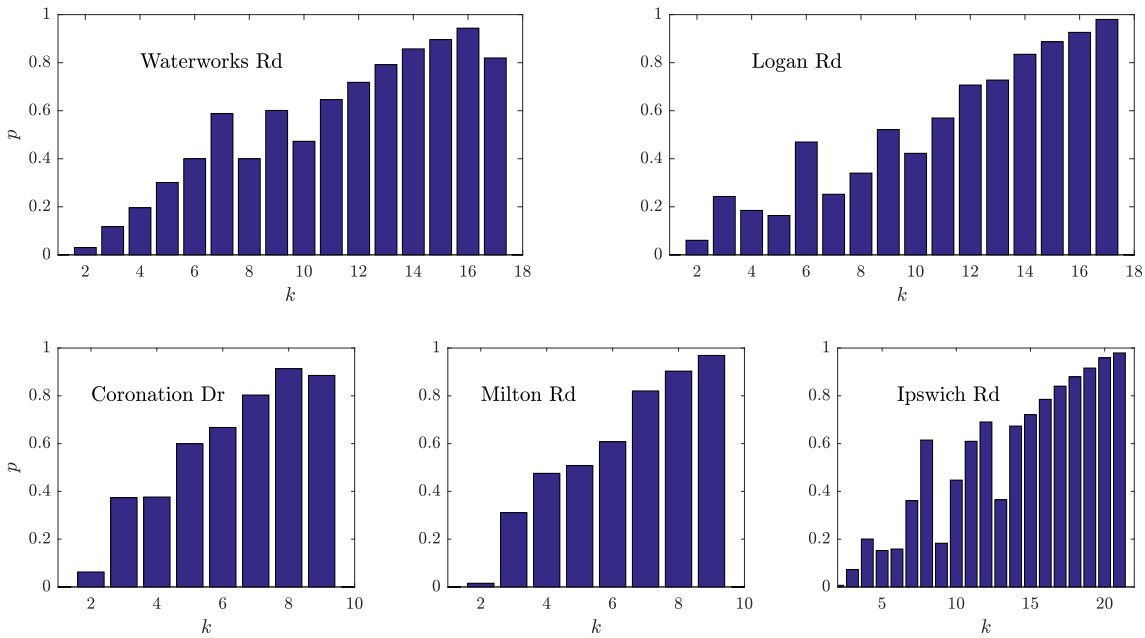


Figure 3.14: Probability of detection computed under the independence assumption as per (3.5) for increasing values of k on the five chosen corridors.

3.3.3 Origins for Missed Detections

From the literature, missed detections can be explained as follows:

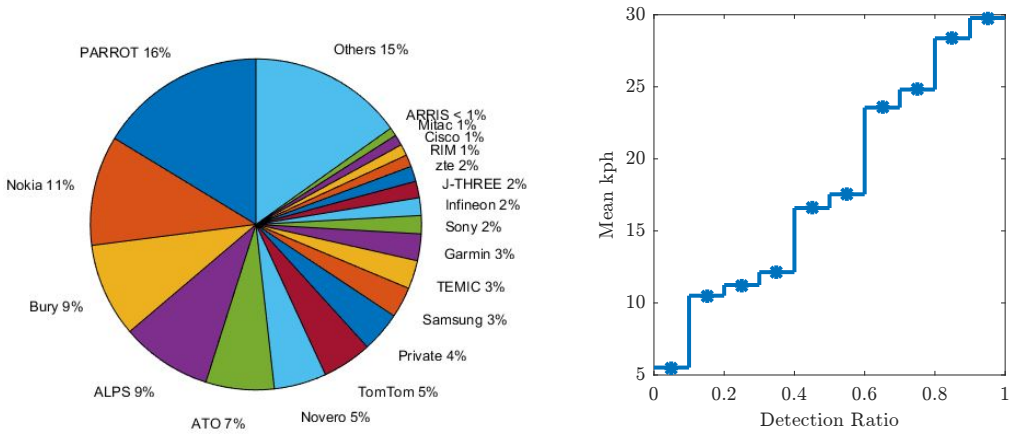
1. Not all scanners and devices are equally powerful, as some have stronger signal than others. From our dataset, we observed that some devices were more likely to be detected, compared to others. This assumption is moreover supported by the work of [166] highlighting the influence of the antenna on the signal strength and detection. However, by looking at the MAC ID for devices with lower detection rate, we could not identify manufacturers corresponding to low detection probability devices. Figure 3.15a, represents the share of the main manufacturers for devices with less than 30% detection rate. It is very similar to Figure 3.11a in Section 3.2.5.

2. A device moving fast enough could pass through the detection area of a BMS in less time than needed for the inquiry cycle (*cf.* Section 3.2.1). Thus, we could expect the average missed detection rate to increase with speed. Figure 3.15b represents the ratio of detections (rounded with 0.1 precision), plotted against the average speed of the vehicles with this detection ratio. This figure contradicts this explanation as a trend of increasing average speed with increasing detection rate clearly appears. However another cause with opposite consequence can conceal the expected increase of missed detection with speed and is presented thereafter.

3. The missed detection rate increases, as the scanning area becomes more crowded with active Bluetooth devices. In fact, it is known that when the number of detectable devices increases, interferences may affect the effectiveness of the detection [165]. This idea is supported by Figure 3.15b. Indeed when traffic density increases highly, speed is known to decrease. Thus a high density of Bluetooth devices would diminish the scanners efficiency while also being indicative of high traffic densities and therefore of lower speed.

4. The position of the BMS is of great importance, as Bluetooth signals are weakened by physical obstacle (e.g. walls and billboard). Brennan Jr, Ernst, Day, Bullock, Krogmeier, and Martchouk (2010) [137] have also shown that the vertical position of the antenna has an influence of the effectiveness of the BMS.

5. Last, not all Bluetooth devices are always in discoverable mode (e.g., some devices may become undiscoverable after a few minutes of non-use).



(a) (b)

Figure 3.15: (a) Manufacturer share amongst devices with a detection ratio lower than 0.3%. The shares are very similar to those in Figure 3.11a. (b) Average speed of the vehicles against their detection ratio. The ratios are rounded toward $+\infty$ with a precision of 0.1.

4 Taxi Data

4.1 Nature

The taxi company *Black & White Cabs* from Brisbane kindly provided us with the GPS tracks of their taxis. The data consist in daily files containing information on Taxis, ordered chronologically:

- The Taxi identification number. (One ID per vehicle).
- The status of the Taxi: *Ignition On*, *Logged On*, *Meter On*, *Ignition Off*.
- The timestamps of the information.
- The latitude of the vehicle.
- The longitude of the vehicle.
- The speed at which the vehicle is driving.
- The direction (angle with 8 degrees increment).
- The identification number of the driver. (This field is always *NULL* if the status is *Ignition On* or *Ignition Off*.)

As an example, in Figure 3.16 are plotted the resulting trajectories for one Taxi. The time between two successive information on the same taxi is of 30 seconds in 83% of the cases and between 29 and 31 seconds in 91% of the cases. The location of the Taxi is measured by a GPS chip therefore the precision of the information can be considered as accurate both for time precision ($\sim$1 second) and spatial precision ($\sim$ 5 meters). However the movement description of the taxi resulting from this dataset have a time resolution of 30 seconds only.

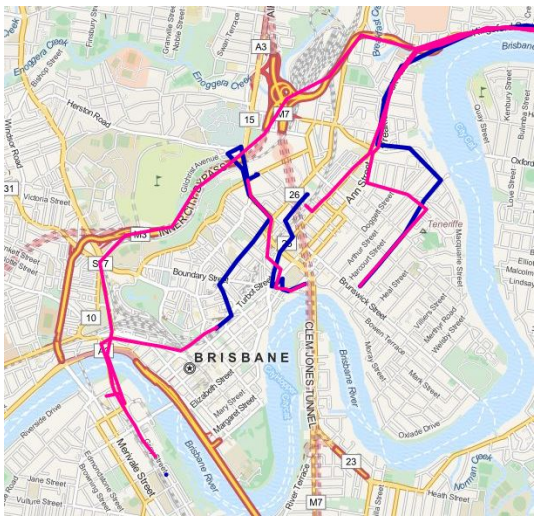


Figure 3.16: GPS tracks for one taxi between 5:00 and 7:00 p.m. on the 24th of July 2014 (Base Layer: OpenStreetMap). Blue lines correspond to status *Logged On* while the magenta lines correspond to the status *Meter On*.

4.2 Matching with Bluetooth Data

These GPS tracks have been matched with Bluetooth devices according to the following methods:

- First, driver identification numbers equal to the *NULL* string are filtered. A unique ID is created for each Vehicle-Driver pair and the data are ordered first, by this ID, then chronologically.
- Second, each of the resulting GPS track is compared against every MAC ID. If, for a given MAC ID a detection occurs while the GPS indicates that at a similar time the taxi was at a location that could not be reached within reasonable speed (less than 200km/h, more than 500m apart, to account for the Bluetooth position and timestamps uncertainty), the match is dismissed. To the opposite, when a detection occurs while the GPS track indicates the taxi was near the detector (plus or minus 30 seconds which corresponds to the sampling rate of the GPS and within 830m around the detectors which correspond to a speed of 100km/h), then the *matching score* of that pairing is increased by one.
- Third, the proportion of the time during which data exist simultaneously for both the Bluetooth device and the Taxi meter, relative to the total time during which either the Bluetooth device has been detected or the taxi driver was logged on, is computed.
- Last, the final matching between a taxi and a Bluetooth device is chosen as the Bluetooth identifier with *relative matching score* (*matching score* divided by the total number of detection for that Bluetooth device) above 0.9 and with highest time proportion, providing that it has not been rejected.

This method enabled to identify over 100 pairings with *relative matching* score over 0.95. By considering sequences with running taximeters only (that is the taxi is occupied or hired), this leads to a set of around 1000 trajectories for a day of data, on which the trajectory recovery methods that will be proposed next, in Chapter 4, Section 2, can be tested.

CHAPTER 4

Trajectories from Bluetooth Data

Contents

1 Retrieving Trips form Bluetooth Detections **71**
- 1.1 From Monthly Data to Trips 71
- 1.2 Application to the Brisbane Dataset 73

2 Retrieving Trajectories from Bluetooth Trips **74**
- 2.1 Intuitions for Reconstruction 74
- 2.2 A Spatially Constrained Shortest Path Algorithm 75

3 Application to Brisbane Case Study **76**
- 3.1 Simulated Trajectories and detections 77
- 3.2 Real Trajectories: Taxi Dataset 80
- 3.3 Conclusion of the case studies 82

4 Trajectories for Further Analysis of the Bluetooth Data **82**
- 4.1 Speed Distribution 82
- 4.2 Discriminating Mode of Travel 82
- 4.3 Missed Detections: Binomial and Gaussian Mixture Models 84
- 4.4 Combining Trajectories with other Datasets 92

Build upon works published in:

- **Section 4 & 5** in G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving dynamic origin-destination matrices from Bluetooth data", in *Transportation Research Board, 93rd Annual Meeting*, Washington DC, Jan. 12–16, 2014. [Online]. Available: http://eprints.qut.edu.au/66511/
- **Section 4** in G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving trip information from a discrete detectors network: The case of Brisbane Bluetooth detectors", in *CAITR*, Sydney, Feb. 17–18, 2014. [Online]. Available: http://eprints.qut.edu.au/83110/
- **Section 3** in G. Michau, A. Nantes, A. Bhaskar, E. Chung, P. Borgnat, and P. Abry, "Bluetooth data in urban context: Retrieving vehicles trajectories", *Submitted in IEEE Transaction on Intelligent Transport Systems*, 2016

THIS chapter explores the opportunity of using vehicle identification systems to recover vehicle trajectories, that is, the sequence of road segments followed by a same vehicle from its origin to its destination. Vehicle trajectories are very valuable transport information as many other features can be inferred directly from their knowledge: OD matrix, travel time, speed, density, congestion. Less directly, trajectories can be used as inputs for route choice model calibration [176] or, in the case of this work, for the estimation of link dependent origin destination matrix.

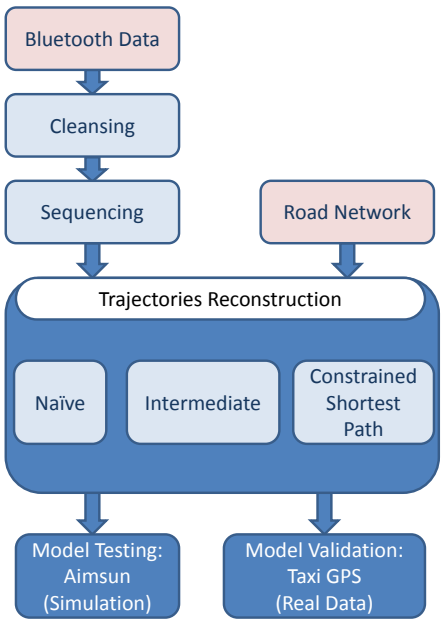


Figure 4.1: Framework for trajectory recovery: Inputs of this process are the Bluetooth data and the corresponding road network. First, the data are cleansed. Second, it is sequenced so that origin-destination and intermediate detections can be identified. Third, from the sequences, the trajectories are reconstructed. Last, the methods are tested and validated on two case studies.

We propose here a process, sketched in Figure 4.1, which does not require the prior knowledge of possible routes between OD pairs to recover user trajectories, to the opposite of previous works [106], [107], [129], [149]. The process of recovering trajectories is split into two parts: First, Section 1 proposes a procedure to recover the Bluetooth origin-destination matrix: Monthly data are divided into sequences representing trips. Second, Section 2 introduces an original algorithm for the computation of spatially constrained shortest paths. In Section 3, the proposed method is tested with a simulated case study and validated with the set of Taxi GPS tracks for which matches with a Bluetooth device have been identified. Finally, in Section 4, the trajectories are used for further characterising of the Bluetooth data, in the continuity of Chapter 3, Section 3.

1 Retrieving Trips form Bluetooth Detections

1.1 From Monthly Data to Trips

The first question to be addressed is the identification and discrimination of the trips within the monthly sequence of detections. For each detected Bluetooth device, the daily sequence of detections is divided into smaller sequences, corresponding to trips. This sequencing is performed with the four following steps:

- First, monthly data are ordered by MAC ID, then chronologically.
- Second, the shared MAC IDs are identified and filtered. As presented in Chapter 3, Section 3.2.6, MAC IDs with an aerial speed exceeding 130 kph, twice a day in average, are

suspicious and thus removed. When computing the aerial speed however, overlapping scanning zones are not taken into account (*cf.* Chapter 3, Section 3.2.2).

- Third, within one sequence, pairs of successive detections occurring at the same scanner are analysed. It appears likely that a user would start a new trip from where it arrived last. As illustrated in Figure 4.2, time intervals between successive detections at the same scanner are hardly ever under 10 minutes (almost a negligible number). There is however a strong gap at 10 minutes, making thus this value a good candidate for a threshold. If this threshold is exceeded, it is assumed that the first detection of the pair is the destination of a trip and the second detection the beginning of a new one. If not, the latest detection is dismissed and the duration field of the first detection is updated accordingly, as the difference between, the timestamps added from the duration of the last detection minus the timestamps of the first.

- Fourth, when successive detections do not occur at the same scanner, the time interval is compared to another threshold Δ. If this second threshold is exceeded, this is considered as the indication of two separate trips (similarly as above), otherwise it is considered as part of the same trip.

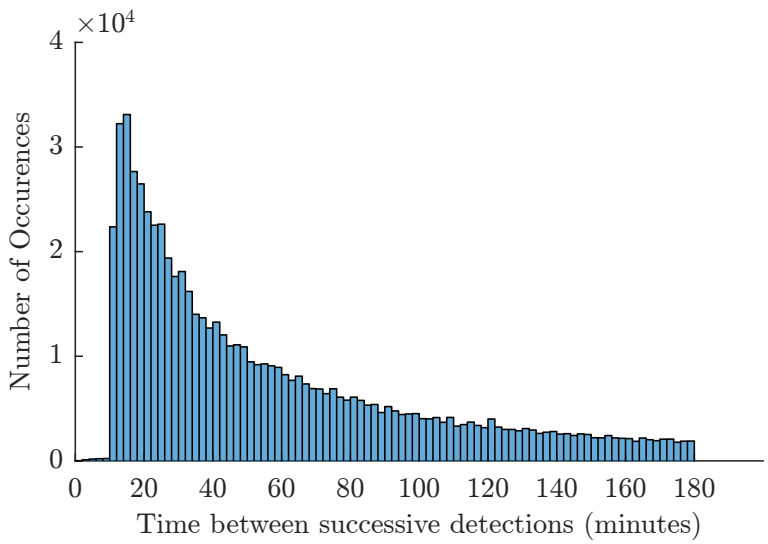


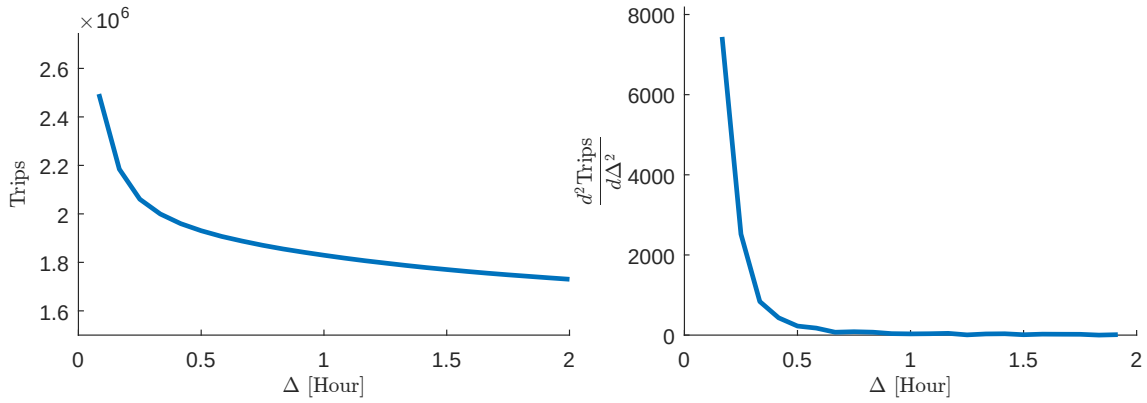
Figure 4.2: Distribution of the time intervals in-between pairs of successive detections at the same scanner.

In other words, if $d_i = (s_i, t_i, \delta_i)_{i\in[1,n]}$ represents the sequence of detections (scanner, timestamp, duration) for a device, we apply algorithm 4.1.

The outputs are m trips $(T_i)_{i\in[1,m]}$ for a device. To select Δ, the algorithm is applied to every device, for one month of data, for increasing values of Δ. The resulting number of trips is plotted in Figure 4.3a and its second derivative in Figure 4.3b. The number of trips evolves almost linearly for Δ varying between 30 min and 2 hours (the second derivative is almost zero), and exponentially before. This change of behaviour at 30 min tends to support the choice of that duration as a threshold. Moreover, this choice of threshold is also supported by the fact that within the area covered by Bluetooth detectors, the distance between adjacent detectors is at most a few kilometres. Thus, it is unlikely that a moving motorised vehicle would not be detected within 30 minutes. If so, then the amount of lost information within those 30 minutes would be such that recovering it would need very strong

Algorithm 4.1 Bluetooth Data to Trip Sequencing Algorithm

```
m ← 1
T_m ← d_1
for i = [1 : n − 1] do
    if (s_{i+1} = s_i and (t_{i+1} − t_i) > 10) or (s_{i+1} ≠ s_i and (t_{i+1} − t_i) > Δ) then
        T_{m+1} ← d_{i+1}
        m ← m + 1
    else
        if s_{i+1} = s_i then
            T_m(end) = (s_i, t_i, (t_{i+1} + δ_{i+1} − t_i))
        else
            T_m = [T_m ; d_{i+1}]
```



(a) (b)

Figure 4.3: (a) Number of trips resulting from the algorithm against sequencing threshold Δ. (b) Second derivative of the function

assumptions with little reliability. For their part, non-motorised vehicles are not within the primary focus of this work. This threshold is moreover consistent with the threshold proposed in [129].

For each detected device, the origin and destination resulting of this sequencing are always the first and last detections. Hence, some information is missing about the parts of the trip occurring before and after those detections, especially if some detections were missed. This approximation is accepted on the basis that recovering those missing parts would need a significant amount of assumptions and modelling, jeopardizing the confidence in the recovered data. Moreover, Wang, Hunter, Bayen, Schechtner, and González (2012) [177] have formalised the concept of detector-to-detector travel information as *transient OD matrix* and demonstrated their usefulness for transport applications [125]. Finally, if the aim is to estimate zone to zone information (traditional OD matrices, or route flows between each OD) or travel information, the aggregation over each zone of the information will mitigate this effect.

1.2 Application to the Brisbane Dataset

When performing the above sequencing for one week of data, from the 25th to the 31st of October 2014, achieved results read as follows:

- 1.2% of detections are from shared MAC IDs (0.1% of the MAC ID set).
- 0.2% of detections occur for devices detected only once (12.6% of the MAC ID set).
- 3.7% of detections are isolated (sequences made out of one detection only).
- 0.003% of detections are in sequences were all detections occurred at the same detector.
- 94.9% of detections belong to sequences that are considered as trips.

A this stage, the set of detections representing trips, can directly give an estimate of the Bluetooth OD matrix. From these data, if one considers the first detection as representative of the origin (e.g., as part of a zone) and similarly for the last detection interpreted as a destination, then the trip table becomes an OD matrix only representative of the Bluetooth flow. If one assumes that the Bluetooth penetration is random then it can give a good sense of the relative volumes of the OD and therefore of their importance. A generalisation of these data to the total traffic can be done by a comparison to traffic counts (e.g., by measuring an average penetration rate over the network and then by multiplying the Bluetooth OD matrix by the inverse of this penetration rate).

Generalising to the total flow however, without involving extra information, can be hazardous as the penetration rate may vary with time and with other parameters (e.g., wealth). A solution is the inference of the total flows from both traffic counts and Bluetooth flows, this will be the matter of Chapter 5. Toward this objective, a first step is the estimation, from the above sequences of detections, of Bluetooth flows on the roads. A way to carry out this estimation is to reconstruct the actual trajectories, on the road network, from the Bluetooth information.

In the next section, a method to retrieve trajectories from the sequences of Bluetooth detections is presented, while in the next chapter, a method to combine such Bluetooth data with traffic counts to estimate link dependent origin destination matrix will be proposed.

2 Retrieving Trajectories from Bluetooth Trips

2.1 Intuitions for Reconstruction

It has already been demonstrated in Chapter 3, Section 3 that a detection event inherently contains a certain spatial and temporal uncertainty. It is safe however, to assume that a detection only occurs if a user is within the surrounding of the detector. As a consequence, a recovered trajectory should pass within a reasonable distance from all the scanners belonging to the same trip, ideally within the detection zone, of radius r. To the opposite, the scanners which did not detect the user do not constrain the trajectory to be reconstructed: they can either be missed detections or simply, they might not belong to the trajectory.

Thus the final trajectory has to satisfy the following properties:

- It should be a continuous set of roads from the network, *i.e.,* it should be a path in the corresponding graph.
- It should include all the (observed) scanners of the trip, preferably in chronological order, except for overlapping scanning areas.

The difficulty here lies in that the scanners are disjoint from the road network and might span multiple roads. In addition, chronlogical ordering of the detections can be unreliable for detectors with overlapping detection zones. Hence, for a given sequence of detections, several paths might satisfy the two above properties. We propose here to consider the shortest path satisfying these conditions. Shortest path appraoches are common in traffic engineering: for the users, it models the optimality of the path amongst the set of possibilities. It is used for trajectory planing within GPS guidance systems [178], [179] or for route choice analysis [180].

The problem at hand here fits within the larger family of constrained shortest path computation that have had a huge body of literature, especially concerning variations on the *travelling salesman* problem [181], [182]. Yet, the specific problem of shortest path constrained by areas in a graph has scarcely been treated. In [183] the authors present this problematic and propose an A* algorithm to solve this problem. Another technique is mentioned in one sentence, suggesting the precomputation of node-to-node paths before performing the paths queries (what we call in the following the *m* message)[183, Section 3.5]. This technique is developed here.

2.2 A Spatially Constrained Shortest Path Algorithm

Table 4.1: Notation Summary

Notation	**Meaning**
$\mathscr{G} = (V, L)$	Oriented graph representing the road network
$V = \{v_k\}_{k \in \lvert V \rvert}$	Set of nodes v_k (traffic intersections are nodes)
N_V	Number of elements in V ($N_V = \lvert V \rvert$)
L	Set of directed edges. An edge is a direct itinerary linking two nodes.
N_L	Number of elements in L ($N_L = \lvert L \rvert$)
$l(v_k, v_m)$	Edge in L linking $v_k \in V$ to $v_m \in V$
$w_{l(v_k,v_m)}$	Length of $l(v_k, v_m)$
r	Scanning radius of the Bluetooth detectors
$S = \{s_k\}_{k \in \lvert S \rvert}$	Set of scanners on the network
$\mathscr{M}_r^V$	The mapping from the space of scanners S to a space of nodes V
$\mathscr{M}_r^V(s) = V_{r,s}$	Set of nodes in V within r to scanner s
$\{\mathscr{M}_r^V\}_{s \in S}$	Set of nodes in V within r of any scanner in S
$p(v_k, v_m)$	Shortest path between nodes v_k and v_m
$d(v_k, v_m)$	Length of the shortest path from v_k to v_m
Π_u	Sequence of observed detections for user u
$\Pi(p)$	Sequence of scanners within r of any node in p
r_{sim}	Detection Radius used in the simulated case study

The road network, as derived in Chapter 3, Section 1, is represented as an oriented graph $\mathscr{G} = (V, L)$ where V is the the set of nodes and L is the set of directed weighted edges, each corresponding to a direct itinerary (or road) linking two nodes (*i.e.,* not going through another node in V). A weight w is associated to each edge, currently consisting in the length of the itinerary. In the future, it could be modified to take into account effective cost of the road segment (travel time, congestion, cost, length, average observed speed, etc...). Note that, with this description of the network, a node is not

necessarily an intersection of roads but an intersection is necessarily a node. Bluetooth scanners, represented by the set S are located on that network according to their actual position and are disjoint from both road segments and nodes.

Let us consider user u with its observed sequence of detections $\Pi_u = \{s_1, \ldots, s_i, \ldots, s_n\}$. Let $V_{r,s_i} = \mathscr{M}_r^V(s_i)$ be the set of vertices within r of the scanner s_i, where $\mathscr{M}_r^V$ represents a mapping from the space of scanners to the space of vertices. Let $p(v_i, V_{r,j})$ be the set shortest paths between a vertex v_i and all vertices in $V_{r,j}$ (e.g., computed with a Dijkstra algorithm). Thus $p(v_i, V_{r,j})$ is a list of shortest paths of size $|V_{r,j}|$. Similarly, we note $p(V_{r,i}, v_j)$ the shortest distance between all vertices in $V_{r,i}$ and a vertex v_j.

By construction of the shortest path, there is a recursive relationship between the shortest path from a vertex v_1 to a vertex v_2 and the shortest path from v_1 to v_3 passing through v_2. If $p(v_1, v_2, v_3)$ is the shortest path between v_1 and v_3 through v_2 then,

$$p(v_1, v_2, v_3) = [p(v_1, v_2)\ ;\ p(v_2, v_3)]. \tag{4.1}$$

If d is the sum over the weights w_l of the links forming the corresponding path (e.g., its length in this case) we have correspondingly,

$$d(v_1, v_2, v_3) = d(v_1, v_2) + d(v_2, v_3). \tag{4.2}$$

This can be generalised as:

$$\min_{\substack{v_i \in V_{r,i} \\ i \in [1,s]}} d(v_1, \ldots, v_s, V_{s+1}) = \min_{v_s \in V_{r,s}} \{ \min_{\substack{v_i \in V_{r,i} \\ i \in [1,s-1]}} d(v_1, \ldots, v_{s-1}, v_s) + d(v_s, V_{r,s+1}) \} \tag{4.3}$$

If we define the *m message* of size $|V_{r,s}|$ as

$$m_{1:s} = \text{Arg} \min_{\substack{v_i \in V_{r,i} \\ i \in [1,s-1]}} d(v_1, \ldots, v_{s-1}, V_{r,s}), \tag{4.4}$$

then, Equation (4.3) can be written recursively as

$$\forall i \in [1, |V_{r,s+1}|], \quad m_{1:s+1}(i) = \text{Arg} \min_{v_s \in V_{r,s}} \left\{ m_{1:s} + d(V_{r,s}, v_{s+1}^i) \right\}, \tag{4.5}$$

where v_{s+1}^i is the i-st vertices in $V_{r,s+1}$ and $d(V_{r,s}, v_{s+1}^i)$ the vector of size $|V_{r,s}|$ of shortest distances between all vertices in $V_{r,s}$ and the node v_{s+1}^i.

Thus, the algorithm runs forward along the sequence of scanners, computing the m message at each iteration using Equation (4.5). At the end, it will have the sequence of vertices $v_1, \ldots, v_S$ of the shortest path reaching each of the scanning areas. Figure 4.4 illustrates the steps of the algorithm.

We propose here to precompute all the shortest paths in between nodes in the neighbourhood of the detectors ($\{\mathscr{M}(s)\}_{s \in S}$). With a Dijkstra algorithm for example, this step has a $\mathscr{O}(|V|^2)$ complexity. The above described algorithm has then a complexity in $\mathscr{O}(|V|)$. To account for the fact that detectors might have overlapping detection areas, in which case the shortest path would not be defined, detectors with overlapping detection areas are combined into virtual detectors. The detection area of a virtual detector is the sum of the detector areas it is made of.

3 Application to Brisbane Case Study

To test and assess the validity of the proposed route recovery procedure, a set of data composed of both trajectories and detections is needed. However, the Bluetooth dataset in Brisbane only consists

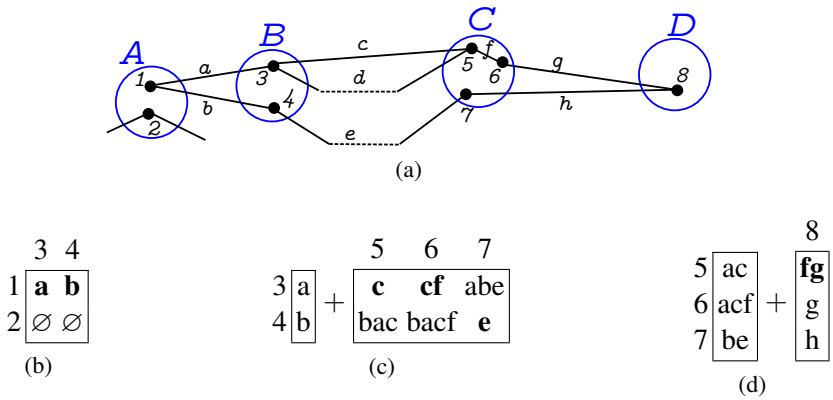


Figure 4.4: Example of a simple network (a) and illustration of the CSP method from scanning area A to D. (b) In the first step, the algorithm computes the shortest paths to scanning area B (composed of nodes 3 and 4), from area A. Bold paths are the shortest. (c) In a second step the algorithm computes the shortest path to area C (nodes 5, 6 and 7) from area B. The path of the first step are added. (d) At the end of the third step, destination D has been reached and the shortest path is $acfg$.

of detections. Thus, the ground truth against which recovered trajectories can be compared is not available. Therefore, we developed two case studies:

1. A simulated case study of Brisbane inner-city. The road network has been imported to AIMSUN[184], an agent-based simulation software, where individual vehicles are modelled and their position retrieved at each time step, giving thus a set of trajectories. Bluetooth detections are then simulated and the trajectory recovery process tested.

2. The process is further compared with a set of real-world data: the taxi dataset described in Section 4. By matching Taxi GPS tracks with corresponding Bluetooth devices, this permits to validate the model.

The spatially constrained shortest path ("CSP") method, proposed here in Section 2.2, is further compared to two simpler methods to answer three additional questions: Is it really necessary to consider intermediate detections? What would be the results with sparser detector networks? Do drivers follow the shortest path from origin to destination? The simpler methods are: the shortest path from first to last detection ignoring the others ("naïve"), and the shortest path considering a single intermediate detection in the middle of the sequence ("intermediate"). These two methods have similar complexity but are easier to implement.

3.1 Simulated Trajectories and detections

3.1.1 Simulated Trajectories

The Brisbane inner-city network has been imported to AIMSUN. It consists of a 10×8 km^2 area, with 8000 links and 440 linear kilometres (see Figure 4.5). The input origin-destination matrix has been calibrated with the *Brisbane Strategic Transport Model* results [185], in peak hours. The simulation is performed over one hour, times and positions of each vehicle are then exported every 0.8 second for analysis. The Bluetooth detectors from Brisbane City Council networks are positioned on the network. The inner city area, as used in this simulation, has 265 scanners.

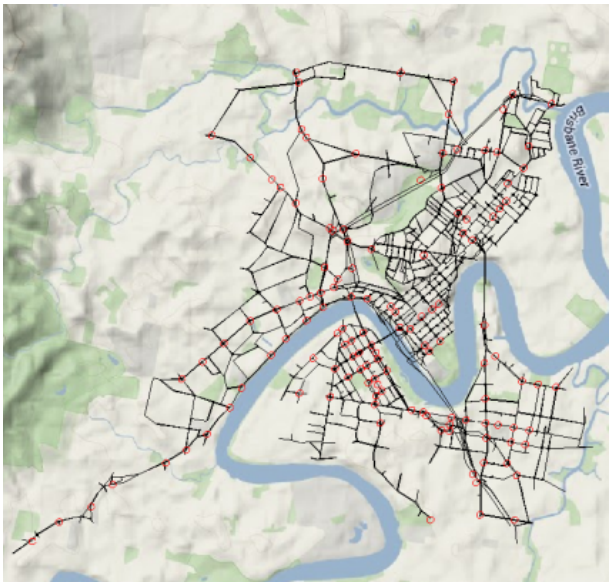


Figure 4.5: Brisbane Inner City Network 10x8 km^2 as input of the traffic model. The red circle are the 265 Bluetooth scanners positioned along the roads.

3.1.2 Simulated Detections

A set of randomly selected vehicles is extracted from the data. The sequence of detectors corresponding to their trajectory is computed as follow: Each detector is projected on the trajectory. Then detectors are ordered by distance between the origin and their projection. The distance between the detectors and their projection is used to compute the detection probability. Detections event are randomly drawn according to these probabilities. The three tested methods provide exactly the same results for trajectories composed of up to three detectors. Thus, vehicles with less than three detectors on their path are discarded. The test set is composed of 1000 vehicles.

The probability of detection chosen for this simulation is defined by two parameters:

- The rate of detection p, which defines the probability of a car to be detected by the detectors. For the simulation, p is varied from 0.4 to 0.9. Results however are presented only for $p = 0.75$ as they are representative of the whole.
- A detection radius r_{sim}, within which the detection rate is assumed to be constant and equal to p. Here r_{sim} has been explored from $10m$ to $150m$.

3.1.3 Results of the Simulated Case Study

To assess the efficiency of the proposed methods for recovering trajectories, we compare each recovered trajectory to the recoverable part of the original one, that is, between the first and the last detections. To measure the efficiency, an indicator, denoted here by *error*, is computed: for each trajectory, we compute the fraction of the length incorrectly recovered and then we average these fractions over the full set.

The results depend on the three parameters r, r_{sim} and p. r is the radius used in the recovery method, while r_{sim} and p are the parameters used to define the probability of detection in the simulation. Thus the recovery process is tested, first, for various values of r from 10m to 150m. We define,

for each simulation, the optimal r as the one leading to the smallest *error* and the least optimal r as the one leading to the largest *error*.

Second, the previous testing method is repeated while the parameters of the simulation vary, to test the robustness of the procedure. Indeed, in a real-world case study, as the one in Brisbane, the behaviour of the scanners is not perfectly known. Here p varies from 0.4 to 0.9 and r_{sim} from 10m to 150m.

Figure 4.6 shows the evolution of the *error* against the simulated detection radius r_{sim}. For each method, the lowest *error*, for the optimal r, is plotted in full line with its value written. The largest *error*, for the least optimal r, is plotted in dotted line. The area in-between represents the values the *error* can take for varying r.

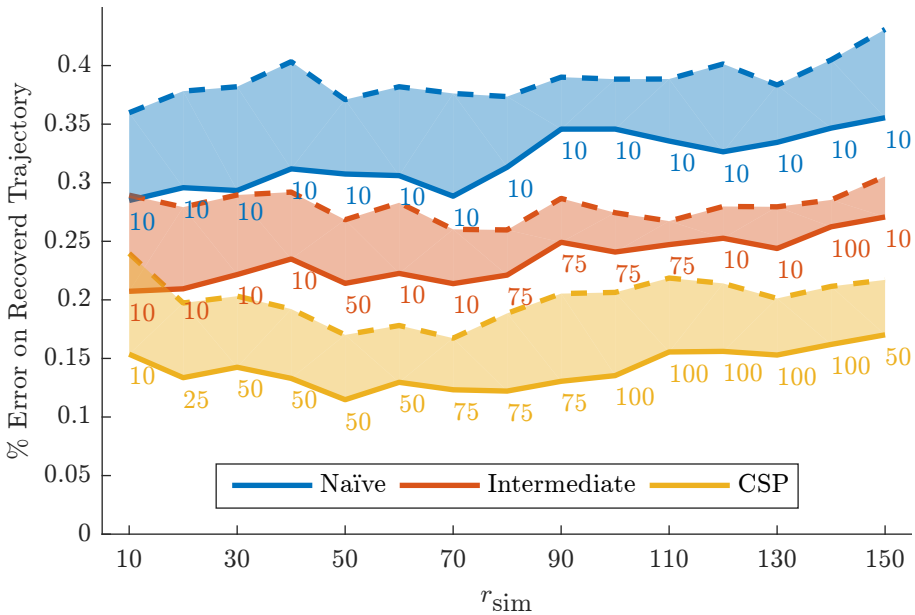


Figure 4.6: Mean error on the recovered trajectories for the three methods. For each method, its corresponding area is defined by the minimum (full lines) and maximum (dotted lines) values of the error for $r \in [10, 150]$. The optimal value of r is written for each r_{sim}.

Figure 4.6 indicates that the proposed method ("CSP") is the most efficient amongst the three with an *error* ranging between 11% and 17% for the optimal r. In addition, there is a good fit between the optimal value of r and the simulated detection radius r_{sim} except for the 150m value. It appears that in overall, to set the value of r to 150m, for the CSP method, is never optimal (*cf.* Table 4.2). It may create too much uncertainty on the position of the user for an efficient recovery.

A second observation is, that the efficiency of the reconstruction does not significantly depend on the chosen value of r, as long as it is not too far from the simulated detection radius r_{sim}. Even so, the worst case is *only* a 24% error.

Last, it has to be remembered that those *errors* are computed only on recoverable parts of the trajectories (between first and last detections). If we take into account the full trajectories, the error varies between around 30% for the optimal values of r to around 55% in the worst cases, as shown in Table 4.2. In this table, (*Err Traj*) represents the mean error on the full trajectory, (*Err Traj Reco*) the mean error on the recoverable part of the trajectory and (*Inc Detect*) the mean proportion of observed detectors not within r of the recovered trajectories (Inconsistent Detectors). This measure is not applicable for the CSP method, as by construction, the recovered route passes through every

observed detector.

This simulation did not aim to reproduce perfectly the behaviour of Bluetooth scanners. It could be extended in the future to take into account more parameters when defining the probability of detection (e.g., a more complex detection probability distribution involving the time spent by the vehicles in the detection zones and their speed or the simulation of an inquiry cycle for the Bluetooth connection process).

Table 4.2: Results for the three methods

r_{sim}	Method	Solution	r	Err Traj	Err Traj Reco	Inc Detect
10	Naïve	Best	10	60.8%	28.5%	13.4%
		Worst	150	64.8%	36.0%	12.5%
	Interm.	Best	10	55.5%	20.7%	5.1%
		Worst	150	59.8%	28.9%	5.2%
	CSP	Best	10	51.8%	15.4%	N.A.
		Worst	150	56.4%	24.0%	
50	Naïve	Best	10	63.3%	30.7%	15.4%
		Worst	150	52.3%	37.1%	19.3%
	Interm.	Best	50	40.0%	21.4%	9.2%
		Worst	150	43.9%	26.8%	9.2%
	CSP	Best	50	31.8%	11.5%	N.A.
		Worst	150	35.6%	17.0%	
100	Naïve	Best	10	63.9%	34.6%	19.1%
		Worst	150	50.8%	38.8%	20.6%
	Interm.	Best	75	39.2%	24.1%	10.9%
		Worst	150	41.5%	27.4%	9.5%
	CSP	Best	100	30.1%	13.5%	N.A.
		Worst	10	54.0%	20.6%	
150	Naïve	Best	10	63.3%	35.5%	20.0%
		Worst	150	52.9%	43.1%	22.7%
	Interm.	Best	10	58.0%	27.1%	9.7%
		Worst	150	42.5%	30.5%	10.7%
	CSP	Best	50	31.8%	17.0%	N.A.
		Worst	10	54.0%	21.7%	

3.2 Real Trajectories: Taxi Dataset

To validate the present model on real data, a set of real trajectories with their corresponding Bluetooth detections needs to be available. Such a database is not directly available for the Brisbane Bluetooth data but we inferred one, using the GPS tracks of the *Black & White Cabs* matched with Bluetooth devices as presented in Chapter 3, Section 4.

Similarly to the simulated case study in Section 3.1, the efficiency of each strategy is assessed through the same indicators. In this case however the fraction of the trajectory that has not been correctly reconstructed is the fraction of GPS points farther than 10m to the reconstructed trajectory. The results are presented in Figure 4.7 and Table 4.3.

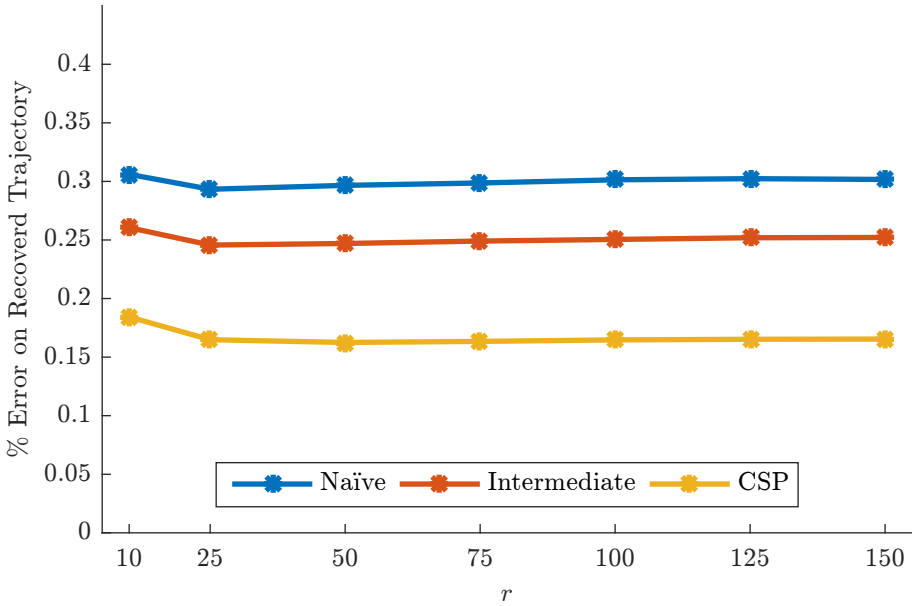


Figure 4.7: Mean error on the recovered Taxi trajectories for the three methods against r.

The proposed CSP method is the most efficient with a mean error of 16.3% when $r = 50m$. Again, the results do not depend significantly on the chosen value for r, unless it is taken very small ($10m$). The error on the full trajectory, however, is at best 55%, a much larger value than in the simulated case study. This has two possible causes: first the network is much broader (the whole city of Brisbane), thus with a smaller density of Bluetooth detectors than the centre. Second, the taxi might pick-up and drop-off passengers in residential streets that are not covered by Bluetooth detectors and not represented in the graph representing the road network.

As a final remark, the measure of inconsistent detectors with the recovered trajectories is not exactly 0% for the CSP method as for few sequences no trajectory passing through every observed scanning area existed.

Table 4.3: Results for the Taxi Case Study

Method	**Solution**	**r**	**Err Traj**	**Err Traj Reco**	**Inc Detect**
Naïve	best	25	72.0%	30.9%	28.6%
	worst	150	73.0%	31.7%	44.6%
Interm.	best	25	66.4%	25.6%	19.1%
	worst	150	67.5%	27.1%	23.4%
CSP	best	50	56.4%	**16.3%**	0.01%
	worst	10	58.3%	18.6%	0.01%

3.3 Conclusion of the case studies

These two case studies lead to conclude that Bluetooth detector networks, as the one in Brisbane, can be used to recover trajectory information from individuals. We developed an algorithm intended for vehicle routing with stopover areas and demonstrated that it is also adapted for trajectory retrieval from detection sequences with spatial and temporal uncertainties. Up to 84% of the recoverable part of the trajectories is correctly reconstructed. By comparing with methods taking less detections into account, we showed moreover that satisfying results could be obtained with sparser detector networks: with only two detections trajectories are recovered at 70%. Yet, a sparser network also means a larger portion of the trajectory, before and after the first and last detections, that will not be retrieved. Last, by comparing the proposed CSP method with the classic shortest path approach, we proved that the shortest path does not always correspond to that chosen by the users. It confirms that trajectory analysis in urban context can not be reduced to a simple shortest path problem.

4 Trajectories for Further Analysis of the Bluetooth Data

In the previous section, it has been shown that the actual trajectories of Bluetooth equipped vehicles can be retrieved with satisfying precision. Using these results, additional characteristics of the Bluetooth data can be inferred in the continuity of those presented in Chapter 3, Section 3.

First, one can get a realistic speed distribution. This is presented in Section 4.1. In Section 4.2, using those speed estimations, we show that modes of travel can be identified. Last, in Section 4.3, by comparing the observed detections, to the detectors along each trajectory, the missed detection rate is further analysed.

Furthermore, the trajectories can be involved in the estimation of the total traffic flows in Brisbane (OD matrix estimation and LOD matrix estimation in particular); this will be the heart of Chapter 5 from a theoretical perspective and of Chapter 6 for the Brisbane case study.

4.1 Speed Distribution

The distribution of the trajectory average speed is shown in Figure 4.8. This distribution has a peak at around 38 kph, an average of 40 kph, and a median of 37 kph. Those values are in accordance with observation [175] and with the speed distribution in Figure 3.10b, Chapter 3, Section 3.2.4, where the speeds were based on aerial distance.

The complete set of trajectories contains data from Bluetooth devices, regardless of the mode of travel of the users. This could be explained why this distribution looks skewed toward lower speeds, compared, for example, to a Gaussian distribution.

In the next section, we show that this distribution can in fact be interpreted as the sum of three distributions.

4.2 Discriminating Mode of Travel

From a transport engineering perspective, three major behaviours can be expected:

1. First, trips from non-motorised modes of travel are expected. These *non-motorised trips* are characterised by low speeds and short length.

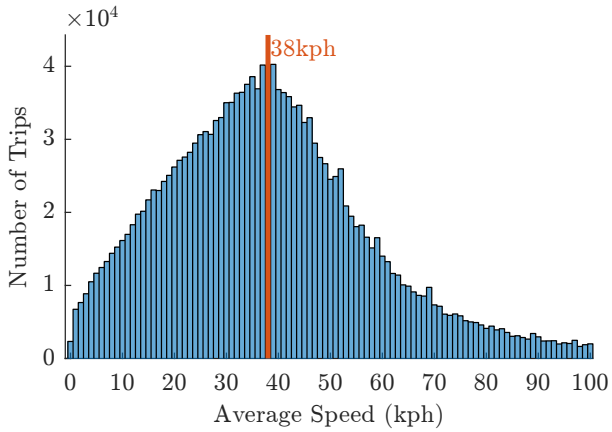


Figure 4.8: Distribution of the average speed of the recovered trajectories. The mode is at 38 kph, the average at 40kph and the median at 37 kph, roughly in accordance with what could have been observed in Brisbane [175] (Data: 25th to 31st of October 2014).

2. Second, trips from motorised mode of travel with a single defined objective (commuting for example). These *direct trips* are characterised by higher speed and average length.

3. Third, trips also from motorised mode of travel but composed of several sub-trips, (for example errands or users working in their cars, e.g., taxi and postmen). These *errand trips* can be identified with their speed dropping occasionally and they tend to have longer length.

Thus, we propose first, to filter for *non-motorised trips*. To do so, the speed in-between successive detections is computed for pairs of detectors farther than 300m one from the other (*cf.* Chapter 3, Section 3.2.2). Then, given a speed threshold, if every computed speed of the trajectory is below that threshold it will be considered as belonging to the non-motorised set. To choose this threshold, similarly to Section 1.1, we perform the algorithm for increasing value of the threshold and characterise the results.

Table 4.4 presents, for increasing threshold speed (from 2 to 25 kph), the percentage of trip filtered, the length of the longest trip filtered and the percentage of trips shorter than several distances from 2 to 14 km. These results have been obtained with trajectories retrieved from the Bluetooth data from the 25th to the 31st of October 2015 but are consistent with those obtained at other time, in particular those presented in [19] with data from 250 Bluetooth detectors in 2013.

Table 4.4: Non-Motorised Mode: Trajectory Lengths versus Threshold Speed

Thrshld (kph)	**% of Trips**	**Longest (km)**	**<2km (%)**	**<4km (%)**	**<6km (%)**	**<8km (%)**	**<10km (%)**	**<12km (%)**	**<14km (%)**
2	0.15	5	99.3	100.0	100.0	100.0	100.0	100.0	100.0
4	0.38	7	97.0	99.9	100.0	100.0	100.0	100.0	100.0
6	0.64	7	93.3	99.8	100.0	100.0	100.0	100.0	100.0
8	0.88	7	90.4	99.5	100.0	100.0	100.0	100.0	100.0
10	1.14	10	87.4	98.3	99.8	100.0	100.0	100.0	100.0
15	1.86	16	81.2	95.2	98.9	99.7	100.0	100.0	100.0
20	2.82	36	76.2	92.4	97.5	99.1	99.8	99.9	100.0
25	4.03	36	70.6	89.2	95.7	98.2	99.3	99.7	99.9

From these results, it appears that choosing a speed of 15kph as a threshold is a sound choice to catch trajectories corresponding to non-motorised modes. Indeed, those trajectories represent less

than 2% of the global set, a value consistent with statistics on Bluetooth equipment proportion among pedestrians and public transport passengers in [135]. In addition these sequences are very short: 95% of these trajectories are shorter than 4km and 99% shorter than 6km.

In a second step, in order to discriminate *errand trips* from the remaining set of trajectories, speeds in-between pair of successive detections are computed. Trajectories with at least one speed lower than 6kph (walking speed) is considered as belonging to the *errand trips* set. The remaining trajectories are considered as *direct trips*. The speed distributions of those three sets are represented in Figure 4.9.

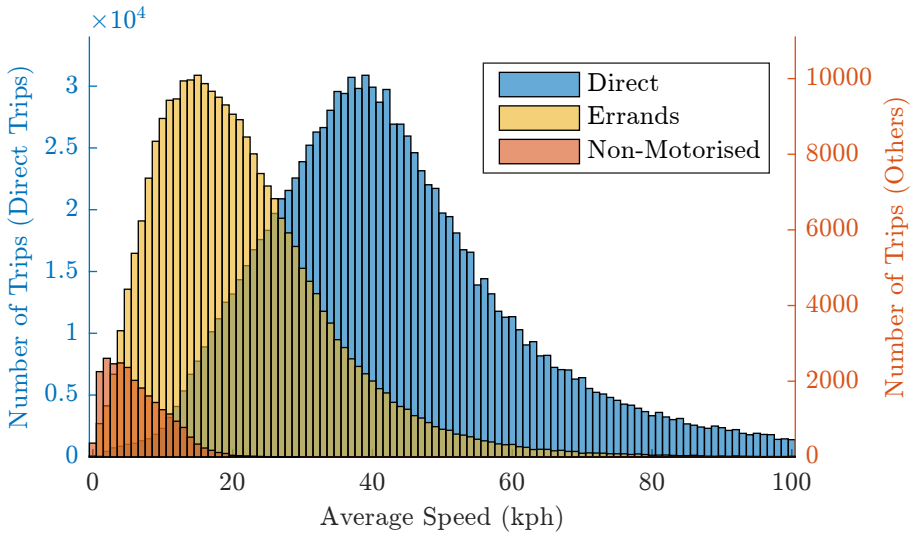


Figure 4.9: Distribution of the average speed of the recovered trajectories for the three sets (Data: 25th to 31st of October 2014).

Eventually, for the Brisbane dataset (25th to 31st of October 2014), 94.9% of detection are considered as trips and trips are separated as follow:

- 18% of the trips are made of only two detections.
- 1.5% are in the *non-motorised trips* set. The median speed is of 5.5 kph, the average length 1.3 km.
- 14.8% are in the *errands trips* set. The median speed is 20 kph, the average length 14 km.
- 65.7% are in the *direct trips* set. The median speed is 41 kph, the average length 9 km.

4.3 Missed Detections: Binomial and Gaussian Mixture Models

4.3.1 Bluetooth equipped Vehicles Detection Probability Distribution

From recovered trajectories, it is possible to realise a deeper analysis of missed detections by comparing the observed detections to the set of detectors along a trajectory. However, to the opposite of the corridor analysis (Chapter 3, Section 3.3) each trajectory can involve a different number of detectors. Let $X = (G,H) = (g_i, h_i)_{(i\in[1:N])}$ be the set of N recovered trajectories with g_i the number of observed detections for the i-th Bluetooth device and h_i the number of detectors along its trajectory. We are

looking here at trajectories recovered from first and last detections. Hence the detection probability are conditioned by these two detections and the quantities of interest are therefore $(g_i - 2)$ and $(h_i - 2)$.

When computing the detection ratio of the i-th car as $\big((g_i - 2)/(h_i - 2)\big)$, it will necessarily be a multiple of $\big(1/(h_i - 2)\big)$. Hence, in order to compare the detection ratios between devices with a precision of ε (e.g., $\varepsilon = 0.1$), every device needs to verify $(h_i - 2) > (1/\varepsilon)$ (e.g., 12 detectors at least).

As a consequence, for computing the histogram of the detection ratios as in Figure 4.10, and to avoid statistical artefacts, the width of the histogram has to be of at least $(\min_i 1/(h_i - 2))$. In Figure 4.10, four of those histograms are represented. On each of these histograms, trajectories with less detectors than the inverse of the bin width have been dismissed.

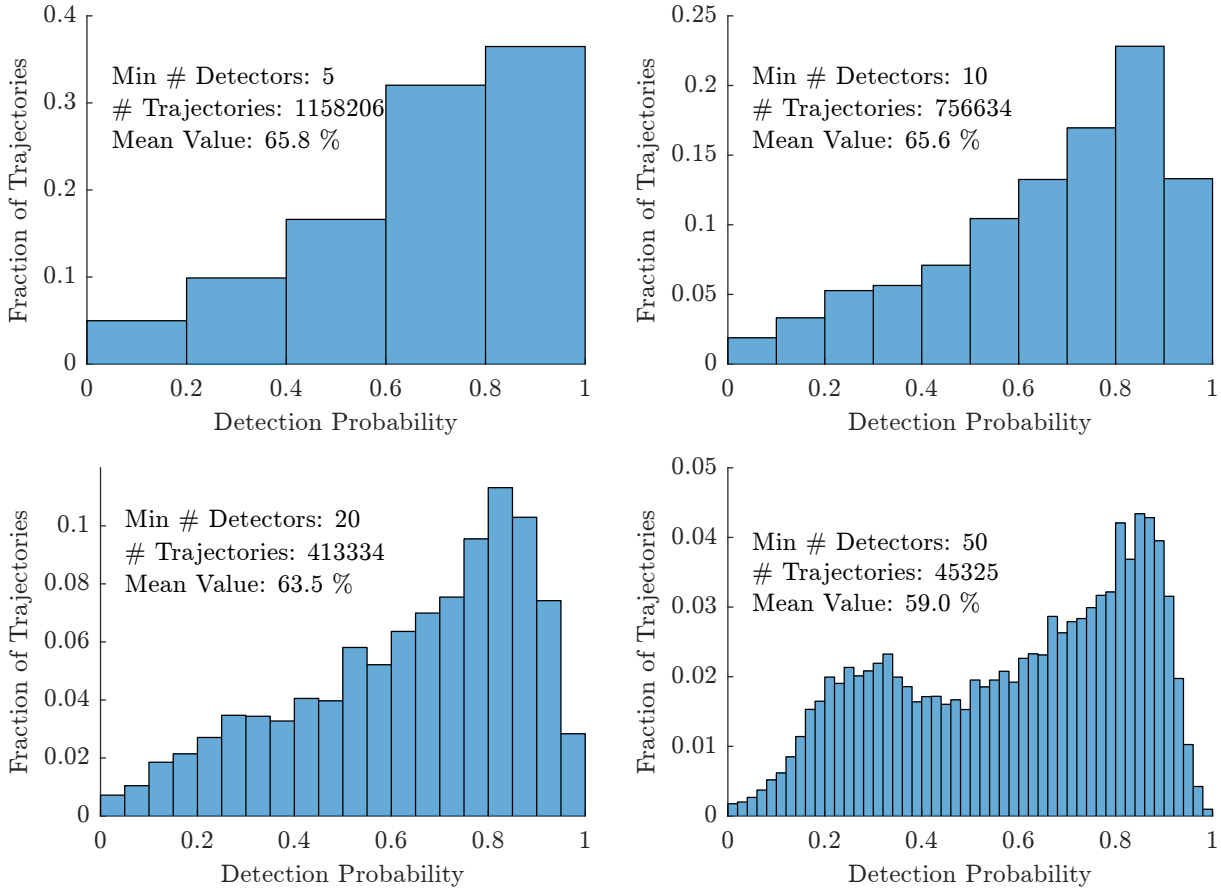


Figure 4.10: Histogram of the Bluetooth device detection ratios for increasing minimum value of $(\min_i(h_i - 2))$: (a) $\min_i(h_i - 2) = 5$ (b) $\min_i(h_i - 2) = 10$ (c) $\min_i(h_i - 2) = 20$ (d) $\min_i(h_i - 2) = 50$.(Data: 25th to 31st of October 2014).

Interestingly, on these histograms, once enough detectors are taken into account, two modes appear, one around 30% and one around 80%. While the reasons of these two modes can only be speculated, further characterising this distribution, for example as a mixture of two distributions is the matter of the next section.

Beforehand, for each of these histograms (Figure 4.10), a global detection ratio can be computed and is represented in Figure 4.11, for continuously increasing value of $(\min_i h_i)$ from 3 to 100.

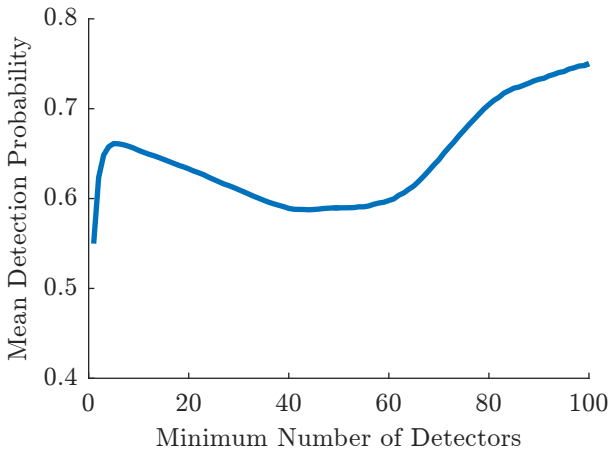


Figure 4.11: Average detection ratio computed over Bluetooth devices i with $h_i > x$ where x is the value read in abscissa. (Data: 25th to 31st of October 2014).

Figure 4.11 illustrates that the value of $(\min_i h_i)$ has an importance when computing the detection ratio. For very large number of detectors taken into account, the statistic is biased toward Bluetooth devices with higher detection probability. In fact, Bluetooth devices with lower detection probability have a higher chance of not being detected in the time interval corresponding to the threshold at which sequences are separated into different trips. Thus their trajectories will appear shorter in average. To the opposite, for a low number of detectors taken into account, two effects affect the global detection ratio. First, the Bluetooth devices had to be detected twice (at origin and destination) thus biasing the detection ratio toward higher probabilities, effect that is mitigated as the number of detectors on the trajectory increases. Second, as the length of the considered trajectories increases, the number of devices with low/very low detection probability diminishes, thus, the value of the ratio increases. Hence the observed horizontal S-shape of the curve in Figure 4.11.

4.3.2 Expectation Maximisation (EM) Algorithm for Mixture of two Distribution

The histograms, as plotted in Figure 4.10, led us to conjecture the existence of two classes of Bluetooth devices. One way to characterise these two classes is by means of a distribution mixture. In the following we propose to use the EM algorithm to fit, first a Binomial mixture and second a Gaussian mixture to the Bluetooth data.

The EM algorithm has been developed by Dempster, Laird, and Rubin (1977) [186] in order to tackle estimation problems, where the underlying model involves unobserved variables. In this case, we assume that the Bluetooth devices belong to two classes with different detection probabilities. The unobserved variable is thus the distribution of the devices among each class while the average detection probability of each class are the parameters to be estimated. The underlying idea of the EM algorithm is to compute the expected log likelihood of the model with guessed parameters for all possible affectations of the observations to the two classes. Then, this expectation is maximised with respect to the parameters of the model in order to get a more precise estimation. It is thus an iterative algorithm.

This algorithm is particularly adapted for this kind of mixture problem when the mixture itself is seldom observed, as in [187], [188].

The theory behind the EM algorithm is recalled in Appendix B and we will present here the results for three different mixtures: Binomial mixture, Gaussian mixture and Gaussian mixture with the means inferred from the binomial mixture.

Last, we demonstrate that the mixtures also apply to the corridors chosen in Chapter 3, Section 3.3, under the condition that the coefficient mixture, derived by the algorithm, is corrected to account for the sampling bias.

The EM algorithm for the Binomial Mixture Model

Let us consider a set of N observations X defined as $X = (G,H) = (g_i,h_i)_{(i\in[1:N])}$ where g_i is the number of detections which occurred and h_i the number of detectors on the trajectory respectively.

Then, the binomial mixture for X read as:

$$(\forall i \in [1,N]), \tag{4.6}$$
$$BM(g_i|h_i,\pi_1,\theta_1,\theta_2) = \pi_1 \binom{h_i-2}{g_i-2} \theta_1^{g_i-2}(1-\theta_1)^{(h_i-g_i)} + (1-\pi_1)\binom{h_i-2}{g_i-2}\theta_2^{g_i-2}(1-\theta_2)^{(h_i-g_i)}$$

where π_1 is the mixture coefficient, θ_1 and θ_2 the success probabilities of the two distributions.

From the theoretical results recalled in Appendix B, the EM algorithm for the binomial mixture can be designed as in Algorithm 4.2. Let us define $MaxIter$, an upper bound on the iteration number at which the algorithm should stop if it did not reach convergence, Tol, a threshold on the iterates below which the convergence of the algorithm is assumed, and an initial guess of Π and Θ, ($\pi_1^0, \theta_1^0, \theta_2^0$). Then, we propose Algorithm 4.2 as the EM algorithm applied to the Binomial Mixture Model. In this algorithm, the power functions are applied element-wise.

Algorithm 4.2 EM Algorithm for the Binomial Mixture Model

Require: $\theta_1^0,\ \theta_2^0,\ \pi_1^0,\ MaxIter,\ Tol, X = (G,H) = (g_i,h_i)_{(i\in[1:N])}$
$k \leftarrow 0$
while $k < MaxIter$ **and** $(k > 0$ **and** $(|\pi_1^k - \pi_1^{k-1}| > Tol$ **or** $|\theta_1^k - \theta_1^{k-1}| > Tol$ **or** $|\theta_2^k - \theta_2^{k-1}| > Tol))$ **do**
  $P_1 = \pi_1^k \cdot (\theta_1^k)^{G-2} \circ (1-\theta_1^k)^{(H-G)}$
  $P_2 = (1-\pi_1^k) \cdot (\theta_2^k)^{G-2} \circ (1-\theta_2^k)^{(H-G)}$
  $E = P_1./(P_1+P_2)$
  $k \leftarrow k+1$
  $\pi_1^k = \frac{1}{N}\sum E$
  $\theta_1^k = \frac{\sum(E\circ G)}{\sum(E\circ H)}$
  $\theta_2^k = \frac{\sum((1-E)\circ G)}{\sum((1-E)\circ H)}$
return $\pi_1^k, \theta_1^k, \theta_2^k$

The EM algorithm for the Gaussian Mixture Model

In case of the Gaussian Mixture model, let us define now X as $X = (x_i)_{(i\in[1:N])} = ((g_i-2)/(h_i-2))_{(i\in[1:N])}$, a set of N observations then the mixture can be written as:

$$(\forall i \in [1,N]), \quad GM(x_i|\pi_1,\mu_1,\sigma_1,\mu_2,\sigma_2) = \frac{\pi_1}{\sigma_1\sqrt{2\pi}}\exp^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + \frac{(1-\pi_1)}{\sigma_2\sqrt{2\pi}}\exp^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}} \tag{4.7}$$

Similarly to Algorithm 4.2, defining $MaxIter$, Tol, and an initial guess of Π and Θ, ($\pi_1^0, \mu_1^0, \mu_2^0, \sigma_1^0, \sigma_2^0$), we propose Algorithm 4.3 as the EM algorithm applied to the Gaussian Mixture Model.

Algorithm 4.3 EM Algorithm for the Gaussian Mixture Model

Require: $\mu_1^0,\ \mu_2^0,\ \sigma_1^0,\ \sigma_2^0,\ \pi_1^0,\ MaxIter,\ Tol, X = (x_i)_{(i\in[1:N])}$
$k \leftarrow 0$
while $k < MaxIter$ **and** $(k > 0$ **and** $(|\pi_1^k - \pi_1^{k-1}| > Tol$ **or** $|\mu_1^k - \mu_1^{k-1}| > Tol$ **or** $|\mu_2^k - \mu_2^{k-1}| > Tol$ **or** $|\sigma_1^k - \sigma_1^{k-1}| > Tol$ **or** $|\sigma_2^k - \sigma_2^{k-1}| > Tol))$ **do**
$P_1 = \pi_1^k \cdot \frac{1}{\sigma_1^k\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu_1^k)^2}{2(\sigma_1^k)^2}}$
$P_2 = (1-\pi_1^k) \cdot \frac{1}{\sigma_2^k\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu_2^k)^2}{2(\sigma_2^k)^2}}$
$E = P_1./(P_1+P_2)$
$k \leftarrow k+1$
$\pi_1^k = \frac{1}{N}\sum E$
$\mu_1^k = \frac{\sum(E\circ X)}{\sum E}$
$\mu_2^k = \frac{\sum((1-E)\circ X)}{\sum(1-E)}$
$\sigma_1^k = \sqrt{\frac{\sum(E\circ(X-\mu_1^k)^2)}{\sum E}}$
$\sigma_2^k = \sqrt{\frac{\sum((1-E)\circ(X-\mu_2^k)^2)}{\sum(1-E)}}$
return $\pi_1^k, \mu_1^k, \mu_2^k, \sigma_1^k, \sigma_2^k$

4.3.3 Binomial Mixture Model on Retrieved Trajectories

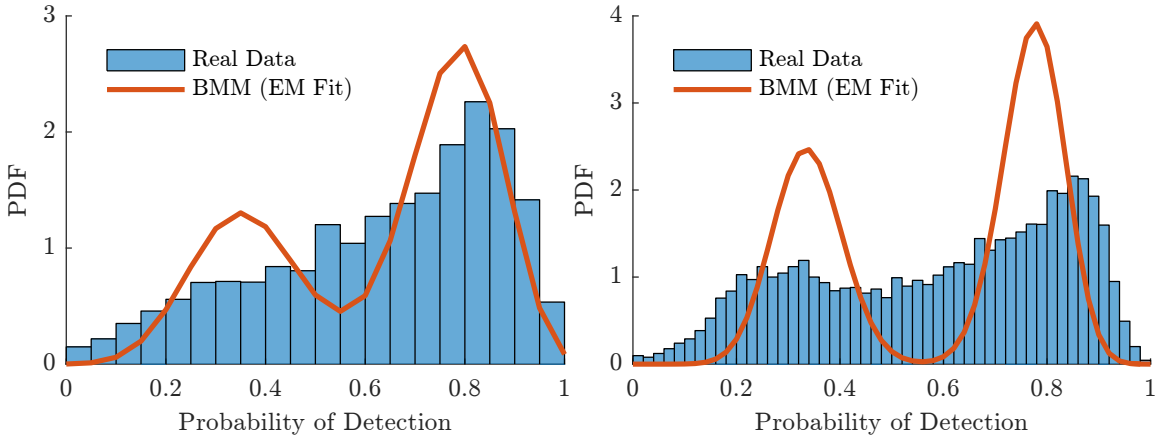
To start with, we applied Algorithm 4.2 to the trajectories retrieved from the Brisbane dataset. As presented in Section 4.3.1, the two modes distribution only arises when trajectories with enough detectors on their path are selected. Thus, we applied the algorithm twice, first on trajectories with at least 20 detectors $\big(min_i(h_i-2) = 20\big)$ with the above notations and second on trajectories with at least 50 detectors $\big(min_i(h_i-2) = 50\big)$. Results are presented in Figure 4.12. In these figures, the histograms, similarly to those in Section 4.3.1 represents the Probability Density Function of the detection ratio among the trajectories (histogram of $(g_i-2)/(h_i-2)$). The PDF corresponding to the binomial mixture model is evaluated for each bin and the x-axis is renormalised between 0 and 1. That is:

$$\begin{aligned}
&\forall k \in [0, Nmax] \\
&\quad \mathrm{pdf}_{BMM}(k/Nmax) = \quad \tilde{\pi}_1 \binom{Nmax}{k} (\tilde{\theta}_1)^k (1-\tilde{\theta}_1)^{(Nmax-k)} \\
&\qquad\qquad + (1-\tilde{\pi}_1) \binom{Nmax}{k} (\tilde{\theta}_2)^k (1-\tilde{\theta}_2)^{(Nmax-k)}
\end{aligned} \tag{4.8}$$

where $Nmax = min_i(h_i-2)$, corresponding to the number of bins, *i.e.,* 20 in Figure 4.12a and 50 in Figure 4.12b.

The adequacy of the models to the data is evaluated through the Normalised Negative Log-Likelihood (NNLLh) computed as:

$$NNLLh = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\mathrm{pdf}_{BMM}(\frac{g_i-2}{h_i-2})}{Nmax}\right) \tag{4.9}$$



(a) (b)

Figure	$\tilde{\pi}_1$	$\tilde{\theta}_1$	$\tilde{\theta}_2$	NNLLh
(a)	35%	0.36	0.78	5.92
(b)	42%	0.34	0.77	8.35

(c)

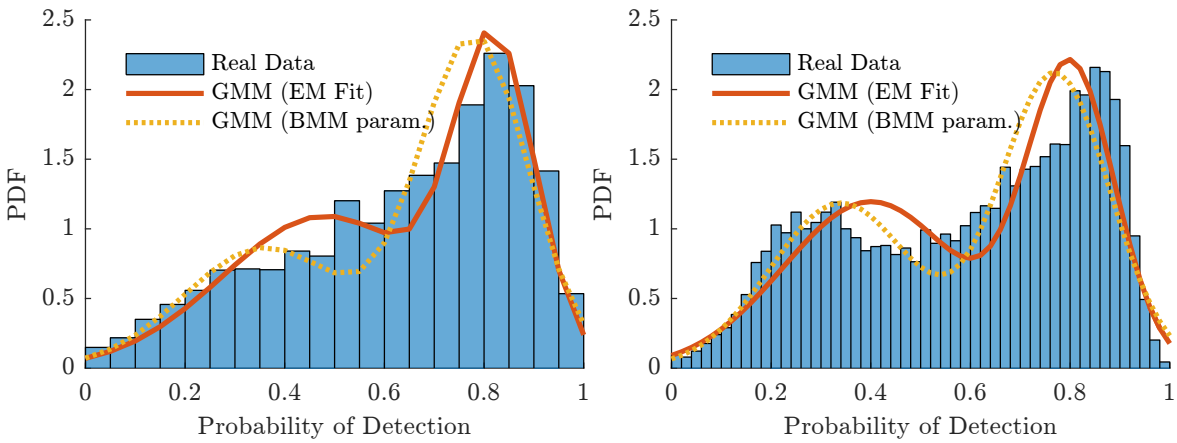
Figure 4.12: Histogram of the Bluetooth detection probability with the proposed Binomial Mixture Models (a) for $min_i(h_i - 2) = 20$ (b) for $min_i(h_i - 2) = 50$ (Data: 25th to 31st of October 2014).

These models agree on both cases for the identified modes of detection probabilities: a first mode close to 0.35 and a second close to 0.78. However, the models do not fit visually the data well. This is a consequence of the binomial distributions whose widths are fixed. The fact that a binomial mixture might not explain well the observed distribution probably stems from some variabilities of the detection probabilities, within the two classes, due to other factors such as, for example, the speed, the weather, the detectors and their positions. Thus, in order to allow for some variability around the two modes, a Gaussian mixture model is proposed next.

4.3.4 Gaussian Mixture Model on Retrieved Trajectories

The Gaussian mixture model is here applied with Algorithm 4.3 to those same trajectories, in the same condition as in previous section and results are similarly presented in Figure 4.13. In this figure, the red and full lines represent the results obtained from the EM algorithm (*GMM (EM fit)*). Both of the models, with $(min_i(h_i - 2) = 20)$ and $(min_i(h_i - 2) = 50)$ have similar results, yet the means are not very consistent with the BMM model. In particular, the first component of the mixture seems to have an overestimated mean compared to what is observed, both with the BMM and visually.

Consequently, in a second step, the mixture and mean parameters, as estimated by the BMM, have been used as fixed parameters and the optimisation has been performed on variances only. Those results are referred as *GMM (BMM param.)*. Interestingly, theses new fits have almost the same NNLLh, that indicates that both models fits the data similarly. However this approach allows, in addition, for consistency with the BMM model and is therefore more appealing.



(a) (b)

Figure	$\tilde{\pi}_1$	$\tilde{\mu}_1$	$\tilde{\mu}_2$	$\tilde{\sigma}_1$	$\tilde{\sigma}_2$	NNLLh
(a) EM fit	54%	0.48	0.82	0.0422	0.0065	2.84
(a) BMM param.	35%	0.36	0.78	0.0260	0.0130	2.85
(b) EM fit	53%	0.40	0.80	0.0312	0.0078	3.79
(b) BMM param.	42%	0.34	0.77	0.0200	0.0120	3.79

(c)

Figure 4.13: Histogram of the Bluetooth detection probability with the two proposed Gaussian Mixture Models (a) for $min_i(h_i - 2) = 20$ (b) for $min_i(h_i - 2) = 50$ (Data: 25th to 31st of October 2014).

4.3.5 BMM and GMM on Corridors: Bias and Correction of the Mixture Coefficient

These two algorithms can, very similarly, be applied to the corridors presented in Chapter 3, Section 3.3.

In this case however, the first and last detectors at which Bluetooth devices have to be detected are fixed. Hence, the mixture as computed by the EM algorithm will be biased towards Bluetooth devices with higher detection probability. We propose therefore a correction to be applied to the mixture coefficient as follow.

Let N denote the total population of Bluetooth devices composed of two classes with mixture coefficient $\pi_1^\star$ for the first component. Then the number of Bluetooth devices in each class, N_1 and N_2 are written as:

$$\begin{cases} N_1 &= \pi_1^\star \cdot N \\ N_2 &= (1 - \pi_1^\star) \cdot N \end{cases} \tag{4.10}$$

In case of the BMM model, each class has its detection probability θ_1 or θ_2. Moreover, the sample of users used for the statistics on the corridors corresponds to the set of users detected at the first and

at the last detector of the corridor. Thus, we actually have,

$$\begin{cases} N_{1,sampled} &= N_1 \cdot \theta_1^2 &= \pi_1^\star \cdot N \cdot \theta_1^2 \\ N_{2,sampled} &= N_2 \cdot \theta_2^2 &= (1-\pi_1^\star) \cdot N \cdot \theta_2^2 \end{cases} \tag{4.11}$$

Therefore, the mixture model algorithm will estimate a mixture coefficient π_1 that verifies:

$$\begin{aligned} \pi_1 &= \frac{N_{1,sampled}}{N_{1,sampled} + N_{2,sampled}} \\ &= \frac{\theta_1^2 \pi_1^\star}{\theta_1^2 \pi_1^\star + \theta_2^2 (1-\pi_1^\star)} \end{aligned} \tag{4.12}$$

As a consequence, once π_1 is estimated, the real mixture coefficient is derived as:

$$\pi_1^\star = \frac{\theta_2^2 \pi_1}{\theta_2^2 \pi_1 + \theta_1^2 (1-\pi_1)}. \tag{4.13}$$

For the GMM model, $(\theta_k)_{k\in[1,2]}$ can be replaced with $(\mu_k)_{k\in[1,2]}$, the mean and the expectation of a Gaussian distribution being equal.

Eventually, the Gaussian Mixture Model and the Binomial Mixture Model are applied on the Waterworks Road corridor. This corridor is chosen because it is the longest corridor among the five proposed and Figure 4.10 highlighted that the two modes really appear when at least 20 detectors are involved. The results are presented in Table 4.5 and in Figure 4.14.

Table 4.5: Mixture Models Applied to the Waterworks Road corridor, 20 detectors and 19861 Bluetooth Devices

Model	π_1	$\pi_1^\star$	μ_1	μ_2	σ_1	σ_2	**NNLLh**
BMM	11%	34%	0.41	0.84	N.A.	N.A.	5.08
GMM	22%	40%	0.57	0.86	0.0411	0.0044	2.17
GMM (BMM param.)	11%	34%	0.41	0.84	0.0300	0.0062	2.20

These results are consistent with the mixture derived on the whole set of data, once the mixture coefficient correction is applied to the model. This proves also that the sampling biases have a strong impact on the results and require corrections.

The question of the sampling bias in the case of the mixture derived from the whole set of trajectories remains however. In fact, Bluetooth devices with lower detection probability will also have a lower probability to belong to the sample and should have, in average, shorter trajectories. This effect could be corrected if one knew the distribution of the length of the trips in Brisbane and thus of the expected number of detectors along a trajectory. In this case however, such distribution is not available. It is not possible to rely on the Bluetooth dataset for its inference because it is the dataset whose bias we aim to correct. The impact of such bias is mitigated however when the length of the trajectories increases as most of Bluetooth devices in use eventually end up being detected. In the Brisbane case, most of the trajectories are several kilometres long, thus the bias should have a limited impact.

Otherwise, these results are a second proof that Bluetooth detections are not random independent events. Two classes of detections, corresponding to two Gaussian distributions, coexist, but the reasons behind those two classes are yet not clear. Future works could focus on this point.

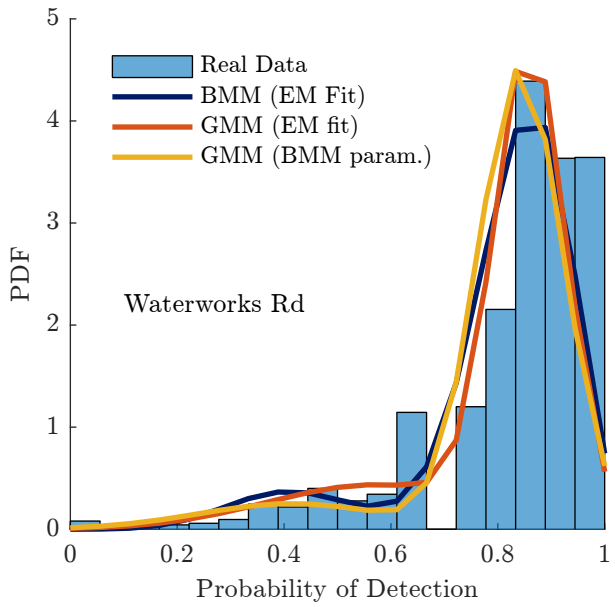


Figure 4.14: Representation as an histogram of the Bluetooth devices detection probability with the three proposed mixture models (Data: 25th to 31st of October 2014).

4.4 Combining Trajectories with other Datasets

In this chapter, we have seen that trajectories can be recovered from Bluetooth data with up to 84% precision between first and last detections. With those trajectories, further analysis of the Bluetooth data as a complementary traffic information source has been performed. This shows that Bluetooth devices can be segregated into three groups that have the characteristics of, first, *non-motorised trips*, second, motorised and *direct trips*, and last motorised *errands trips*.

Furthermore, it has been shown that Bluetooth devices do not always have the same detection probability and that those probabilities can be described with a mixture of two Gaussian distributions. It could be the matter of future works to look deeper into these distributions and to look for justification of this mixture. In addition, also in future works, these detection probabilities could be involved in the trajectory recovery process in order to discriminate the most likely trajectory among a set of possible one, if one were to consider more paths than the spatially constrained shortest one.

In any case, the trajectories, once recovered can now be combined with other traffic datasets, in particular with traffic counts, in order to improve the estimation processes of the travel demand. In the next chapter, we will propose an original framework in order to jointly estimate the OD demand and the usage of the network from both the traffic counts and the set of probe trajectories. To do so, a new tool, the link dependent origin destination matrix, will be presented. A procedure for its estimation will be proposed that account for the typology of the road infrastructure given by the graph $\mathscr{G} = (V, L)$ presented in Chapter 3, Section 1 while being consistent with observed traffic information.

CHAPTER 5

Link dependent Origin Destination Matrices

Contents

1 From OD Matrix to LOD Matrix Estimation: Summary of the Problem **95**
2 Road Network and LOD matrix **96**
2.1 Problem Statement 96
2.2 Road Network as a Graph 96
2.3 Model, Measures and Estimates 98
3 Functional Optimisation Formulation **98**
3.1 Objective Function 98
3.2 Algorithm 101
4 Simulated Case Study **103**
4.1 Experimental setup 103
4.2 Performances 105
4.3 Lower Time Granularity 109
5 Alternative Formulations of the Problem **109**
5.1 Optimisation on Bluetooth Penetration Factors 109
5.2 A simple Forward-Backward approach 110
6 Conclusion **110**

Build upon works published in:

- G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 19–24, 2015, pp. 5480–5484. DOI: 10.1109/ICASSP.2015.7179019 (*cf.* Appendix C)
- G. Michau, P. Borgnat, N. Pustelnik, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", presented at the XXV GRETSI, Lyon, France, Sep. 8, 2015. [Online]. Available: http://eprints.qut.edu.au/86449/ (*cf.* Appendix D)
- G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Graph Signal Processing Workshop*, Philadelphia, May 25–27, 2016
- G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Complex Networks*, Marseilles, France, Jul. 11–13, 2016
- G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, A. Bhaskar, and E. Chung, "A primal-dual algorithm for link dependent origin destination matrix estimation", *Submitted in IEEE Transactions on Signal and Information Processing Over Networks*, 2016

1 From OD Matrix to LOD Matrix Estimation: Summary of the Problem

THE traditional problem of OD matrix estimation aims to solve Problem (2.4) rewritten here (see Chapter 2, Section 1.3.2):

$$\begin{aligned}(\widehat{\underline{\underline{T}}}, \widehat{\underline{q}}) \in & \underset{\underline{\underline{T}},\underline{q}}{\text{Argmin}} \left\{\gamma_1 \mathscr{D}_1(\tilde{\underline{\underline{T}}}, \underline{\underline{T}}) + \gamma_2 \mathscr{D}_2(\tilde{\underline{q}}, \underline{q})\right\} \\ & \text{s.t.} \quad \underline{q} = F(\underline{\underline{T}}),\end{aligned} \tag{2.4}$$

where the road network is represented by a graph $\mathscr{G} = (V, L)$, with $\underline{\underline{T}}$ the corresponding OD matrix, of size $N_V \times N_V$. The assignment function F relates estimated OD flows to network links, for comparisons against traffic counts $\tilde{\underline{q}}$. We suppose here that these traffic counts are measured by N_L magnetic loops (on each link $l \in L$).

The two main difficulties in solving Problem (2.4) stem, first, from its being ill-posed: the size of $\underline{\underline{T}}$ is $N_V \times N_V$ whereas the size of $\tilde{\underline{q}}$, the measures, is N_L and second, from F being unknown, and thus often modelled.

Despite the fact that $\underline{\underline{T}}$ is of size $N_V \times N_V$, solving Problem (2.4) is in fact an inverse problem of size $N_V \times N_V \times N_L$, because the required assignment step, here characterised by means of the function F, actually involves the number of links in the network (e.g., Equation (2.2) in Chapter 2, Section 1.2.3).

The goal of the present chapter is to directly account for the real dimensionality of the problem by proposing a new and original description tool for traffic, that directly includes assignment: the Link dependent Origin Destination matrix (LOD matrix). LOD matrix represents the OD flows already assigned to each link of the network, thus incorporating the assignment, or equivalently making its independent specification unnecessary. We also propose to estimate LOD matrix as an inverse problem of dimension $N_V \times N_V \times N_L$. We rely on traffic counts $\tilde{\underline{q}}$ and, in addition, on a new set of data: a set of trajectories, that is, for a fraction of the users, the set of roads used to travel from origin to destination. Trajectories are interpreted as a sample of the LODM. Trajectory collection is now made possible by new technologies such as, among others GPS [112], Bluetooth (see Chapter 4 or [14], [106], [149]) or floating cars data [189]. The actual technology matters little; we propose, though, to refer to the Bluetooth technology, as it is the technology of interest here. Also, it currently provides trajectory datasets with the highest penetration rate, compared to other technologies.

Section 2 formulates the transport problem, from an engineering perspective and develops an expression similar to that of Problem (2.4). Section 3 aims to define the objective function of this minimisation problem. To do so, it details five significant properties imposed either by the network or for consistency with observed data and turns them into five components of the objective function. This formalises the LOD matrix estimation problem from traffic counts and sampled trajectories. A proximal primal-dual algorithm is devised to minimise the resulting linear combinations of these five functions, which are convex but non necessarily differentiable. To finish with, the feasibility of the proposed approach and the assessment of its estimation performance are investigated in Section 4, on a case study consisting of network and traffic simulations, designed to match network and traffic in large western metropolitan cities.

This formulation has been built upon preliminary works, presented in conference proceedings [17] and [20], discussed in Section 5 and detailed in Appendices C and D.

Notations Reminder

The following notations are recalled (for full description see Chapter Notations) :

- $\underline{X}$, $\underline{\underline{X}}$ and $\underline{\underline{\underline{X}}}$ respectively refer to vectors, matrices and tensors.
- The Hadamard (element-wise) product of $\underline{\underline{\underline{C}}}$ and $\underline{\underline{\underline{X}}}$ is denoted $\underline{\underline{\underline{C}}} \circ \underline{\underline{\underline{X}}}$.
- The symbol $\bullet$ is used to denote the dimension that does not contribute to a sum: e.g., the sum over first and third dimensions is written $\sum_{i\bullet l} \underline{\underline{\underline{X}}}$.
- We denote by $\|\cdot\|$ the element-wise norm for matrices: e.g., $\|\underline{\underline{X}}\|_1 = \sum_{ij} |X_{ij}|$ and $\|\underline{\underline{X}}\|_2 = \left(\sum_{ij} X_{ij}^2\right)^{\frac{1}{2}}$.
- The notation $\tilde{}$ (as in $\tilde{X}$) is used preferentially for measured variables.
- The notation * (as in $X^\star$) is used preferentially to refer to the original variable (or true variable).
- The notation $\widehat{}$ (as in $\widehat{X}$) is used preferentially for estimates.

2 Road Network and LOD matrix

2.1 Problem Statement

In the network graph $\mathscr{G} = (V, L)$, the finite set of nodes V models intersections of the road network and each node also defines a possible origin or destination. L is the set of directed edges, each corresponding to a direct itinerary (or road) linking two nodes (*i.e.,* not going through another node in V). The number of road users is denoted N. A schematic (small size) such graph is illustrated in Figure 5.1a.

On such a graph, LOD matrix consists in a tensor of size $N_V \times N_V \times N_L$, labelled $\underline{\underline{\underline{Q}}} = (Q_{ij}^l)_{(i,j)\in V^2, l\in L}$. As illustrated in Figure 5.1, each trajectory, of origin i and destination j adds a count of 1 in Q_{ij}^l if the link l is on the trajectory. Therefore, $\underline{\underline{\underline{Q}}}$ consists, for each link $l \in L$, in an OD matrix of size $N_V \times N_V$.

Information on trajectories is stored into a tensor labelled $\underline{\underline{\underline{\tilde{B}}}}$, of size $N_V \times N_V \times N_L$. $\underline{\underline{\underline{\tilde{B}}}}$ can be read as a LODM consisting in a sampled version of $\underline{\underline{\underline{Q}}}$, for a fraction of the total traffic. Traffic counts consist of the total volume of traffic on each link $l \in L$, labelled $\underline{\tilde{q}}$, of size N_L, irrespective of OD pairs. Traffic count can be, for instance, measured by magnetic loops.

An approach will now be devised to estimate $\underline{\underline{\underline{Q}}}^\star$, the real LOD matrix, by means of nonsmooth convex optimisation from $\underline{\underline{\underline{\tilde{B}}}}$ and $\underline{\tilde{q}}$. The convex function represents on the one hand the relationships that link the estimate $\underline{\underline{\underline{\widehat{Q}}}}$ with the measures ($\underline{\underline{\underline{\tilde{B}}}}$, $\underline{\tilde{q}}$) and, on the other hand, properties of the road network and traffic constraints (e.g., car conservation at intersections).

2.2 Road Network as a Graph

The structure of the graph is given by the *incidence* and *excidence* matrices denoted respectively $\underline{\underline{I}}$ and $\underline{\underline{E}}$ of size $N_V \times N_L$. These matrices describe the relations between nodes and edges: For every

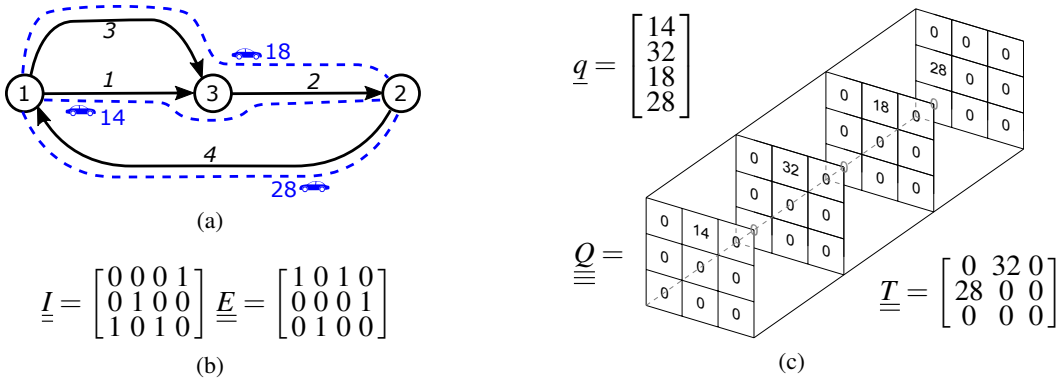


Figure 5.1: Example of a simple network ($N_V = 3, N_L = 4$ and $N = 60$) (a) with the associated tools describing the topology (b) and the traffic (c).

$(k,l) \in V \times L$,

$$\begin{aligned} I_k^l &= \begin{cases} 1 & \text{if the link } l \text{ is arriving to the node } k, \\ 0 & \text{otherwise,} \end{cases} \\ E_k^l &= \begin{cases} 1 & \text{if the link } l \text{ is starting from the node } k, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{5.1}$$

Note that in graph theory, the *Incidence Matrix* usually refers to the difference $(\underline{\underline{I}} - \underline{\underline{E}})$; however we need both matrices separately in this work.

Let us also define tensors $\underline{\underline{\underline{I_1}}}$ and $\underline{\underline{\underline{I_2}}}$ (resp. $\underline{\underline{\underline{E_1}}}$ and $\underline{\underline{\underline{E_2}}}$) corresponding to the replication of $\underline{\underline{I}}$ (resp. $\underline{\underline{E}}$) along a third index such that,

$$\begin{aligned} (\forall m \in V) \quad (I_1)_{km}^l &= \begin{cases} 1 & \text{if link } l \text{ is arriving to node } k, \\ 0 & \text{otherwise,} \end{cases} \\ (\forall k \in V) \quad (I_2)_{km}^l &= \begin{cases} 1 & \text{if link } l \text{ is arriving to node } m, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{5.2}$$

Using these notations, we relate the LOD matrix ($\underline{\underline{\underline{Q}}}$) to the classical OD matrix $\underline{\underline{T}}$ of size $N_V \times N_V$ where each element T_{ij} contains the traffic flow originating from the node i and having j for destination as follows:

$$\underline{\underline{T}} = \sum_{\bullet\bullet l} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{Q}}} = \sum_{\bullet\bullet l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{Q}}} \tag{5.3}$$

under the assumptions that the trajectories represented in $\underline{\underline{\underline{Q}}}$ do not have cycles. The interested reader can refer to Appendix E, Section 1 for more details on those two relationships.

We denote by $\underline{O}$ (resp. $\underline{D}$) the *origin* (resp. *destination*) vector, of size N_V as the sum of $\underline{\underline{T}}$ over the second (resp. first) dimension. It represents the flows originating (or having for destination) each node of the graph. Formally,

$$\begin{cases} \underline{D} = (\Sigma_{i\bullet} \underline{\underline{T}})^\top, \\ \underline{O} = \Sigma_{\bullet j} \underline{\underline{T}}. \end{cases} \tag{5.4}$$

2.3 Model, Measures and Estimates

In realistic networks, roads are ranked by transport engineers depending on several parameters (e.g., speed limit, capacity and priority at intersections). Usually, when road monitoring is planed, roads of highest ranks only are equipped with monitoring devices. Low ranks roads are excluded from traffic studies for their traffic is low and not crucial to urban mobility. For this problem, we consider an urban road network with highest rank roads only (as proposed in Chapter 3, Section 1) and we assume that every road is equipped with a magnetic loop, counting the number of cars using it. It implies therefore that every element in $\underline{\tilde{q}}$ is known.

The magnetic loops are usually subject to counting errors and it is modelled here by a noise $\underline{\varepsilon}$. Hence the measured quantity $\underline{\tilde{q}}$ reads:

$$\underline{\tilde{q}} = \underline{q}^\star + \underline{\varepsilon} \tag{5.5}$$

where $\underline{q}^\star$ is the true traffic volumes:

$$\underline{q}^\star = \sum_{ij\bullet} \underline{\underline{\underline{Q}}}^\star \tag{5.6}$$

Second, we also assume that Bluetooth devices are not turned on and off while users are travelling. If this assumption holds, the penetration rate can be defined per OD as the number of Bluetooth equipped vehicles divided by the total traffic for this particular OD and is denoted η_o of size $N_V \times N_V$. Measures in Brisbane have shown that the average Bluetooth penetration rate is around 25% (see Chapter 3, Section 3.2.3). Moreover, $\underline{\underline{\underline{\tilde{B}}}}$ appears as a sampled version of $\underline{\underline{\underline{Q}}}^\star$. The relation between the tensors $\underline{\underline{\underline{\tilde{B}}}}$ and $\underline{\underline{\underline{Q}}}^\star$ can thus be modelled by a Poisson law, typically used for counting processes (*cf.* Chapter 2, Section 1.3.2):

$$(\forall i, j, l \in V \times V \times L) \quad B_{ij}^l = \mathscr{P}\left((\eta_o)_{ij}\, Q_{ij}^{\star l}\right). \tag{5.7}$$

More formally, the problem we propose to solve can be expressed, similarly to Problem (2.4), by the following general problem:

$$\underline{\underline{\underline{\widehat{Q}}}} \in \underset{\underline{\underline{\underline{Q}}}}{\text{Argmin}} \left\{ \gamma_1 \mathscr{D}_1(\underline{\underline{\underline{\tilde{B}}}}, \underline{\underline{\eta_o}}\, \underline{\underline{\underline{Q}}}) + \gamma_2 \mathscr{D}_2(\underline{\tilde{q}}, \sum_{ij\bullet} \underline{\underline{\underline{Q}}}) + \sum_{k\geq 2} \gamma_k f_k(\underline{\underline{\underline{Q}}}) \right\} \tag{5.8}$$

where $\mathscr{D}_1$ and $\mathscr{D}_2$ are functions designed to quantify deviations between estimates and observed data while the functions f_k models important properties that the estimate should satisfy.

3 Functional Optimisation Formulation

Instead of using the traditional four-steps model resolution, iterating over a process involving *a priori* information, modelling of the traffic, estimating the variables of interest, comparing to the observed measures and tuning the models, we propose here to solve Problem (5.8) stated as a function minimisation. In the following, the terms composing the objective function are presented.

3.1 Objective Function

The terms of the objective function can be classified in three types: The first type, composed of the functions presented in Section 3.1.1, Section 3.1.2 and Section 3.1.3, is aiming for consistency

between measures and estimate. The second type, with the function of Section 3.1.4, stems from the topology of the network. The third type, with the function of Section 3.1.5, comes from an additional assumption based on knowledge on transport networks.

3.1.1 Traffic Count Data Fidelity f_{TC}

Ensuring the consistency with traffic counts would require that Equations (5.5) and (5.6) are satisfied. Assuming a random unbiased Gaussian noise $\underline{\varepsilon}$ for the magnetic loops, as in Equation (5.5), the constraint of Equation (5.6) can be relaxed and leads to the following function:

$$f_{TC}(\underline{\underline{\underline{Q}}}) = \|\underline{\tilde{q}} - \sum_{ij\bullet} \underline{\underline{\underline{Q}}}\|^2. \tag{5.9}$$

In the classical OD matrix estimation problem, this would correspond to the least square estimators (*cf.* Chapter 2, Section 1.3.2).

3.1.2 Poisson Bluetooth Sampling Data Fidelity f_P

Second, the consistency with Bluetooth measures, as modelled in Equation (5.7) requires the knowledge of the OD-dependent penetration rate $\underline{\underline{\eta_o}}$. This information, of size $N_V \times N_V$, is not directly available from $\underline{q}$ and $\underline{\underline{\underline{\tilde{B}}}}$, therefore we introduce an approximation of this penetration rate of size N_L, noted $\underline{\tilde{\eta}}$ and computed as a link dependent estimate of $\underline{\underline{\eta_o}}$:

$$\underline{\tilde{\eta}} = \frac{\sum_{i,j,\bullet} \underline{\underline{\underline{\tilde{B}}}}}{\underline{\tilde{q}}}. \tag{5.10}$$

The resulting data fidelity term, denoted f_P, models the negative log-likelihood associated with the Poisson model [190]:

$$f_P(\underline{\underline{\underline{Q}}}) = \sum_{ijl} \psi\left(B^l_{ij}, \eta^l Q^l_{ij}\right) \tag{5.11}$$

where, for every $(u,v) \in \mathbb{R}^2$,

$$\psi(u,v) = \begin{cases} -u \log v + v & \text{if } v > 0 \text{ and } u > 0, \\ v & \text{if } v \geq 0 \text{ and } u = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.12}$$

3.1.3 Consistency Constraint f_C

Third, another term ensuring data consistency models that the total flow should be greater than the flow of Bluetooth enabled vehicles. It consists thus in imposing that $\underline{\underline{\underline{Q}}}$ belongs to the following convex set C:

$$C = \left\{\underline{\underline{\underline{Q}}} = \left(Q^l_{ij}\right)_{(ijl)\in V\times V\times L} \in \mathbb{R}^{N_V\times N_V\times N_L} \mid Q^l_{ij} \geq B^l_{ij}\right\}. \tag{5.13}$$

The corresponding convex function is the indicator function ι_C:

$$f_C(\underline{\underline{\underline{Q}}}) = \iota_C(\underline{\underline{\underline{Q}}}) = \begin{cases} 0 & \text{if } \underline{\underline{\underline{Q}}} \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.14}$$

3.1.4 Kirchhoff's Law f_K

This property is the classical law for flows on network, the Kirchhoff's law, describing the conservation of cars at intersections. It takes into account the network topology. It requires that, for each OD pair and at every node, the number of cars is conserved when properly accounting for origins and destinations. For every origin $i \in V$, destination $j \in V$ and node $k \in V$ of the network, this yields to,

$$\sum_l E_k^l Q_{ij}^l - \underbrace{\delta_{ik} T_{ij}}_{\substack{\text{origin}\\ \text{(source)}}} = \sum_l I_k^l Q_{ij}^l - \underbrace{\delta_{jk} T_{ij}}_{\substack{\text{destination}\\ \text{(sink)}}} . \tag{5.15}$$

where δ_{ij} is the Kronecker delta, that is:

$$\forall (i,j) \in V \times V \qquad \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \tag{5.16}$$

This constraint can then be reformulated as

$$\left(\forall (i,j,k) \in V^3\right) \quad \sum_l A_{ijk}^l Q_{ij}^l = 0 \tag{5.17}$$

where the $N_V \times N_V \times N_V \times N_L$ tensor $\underline{\underline{\underline{A}}}$ is defined as

$$\left(\forall (i,j,k,l) \in V^3 \times L\right) \quad A_{kij}^l = \left(E_k^l - I_k^l\right) - \left(\delta_{ik} - \delta_{jk}\right) E_i^l. \tag{5.18}$$

It results in a convex function:

$$f_K(\underline{\underline{\underline{Q}}}) = \sum_{ijk} \left(\sum_l A_{ijk}^l Q_{ij}^l \right)^2. \tag{5.19}$$

Compared to the preliminary works presented in Appendix C or [17] and Appendix D or [20], the Kirchhoff's law is here applied per OD pair rather than globally. Indeed, the Kirchhoff's law needs also to be satisfied at each node, independently of the origin and destination of the cars. Satisfying Equation (5.17) automatically implies that the Kirchhoff's law used in these preliminary works is satisfied, as demonstrated in Appendix E, Section 2.2.

3.1.5 Total Variation f_{TV}

Finally, from a transport perspective, it seems realistic to assume that for two paths having close origins (resp. destinations) and same destination (origin), trajectories should be alike (e.g., use of similar roads):

$$(\forall i \sim i')(\forall j \in V)(\forall l \in L) \qquad Q_{ij}^l \sim Q_{i'j}^l \tag{5.20a}$$

$$(\forall j \sim j')(\forall i \in V)(\forall l \in L) \qquad Q_{ij}^l \sim Q_{ij'}^l \tag{5.20b}$$

These relationships can be gathered within the convex function f_{TV} defined as the total variation:

$$f_{TV}(\underline{\underline{\underline{Q}}}) = \sum_{i \sim \mathcal{N}_{i'}} \sum_{j,l} \omega_{ii'} |Q_{ij}^l - Q_{i'j}^l| + \sum_{j \sim \mathcal{N}_{j'}} \sum_{i,l} \omega_{jj'} |Q_{ij}^l - Q_{ij'}^l| \tag{5.21}$$

where $\mathcal{N}_{i'}$ models the neighbourhood of i' and where $\omega_{ii'}$ are positive weights on edges detailed in Equation (5.23).

The use of the ℓ_1-norm is justified for its *edge preservation* properties. Indeed, it has been shown in [7], [191] that the ℓ_1-norm is adapted for cases where one seeks for spatial correlations while allowing some irregularities, e.g., *edges*, in image analysis. From a traffic perspective, we want to favour users from similar origin (resp. destination) and with same destination (resp. origin) to use similar routes, but for some variations: For example, users can use different roads is their origins (or destination) are in between two major roads. Those origins and destinations can be interpreted as *edges* in image analysis.

Equation (5.21) can further be simplified using a weighted effective incidence matrix, denoted $\underline{\underline{J}}$, defined as

$$\big(\forall (k,l) \in V \times L\big) \qquad J_k^l = W^l (I_k^l - E_k^l) \tag{5.22}$$

and thus having a size $|V| \times |L|$, where each element W^l denotes the weight for the link l. If l is a link between k and m then:

$$W^l = \omega_{km} = e^{-\frac{d_l}{d_0}}. \tag{5.23}$$

d_l models the length of the link l and d_0 is the average distance of the nodes. Equation (5.21) can then be rewritten as

$$f_{TV}(\underline{\underline{\underline{Q}}}) = \sum_l \|\underline{\underline{J}}^\top \underline{\underline{Q}}^l\|_1 + \sum_l \|\underline{\underline{J}}^\top (\underline{\underline{Q}}^l)^\top\|_1. \tag{5.24}$$

where $\underline{\underline{Q}}^l$ models the l-th extracted matrix from $\underline{\underline{\underline{Q}}}$. Its dimension is thus $|V| \times |V|$.

3.2 Algorithm

To sum up, the objective is to find an estimate $\widehat{\underline{\underline{\underline{Q}}}}$ of $\underline{\underline{\underline{Q}}}^\star$ satisfying

$$\widehat{\underline{\underline{\underline{Q}}}} \in \underset{\underline{\underline{\underline{Q}}}}{\text{Argmin}} \left\{ \gamma_{TC} f_{TC}(\underline{\underline{\underline{Q}}}) + \gamma_P f_P(\underline{\underline{\underline{Q}}}) + \gamma_C f_C(\underline{\underline{\underline{Q}}}) + \gamma_K f_K(\underline{\underline{\underline{Q}}}) + \gamma_{TV} f_{TV}(\underline{\underline{\underline{Q}}}) \right\} \tag{5.25}$$

where $\gamma_\cdot$ are positive weights applied to the objectives to model their relative importance within the global objective.

All five functions involved in Equation (5.25) are convex, lower-semicontinuous (l.s.c.) and proper. Moreover, both functions f_{TC} and f_K are differentiable, with gradients given below:

$$\nabla f_{TC}(\underline{\underline{\underline{Q}}}) = \left(\left(-2 \left(\tilde{q}^l - \sum_{k,m} Q_{km}^l \right) \right)_{ij}^l \right)_{(ijl) \in V \times V \times L} \tag{5.26}$$

and

$$\nabla f_K(\underline{\underline{\underline{Q}}}) = \left(\left(2 \sum_k A_{jik}^l \sum_e A_{ijk}^e \, Q_{ij}^e \right)_{ij}^l \right)_{(ijl) \in V \times V \times L}. \tag{5.27}$$

Their Lipschitz constants are denoted β_{TC} and β_K respectively [192]. The other three functions however are not differentiable and f_{TV} involves a linear transformation H such as:

$$f_{TV}(\underline{\underline{\underline{Q}}}) = \|H(\underline{\underline{\underline{Q}}})\|_1 \tag{5.28}$$

with H defined as:

$$\begin{aligned} H \colon \mathbb{R}^{|V| \times |V| \times |L|} &\to \mathbb{R}^{|L| \times |V| \times |L|} \times \mathbb{R}^{|L| \times |V| \times |L|} \\ \underline{\underline{\underline{Q}}} &\mapsto \left(\left(\underline{\underline{J}}^\top \underline{\underline{Q}}^l\right)_{l \in L}, \left(\underline{\underline{J}}^\top (\underline{\underline{Q}}^l)^\top\right)_{l \in L} \right) \end{aligned} \tag{5.29}$$

and whose adjoint is

$$H^*\colon (\underline{\underline{R}},\underline{\underline{S}}) \mapsto \left(\underline{\underline{J}}\,\underline{\underline{R}}^l\right)_{l\in L} + \left((\underline{\underline{J}}\,\underline{\underline{S}}^l)^\top\right)_{l\in L}. \tag{5.30}$$

In the following, we denote χ_H the norm of this operator. For further details about the way to compute this norm, the reader is refered to [192].

Optimisation Problem (5.25) is solved by means of a primal-dual proximal algorithm, as in [12], [13], [193], [194], which is particularly suited when the objective combines differentiable and non-differentiable functions along with linear operators. In such an iterative scheme, the non-differentiable functions are involved through their proximity operator [195] defined as:

$$(\forall u \in \mathscr{H}) \qquad \operatorname{prox}_f(u) = \arg\min_{x\in\mathscr{H}} f(x) + \frac{1}{2}\|u-x\|_2^2 \tag{5.31}$$

where $\mathscr{H}$ denotes a real Hilbert space and f a convex, l.s.c., proper function from $\mathscr{H}$ to $]-\infty,+\infty]$. For further details on proximal algorithms, the reader could refer to [10], [196], [197].

The proximity operator of the indicator of the convex set C has a closed form expression as a projection [198]:

$$\operatorname{prox}_{\gamma_C f_C}(\underline{\underline{Q}}) = \begin{cases} P_C(\underline{\underline{Q}}) = \max(\underline{\underline{Q}},\underline{\underline{\tilde{B}}}) & \text{if } \gamma_C > 0 \\ \underline{\underline{Q}} & \text{if } \gamma_C = 0. \end{cases} \tag{5.32}$$

The proximity operator of function, f_P, also have a closed form expression [190]:

$$\begin{aligned} \operatorname{prox}_{\gamma_P f_P}(\underline{\underline{Q}}) &= \left(\operatorname{prox}_{\gamma_P \psi}(B^l_{ij}, \eta^l Q^l_{ij})\right)_{(ijl)\in V\times V\times L} \\ &= \left(\frac{Q^l_{ij} - \gamma_P\eta^l + \sqrt{|Q^l_{ij} - \gamma_P\eta^l|^2 + 4\gamma_P B^l_{ij}}}{2}\right)_{(ijl)\in V\times V\times L}. \end{aligned} \tag{5.33}$$

The proximal operator of the sum of these two functions satisfies the following property [199]:

$$\operatorname{prox}_{\gamma_C f_C + \gamma_P f_P}(\underline{\underline{Q}}) = P_C(\operatorname{prox}_{\gamma_P f_P}(\underline{\underline{Q}})). \tag{5.34}$$

The ℓ_1-norm, applied to H, as in Equation (5.28), also has a closed form expression for its proximity operator [200]–[203]:

$$\operatorname{prox}_{\gamma_{TV}\|\cdot\|_1}(\underline{\underline{R}},\underline{\underline{S}}) = \left(\operatorname{sign}(\underline{\underline{R}})\ \max\{|\underline{\underline{R}}| - \gamma_{TV}, 0\}, \operatorname{sign}(\underline{\underline{S}})\ \max\{|\underline{\underline{S}}| - \gamma_{TV}, 0\}\right). \tag{5.35}$$

The primal-dual proximal iterations designed for minimising Equation (5.25) are described in Algorithm 5.1. Under some technical assumptions regarding the domain of definition and the following condition:

$$\frac{1}{\tau} - \sigma\chi_H \geq \frac{\beta}{2}, \tag{5.36}$$

where $\beta = \gamma_{TC}\beta_{TC} + \gamma_K\beta_K$ denotes the Lipschitz constant of $\gamma_{TC}f_{TC} + \gamma_K f_K$ and $\sigma > 0$, the sequence $\left(Q^{k+1}\right)_{k\in\mathbb{N}}$ converges to a minimiser of Equation (5.25) [12, theorem (3.1)]. Algorithm 5.1 has one stopping criterion based on the convergence of the estimates. Yet, to limit computation time, we added a limit at 10^5 iterations. This limit is seldom reached and results for which it has been reached are considered as non acceptable solutions.

Algorithm 5.1 Proximal Primal Dual Algorithm for LOD Matrix Estimation

Input: $\gamma_{TC} \geq 0, \quad \gamma_K \geq 0, \quad \gamma_{TV} \geq 0, \quad \gamma_P \in [0,1], \quad \gamma_D \in [0,1]$
Compute: χ_H

$\beta \leftarrow \gamma_{TC}\beta_{TC} + \gamma_K\beta_K$
if $\gamma_{TV} = 0$ **then**
  $\tau \leftarrow \frac{1.99}{\beta}$
else
  Choose (τ, σ) such as $\tau = \frac{0.9}{\frac{\beta}{2}+\sigma\chi_H} \in [\frac{2}{3}\sigma, \frac{3}{2}\sigma]$
$\underline{\underline{\underline{Q}}}^0 \leftarrow 0$
$(\underline{\underline{\underline{R}}}^0, \underline{\underline{\underline{S}}}^0) \leftarrow (0,0)$
for $k \geq 1 \quad k \in \mathbb{N}$ **do**
  $\underline{\underline{\underline{\widetilde{Q}}}}^{k+1} = \underline{\underline{\underline{Q}}}^k - \tau\big(\gamma_{TC}\nabla f_{TC}(\underline{\underline{\underline{Q}}}^k) + \gamma_K \nabla f_K(\underline{\underline{\underline{Q}}}^k)\big) - \tau H^*(\underline{\underline{\underline{R}}}^k, \underline{\underline{\underline{S}}}^k)$
  $\underline{\underline{\underline{Q}}}^{k+1} = \mathrm{prox}_{\gamma_C f_C}\left(\mathrm{prox}_{\tau\gamma_P f_P}\left(\underline{\underline{\underline{\widetilde{Q}}}}^{k+1}\right)\right)$
  $\left(\underline{\underline{\underline{\widetilde{R}}}}^{k+1}, \underline{\underline{\underline{\widetilde{S}}}}^{k+1}\right) = \sigma H(2\underline{\underline{\underline{\widetilde{Q}}}}^{k+1} - \underline{\underline{\underline{Q}}}^k) + \left(\underline{\underline{\underline{R}}}^k, \underline{\underline{\underline{S}}}^k\right)$
  $\left(\underline{\underline{\underline{R}}}^{k+1}, \underline{\underline{\underline{S}}}^{k+1}\right) = \left(\underline{\underline{\underline{\widetilde{R}}}}^{k+1}, \underline{\underline{\underline{\widetilde{S}}}}^{k+1}\right) - \sigma \cdot \mathrm{prox}_{\gamma_{TV}/\sigma, \ell_1}\left(\frac{1}{\sigma}\underline{\underline{\underline{\widetilde{R}}}}^{k+1}, \frac{1}{\sigma}\underline{\underline{\underline{\widetilde{S}}}}^{k+1}\right)$
  if $\frac{\|\underline{\underline{\underline{Q}}}^{k+1} - \underline{\underline{\underline{Q}}}^k\|_2}{\|\underline{\underline{\underline{Q}}}^{k+1}\|_2} < 10^{-6}$ **or** $k > 10^5$ **then return** $\underline{\underline{\underline{Q}}}^{k+1}$

4 Simulated Case Study

4.1 Experimental setup

4.1.1 Simulation context

To test and validate the proposed method, a simplified road network model has been created. This has been preferred to a real case study for three reasons: tractability, the possibility to access the ground truth and the opportunity to explore the behaviour of the method for varied conditions. However, the connectivity, the number of users and their OD patterns have been chosen to be consistent with those of a real networks.

The number of nodes of the simulated network is $N_V = 50$ nodes. This number is kept relatively low to allow for a thorough exploration of the possible weights $\gamma_{\cdot}$ of problem (5.25). For comparison, the Brisbane Bluetooth scanner network has around 900 intersections equipped with vehicle identification devices. Other works on OD matrix estimation consider often few tens of nodes ($\propto$ 100 OD flows) [204] while very recent works considered up to 300 nodes [60].

For the simulation, nodes are first located randomly on a grid and then links are created while aiming for an average connectivity of 6, a value shared by most of real road networks [205]. This is done first, by means of a minimum spanning tree (computed by the Kruskal's algorithm [206]), then, by adding links randomly to the nodes with lower degree (sum of in and out edges) provided that the added links do not intersect or repeat an existing one.

Thus, the average distance of the nodes, d_0 for such a simulated network is:

$$d_0 = \sqrt{\frac{\text{GridWidth} \cdot \text{GridHeight}}{N_V}}. \tag{5.37}$$

The number of users is fixed to $N = 10^5$. This leads to a mean flow per link of $3 \cdot 10^3$ users. In large cities, it corresponds to around an hour of traffic at peak hours. Each node i has a probability $p_O(i)$ of being an origin and a probability $p_D(i)$ of being a destination. Thus $\underline{p_O}$ and $\underline{p_D}$ satisfy:

$$\mathbb{E}(\underline{D}) = N \times \underline{p_D}, \tag{5.38a}$$

$$\mathbb{E}(\underline{O}) = N \times \underline{p_O}. \tag{5.38b}$$

An origin and a destination are randomly associated to each user, according to the probabilities p_O and p_D. We simulate a preferred direction of travel, to mimic trends observed in urban context (mostly due to commuters). To this end, p_O is decreasing linearly with the X-axis of the grid while p_D is increasing linearly. The shortest path from origin to destination is then assigned to each user.

For each OD pair, a Bluetooth penetration rate is drawn from a Gaussian distribution of mean 30% and standard deviation of 10% (and truncated to be between 0 and 1). This choice accounts for the unknown variability of the ownership distribution of Bluetooth devices from one node to another, depending, as an example, on the wealth of the neighbourhoods of the node. The average is consistent with global penetration rates observed in Brisbane (Section 3.2.3 and [159]). Each user has a probability equal to the Bluetooth penetration rate drawn for its OD of being equipped with a Bluetooth device. This provides $\underline{\underline{\tilde{B}}}$ while the full set of trajectories gives $\underline{\underline{Q}}^\star$ for ground truth. The measured traffic flow per link $\underline{\tilde{q}}$ is obtained from $\underline{\underline{Q}}^\star$, assuming the addition a noise $\underline{\varepsilon}$, for which each independent component is drawn from a Gaussian distribution $\mathcal{N}(0, r \cdot \underline{q}^\star)$ and truncated so that $\underline{\tilde{q}} \geq 0$. For consistency with the noise usually measured on magnetic loops [174], we take $r = 5\%$. Figure 5.2 illustrates the simulated case study with total volumes on the links ($\underline{\tilde{q}}$) and the realisation of p_O and p_D for the 10^5 users on the nodes.

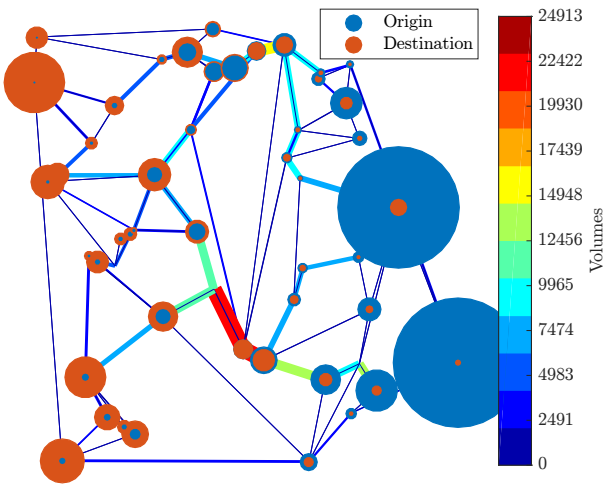


Figure 5.2: Simulated road networks with the projection of $\underline{\tilde{q}}$ on the links: the width is proportional to the flows, also correlated with the color. For nodes, the color distinguishes between origin (blue) and destination flows (red) and the diameter of the nodes is proportional to their value in $\underline{p_O}$ and $\underline{p_D}$.

4.1.2 Algorithmic parameter setup

As discussed in Section 3.2, the objective function (5.25) depends on five parameters $\gamma_\cdot \geq 0$. It appears however that, as f_C can only be 0 or ∞, exploring $\gamma_C \in \{0; 1\}$ is sufficient. Moreover, the minimum in Equation (5.25) is preserved by a linear operation over the four remaining parameters $\gamma_\cdot$, and thus

we choose $\gamma_P \in \{0;1\}$, that is, considering or not the data stemming from sampled trajectories, in the estimation process. We then explore the space of positive real numbers for the three remaining parameters $\gamma_\cdot$.

For those parameters, it has been observed that, for comparison purposes, it is justified to compare scenarii for rescaled values $\gamma_\cdot\beta_\cdot$. Indeed, $\beta_\cdot$ depends on the setup, in particular on $\underline{\tilde{q}}$ and $\underline{\underline{\underline{\tilde{B}}}}$.

The algorithm stops either when the convergence criteria is satisfied (*cf.* Algorithm 5.1), or after a fixed number of iterations, here, 10^5.

4.1.3 Performance evaluation

The efficiency of the estimation algorithm is assessed by comparing its results to the ground truth and the relevance of using such algorithm is established by comparing its best results with two *naive* LODM, estimated directly from the data at hand.

To compare estimates to the ground truth $\underline{\underline{\underline{Q}}}^\star$, we propose two indicators: First, the RMSE:

$$RMSE(\underline{\underline{\underline{\widehat{Q}}}}) = \frac{\|\underline{\underline{\underline{\widehat{Q}}}} - \underline{\underline{\underline{Q}}}^\star\|}{\|\underline{\underline{\underline{Q}}}^\star\|}. \tag{5.39}$$

The RMSE measures the standard deviation between the estimates and the ground truth. Second, to go beyond the simple RMSE, we use the Earth Movers' Distance (EMD) to compare the distribution of the traffic flows between the estimated LODM and the ground truth. It is a metric often used for image comparison and its definition can be found in [207].

The *naive* estimates that can be directly computed from the observed data are denoted $\underline{\underline{\underline{\widehat{Q}}}}^0$ and $\underline{\underline{\underline{\widehat{Q}}}}^1$, computed the Bluetooth LOD matrix multiplied by a penetration factor (inverse of the penetration rate), where the penetration factor is either a global average over the whole network ($\tilde{\alpha}$) or a local average over each link ($\underline{\overline{\alpha}}$):

$$\forall (i,j,l) \in V \times V \times L$$

$$(\widehat{Q}^0)^l_{ij} = \tilde{\alpha}\, B^l_{ij} \quad \text{where} \quad \tilde{\alpha} = \sum_l \underline{\tilde{q}} / \sum_{i,j,l} \underline{\underline{\underline{\tilde{B}}}}, \tag{5.40}$$

$$(\widehat{Q}^1)^l_{ij} = \overline{\alpha}^l\, B^l_{ij} \quad \text{where} \quad \overline{\alpha}^l = \tilde{q}^l / \sum_{i,j,\bullet} \underline{\underline{\underline{\tilde{B}}}}. \tag{5.41}$$

Note that $\underline{\underline{\underline{\widehat{Q}}}}^0$ is related to the solution usually proposed in the literature about ODM estimation from Bluetooth data [129].

4.2 Performances

4.2.1 Impact of the Regularisation Parameters

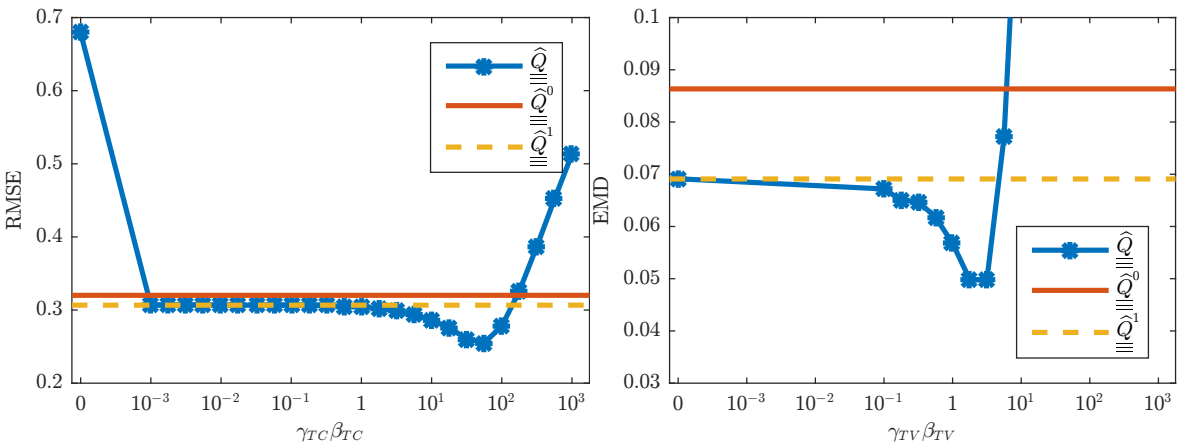
Solutions to Problem (5.25) have been explored through a systematic exploration of the values of $\gamma_\cdot$. This exploration aims to find a set of parameters $\gamma_\cdot$ for which the estimate has minimal criteria and we denote by $\underline{\underline{\underline{\widehat{Q}}}}_{RMSE}$ the estimate $\underline{\underline{\underline{\widehat{Q}}}}$ minimising the RMSE and by $\underline{\underline{\underline{\widehat{Q}}}}_{EMD}$ the one minimising the EMD.

That is:

$$\underline{\underline{\underline{\widehat{Q}}}}_{RMSE} \in \underset{\underline{\underline{\underline{\widehat{Q}}}}}{\text{Argmin}}\ RMSE(\underline{\underline{\underline{\widehat{Q}}}}), \tag{5.42a}$$

$$\underline{\underline{\underline{\widehat{Q}}}}_{EMD} \in \underset{\underline{\underline{\underline{\widehat{Q}}}}}{\text{Argmin}}\ EMD(\underline{\underline{\underline{\widehat{Q}}}}). \tag{5.42b}$$

As an example, Figure 5.3a illustrates a one dimensional cut of the evolution of the criterion RMSE, that is, as a function of γ_{TC}, the others being fixed. Similarly, Figure 5.3b illustrates the evolution of the EMD when varying γ_{TV} only It shows that for some values of $\gamma_{\cdot}$, the estimates have



(a) (b)

Figure 5.3: (a) RMSE as a function of $\gamma_{TC}\ \beta_{TC}$ for $\gamma_K = \gamma_{TV} = 0$ and $\gamma_C = \gamma_P = 1$. (b) EMD as a function of $\gamma_{TV}\ \beta_{TV}$ for $\gamma_K = 0$ and $\gamma_{TC} = \gamma_C = \gamma_P = 1$. Note that, on both figures, the first point on the left is for the zero value, to be distinguished from the rest drawn on the logarithmic scale.

minimal criteria while being lower than the criteria of the naive estimates $\underline{\underline{\underline{\widehat{Q}}}}^0$ and $\underline{\underline{\underline{\widehat{Q}}}}^1$. For the four estimates $\underline{\underline{\underline{\widehat{Q}}}}^0$, $\underline{\underline{\underline{\widehat{Q}}}}^1$, $\underline{\underline{\underline{\widehat{Q}}}}_{RMSE}$ and $\underline{\underline{\underline{\widehat{Q}}}}_{EMD}$, Table 5.1 presents the values of the RMSE, and EMD indicators, along with the values of f_{TC} (consistency with observed counts) and f_K (conformity with Kirchhoff's law). These two functions are chosen because they are the most important from a transport perspective. When applicable, the corresponding values for the $\gamma_{\cdot}$ are indicated.

Table 5.1: LOD matrix estimates: Two naive solutions compared to those minimising RMSE and EMD

	γ_{TC}	γ_K	γ_{TV}	RMSE	EMD	f_K	f_{TC}
$\underline{\underline{\underline{\widehat{Q}}}}^0$				0.320	0.086	**0**	55
$\underline{\underline{\underline{\widehat{Q}}}}^1$				0.307	0.069	1142	84
$\underline{\underline{\underline{\widehat{Q}}}}_{RMSE}$	31.6	0.008	0.015	**0.239**	0.047	289	**1**
$\underline{\underline{\underline{\widehat{Q}}}}_{EMD}$	1	0.025	0.027	0.244	**0.045**	133	28

Table 5.1 demonstrates first, that results achieved with the algorithm outperform the naive solutions for the two indicators RMSE and EMD. This demonstrates that the idea of involving additional

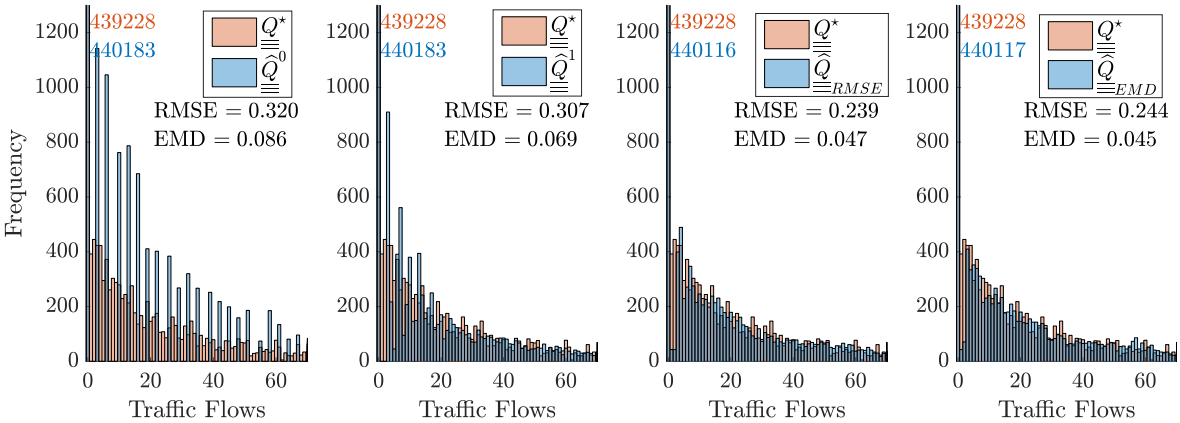


Figure 5.4: Distribution as histograms of the elements in $\widehat{\underline{\underline{Q}}}^0$, $\widehat{\underline{\underline{Q}}}^1$, $\widehat{\underline{\underline{Q}}}_{RMSE}$, $\widehat{\underline{\underline{Q}}}_{EMD}$ (in blue) superposed to the ground truth $\underline{\underline{Q}}^\star$ (in red).

information to the observed data helps indeed to reach better estimates and justifies that the proposed algorithm actually makes sense. Yet, a remaining question is whether implementing the algorithm is worth the effort. Table 5.1 illustrates that $\widehat{\underline{\underline{Q}}}^1$, performs poorly on the relevant transport indicators, in particular the Kirchhoff's law. It is therefore not a satisfying solution from a transport perspective. For its part, $\widehat{\underline{\underline{Q}}}^0$ performs well on the transport indicators, as by construction, it respects perfectly the Kirchhoff's law but its high EMD value indicates that its distribution is not in accordance with that of the ground truth. To investigate this point, Figure 5.4 represents the distribution of the elements in the four estimates as histograms, superposed to the ground truth. It appears indeed that the distribution of $\widehat{\underline{\underline{Q}}}^0$ does not fit the ground truth: it only contains multiples of the penetration rate value $\tilde{\alpha}$. Similarly for $\widehat{\underline{\underline{Q}}}^1$ (Figure 5.4(b)), multiples of the penetration rate can also be identified for low traffic flows. This is the consequence of Equation (5.41): elements in $\tilde{\underline{\underline{B}}}$ are integers only. The traffic values in $\widehat{\underline{\underline{Q}}}^0$ and $\widehat{\underline{\underline{Q}}}^1$ are not relevant and, in conclusion, neither of the naive estimates are satisfying solutions.

The question of which of both solutions $\widehat{\underline{\underline{Q}}}_{RMSE}$ and $\widehat{\underline{\underline{Q}}}_{EMD}$ is the best remains open and amounts to choosing the best γ, which remains an open question. Part of the answer relies on the importance of each term in the objective function: as one might expect, the performances with respect to the transport indicators are consistent with the weights applied to the corresponding functions: The solution $\widehat{\underline{\underline{Q}}}_{RMSE}$ is reached for higher γ_{TC} and, therefore, performs much better on the f_{TC} criterion while $\widehat{\underline{\underline{Q}}}_{EMD}$ has a relatively higher γ_K and hence satisfies the Kirchhoff's law better. Hence, the question of which solution to choose can depend on the reliability of the link counts: satisfying perfectly f_{TC} might not be relevant. Anyway, choosing between the two amounts to choosing the best γ, a question left for future work.

4.2.2 Impact of each Objective Function

While the results discussed above motivate to use Problem (5.25), the question of the importance of each objective function can be raised. Table 5.2(a) (resp. (b)) summarises the best RMSE (resp. EMD) when only the objectives indexed by the rows and column are used. Thus, diagonal elements

correspond to a single objective function and non diagonal elements involve the two objective functions indexed by the row and column. For example, element (1,2) of Table 5.2(a) corresponds to the best value of RMSE achieved for $\gamma_{TV} = \gamma_P = \gamma_C = 0$ and values of γ_K and γ_{TC} evaluated on a grid. For those tables, light-grey cells correspond to estimates that do not outperform the naive estimates, and darker grey cells, cases for which Algorithm 5.1 reached the maximum iteration limit without convergence.

Both tables lead to the conclusion that neither term gives satisfactory performance when used alone. The Poisson assumption and either the traffic counts function or the Kirchhoff's law are required to outperform the naive estimates. This means that one cannot obtain a good estimate of the traffic flows while there are not at least one term ensuring data fidelity, along with a regularisation term. The fact that the Poisson model is the most important can be expected as $\underline{\underline{\tilde{B}}}$ brings the most valuable information and confirms the importance of probe trajectories in solving traffic engineering problems.

Thus in a second step, the Poisson function has been systematically used with the projection and either one or two extra functions ($\gamma_P = 1$). Performances are presented in Tables 5.3(a) and (b). The indicator function corresponding to the projection on the convex set f_C has also been imposed ($\gamma_C = 1$) for three reasons: First its additional computational cost is negligible compared to the other functions, second, it accelerates the convergence speed of the algorithm by reducing the number of steps required while having little impact on the values of the criteria at convergence. Last, it corresponds to the weakest assumption: the total flow is larger than the measured probe trajectories.

In this second study, any estimate performs better than the naive ones but for the case where only the total variation (TV) is added. Depending on whether a minimum is sought for RMSE or for EMD, it is either the pair Kirchhoff's law and TV, or the pair Traffic Counts and TV, that complements best the Poisson assumption.

To conclude, to outperform the naive solutions, the Poisson model and a regularisation function are needed. Then, the addition of the TV brings the most improvement on the performance. In any case, best performances are achieved when all functions are involved (see Table 5.1). When all other functions are used, using the TV lowers the RMSE from 0.262 to 0.239 and the EMD from 0.067 to 0.045. The TV plays therefore an important role for the Minimisation Problem (5.25).

Yet, as a final remark, the additional computation cost in Algorithm 5.1 caused by the computation of H, its adjoint and the proximal of the ℓ_1-norm increases the computing time by four (convergence reached in ∼4 hours instead of ∼1 on a Core i7 laptop).

Table 5.2: Best RMSE and EMD when only one or two constraints are used

	RMSE				
	TC	K	TV	P	C
TC	0.99	0.98	0.99	0.30	0.67
K		1	1	**0.29**	0.68
TV			1	0.39	0.68
P				1	0.68
C					0.68

(a)

	EMD				
	TC	K	TV	P	C
TC	2.03	1.20	1.20	**0.07**	1.43
K		1.20	1.20	0.08	0.81
TV			1.20	0.19	0.78
P				1.20	0.81
C					0.81

(b)

Table 5.3: Best RMSE and EMD when $\gamma_C = \gamma_P = 1$, and with one or two additional constraints

	RMSE		
	TC	K	TV
TC	0.27	0.26	0.26
K		0.27	**0.25**
TV			0.35

(a)

	EMD		
	TC	K	TV
TC	0.067	0.067	**0.046**
K		0.069	0.069
TV			0.605

(b)

4.3 Lower Time Granularity

The number of users depends on the time granularity. It is therefore interesting, from a transport engineering perspective, to raise the question of whether the performances achieved here still hold for a lower number of users. In this section the main results are presented again for $N = 10^4$ users. This correspond to an average flow of 300 vehicles per link. In large city, this correspond to 5 to 10 minutes of traffic during peak hours. With such a low number of users we are reaching the limits of the model as 10^4 users corresponds to $\sim$4 users (that is, in average, 1.3 probe trajectories) per OD. Therefore the impact of the Poisson assumption, inferring information from $\underline{\underline{\tilde{B}}}$ decreases. Table 5.4 summarises the results in this case and shows that there is still a 14% improvement on the RMSE and a 30% improvement on the EMD. These results are very encouraging as even in the limit cases, the estimates achieved with the algorithm are an improvement with respect to the naive estimates.

Table 5.4: Best Achieved Results with $N = 10\,000$

	γ_{TC}	γ_K	γ_{TV}	RMSE	EMD	f_K	f_{TC}
$\underline{\underline{\hat{Q}}}^0$				0.398	0.021	**0**	13.3
$\underline{\underline{\hat{Q}}}^1$				0.396	0.017	128	**0.1**
$\underline{\underline{\hat{Q}}}_{RMSE}$	1.78	0.25	0.027	**0.341**	0.013	10.7	9.8
$\underline{\underline{\hat{Q}}}_{EMD}$	1	0.45	0.026	0.342	**0.012**	5.8	13.7

5 Alternative Formulations of the Problem

Two preliminary fomulations of the LDO matrix estimation were proposed in [17] and in [20], as tentative formulation of the same problem and have led to find the most elaborate formulation presented above.

5.1 Optimisation on Bluetooth Penetration Factors

In [17], further detailed in Appendix C, the idea is that, given enough time over which observed traffic data are aggregated, every element in the LOD matrix $\underline{\underline{Q}}$ should be sampled in the partial one $\underline{\underline{\tilde{B}}}$. Thus, the aim is to directly estimate the penetration factors $\underline{\underline{\alpha}}$, of size $N_V \times N_V \times N_L$, that is, the inverse of the

penetration rates, per link, origin and destination. The variable $\underline{\underline{\underline{\alpha}}}$ is introduced here a supplementary variable such that

$$\underline{\underline{\underline{Q}}} = \underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}}. \tag{5.43}$$

And the problem we aim to solve, similarly to Problem (5.8), can be expressed as:

$$\underline{\underline{\underline{\hat{\alpha}}}} \in \underset{\underline{\underline{\underline{\alpha}}}}{\text{Argmin}} \left\{ \gamma_1 \mathscr{D}_1(\tilde{\alpha}, \underline{\underline{\underline{\alpha}}}) + \gamma_2 \mathscr{D}_2(\underline{\tilde{q}}, \sum_{ij\bullet} \underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}}) + \sum_k \gamma_k f_k(\underline{\underline{\underline{\alpha}}}) \right\}, \tag{5.44}$$

where $\mathscr{D}_1$ is a function aiming to limit the variability of $\underline{\underline{\underline{\alpha}}}$ around the global average penetration factor $\tilde{\alpha}$, $\mathscr{D}_2$ is a function to quantify the deviation between the observed traffic count and the estimates, very similar to that in Section 3.1.1, and f_k are functions modelling properties of the estimates: a domain of definition indicator function also very similar to the one in Section 3.1.3 and a Kirchhoff's law ensuring traffic conservation at nodes independently of the OD pairs.

In this version of the problem we implicitly assumed that $\underline{\underline{\underline{\tilde{B}}}}$ follow a Gaussian distribution of means $\tilde{\alpha}$. It appears however, from the literature, that assuming Poisson distributions when dealing with counting process is generally favoured to Gaussian ones. Moreover, this method does not estimate any traffic information whenever the value in $\underline{\underline{\underline{\tilde{B}}}}$ are zeros. In order to tackle those two issues, this problem has been reinterpreted with the objective to directly estimate the LOD matrix and to involve a Poisson distribution in a second preliminary work.

5.2 A simple Forward-Backward approach

In [20], a problem very close to the one presented above has been developed. Three major changes can be identified: First, the objective function, with very little change, has been interpreted directly as a function of $\underline{\underline{\underline{Q}}}$ rather than as a function of $\underline{\underline{\underline{\alpha}}}$. This allows for the estimation of the LOD matrix even where $\underline{\underline{\underline{\tilde{B}}}}$ has zero elements. Second, instead of assuming a Gaussian distribution of the Bluetooth sampling, a Poisson distribution is assumed. Third, the Kirchhoff's law, as written in Problem, involved the observed traffic counts to compute the flow going through a node. The aim however is that the estimate satisfies the Kirchhoff's law, independently of the observations. The Kirchhoff's law has thus been rewritten to involve the estimated traffic only. This problem is detailed further in Appendix D.

The two limits of this version of the minimisation problem are first, that the Kirchhoff's law only ensures the conservation of the traffic at each node independently of the OD pairs, and second, that the estimation of the element in $\underline{\underline{\underline{Q}}}$ with zero value in $\underline{\underline{\underline{\tilde{B}}}}$ is not preforming well. Another formulation of Kirchhoff's law tacking into account OD information is used in Section 3.1.4. The minimisation of traffic total variation over the graph (see Section 3.1.5), ensures that roads with similar characteristics have similar traffic, independently of the value in $\underline{\underline{\underline{\tilde{B}}}}$.

6 Conclusion

We have shown the relevance of link dependent origin destination matrix for traffic engineering problems: first, it combines the traffic demand and the assignment, second, it still contains the OD matrix and its estimation through an inverse problem is in the continuity of the classical OD matrix estimation problem. We have further formalised the LOD matrix estimation problem as a nonsmooth

convex optimisation problem with regularisation terms that we have explicitly devised. These regularisations quantify the deviations to the satisfaction of two types properties: First, the consistency between estimates and observations, second properties related to the infrastructure of the network. Moreover we have devised a proximal primal dual algorithm to solve this inverse problem. We have assessed the relative importance and the impact of each constraint. Notably, we have shown that the data fidelity term with the Bluetooth observations along with another regularisation term is required to outperform naive solutions and that the performances are best when the Bluetooth data are modelled with a Poisson distribution rather than with a Gaussian one. This also highlight that the LOD matrix estimation problem and probe trajectories only make sense together: With traffic counts, the traditional dataset used in previous research, as sole observations, we have not managed to reach satisfying performances. Once the regularisation terms stemming from the trajectories is involved, the addition of the Kirchhoff's law or of the traffic counts is enough to outperform the naive solutions. Yet it is only when all the regularisation functions are involved, and the total variation in particular, that best performances are achieved.

This work, along with the preliminary works, highlights how the design and the choice of the objective functions impacts the results. Thus, the problem of finding more efficient objective functions remains open. For example, combining traffic counts and turning fractions at intersections in a new data fidelity term would be a good future development. Moreover, this work evoked the question of time granularity for the estimation process by assessing the impact of the number of users on the performances. However, the estimation method proposed here is static. Thus, two research questions remains open: first, successive estimations with low time granularity would need time dependent regularisation to ensure consistency between estimates. Second, additional regularisations could model the evolution of the traffic with time. For example, implementing a Kalman filter similarly to what have been done on traffic counts based OD matrix estimation [91], or also with supplementary data (e.g., Bluetooth [208] or other sensors [209]). Last, an online algorithm as the one presented in [210, Section 5.2] would be very valuable for traffic management.

CHAPTER 6

Brisbane Case Study

Contents

1 From Toy Models to Large Real Case Studies **115**
1.1 Adapting the Datasets 115
1.2 Objective Function for the Real Case Study 116
1.3 Algorithm 118
2 LOD Matrix in Brisbane **118**
2.1 Size of the Problem 118
2.2 Estimation and Assessment of Brisbane LODM 120
2.3 Example of Traffic Map from LOD matrix 121
3 Conclusion **123**

IN this chapter, the contributions of this thesis are applied to the case study of Brisbane. We use the Bluetooth dataset, the traffic counts dataset and, the road network dataset presented in Chapter 3, for the period from the 25th and the 31st of July 2014. In Chapter 4, we have cleansed the Bluetooth dataset and retrieved the trajectories for this period. We will use here, subsets of these datasets (traffic counts and trajectories), for estimating the Brisbane LOD matrix over one interesting period, from a traffic engineering perspective: The morning peak hours from 6a.m. to 9a.m.. We choose to study traffic on the Tuesday 28th of July 2014, a generic weekday.

To estimate the Brisbane LOD matrix using the method developed in Chapter 5, the datasets presented in Chapter 3 need to be adapted: two characteristics of the real datasets prevent indeed a direct application: First, on real networks, only a fraction of links (or roads) are equipped with magnetic loops (e.g., of the order of a third in the Brisbane network of interest here). Second, in Chapter 5, a bijection between intersections and Bluetooth detectors was assumed. Yet, we have shown in Chapter 3 that, in real cases, Bluetooth detector scanning range might span multiple intersections and that every intersection is not necessarily covered by a Bluetooth scanner.

In the first part of this chapter, we will present how to implement the LOD matrix estimation procedure developed in Chapter 5, with real datasets. Then, in a second part, we will illustrate achieved results using traffic maps based on recovered LOD matrices.

1 From Toy Models to Large Real Case Studies

1.1 Adapting the Datasets

First, the problem of traffic counts being monitored on a subset of roads only can be circumvented by comparing observed and estimated counts on this subset only. Let us however emphasise that the goal still remains to estimate the LOD matrix $\underline{\underline{Q}}$ over all existing links in L. Thus, the subset of links with traffic counts, denoted $\widetilde{L}$, only impacts relationships involving measured traffic counts: the traffic count data fidelity term f_{TC} and the estimation of the penetration rate η.

Second, we propose to connect the set S of Bluetooth detectors to the road network with the addition of virtual links from each Bluetooth scanner to every intersection within its detection range. We set the weight (or length) of these links to 0. From now on, we denote $\mathscr{G} = (V, L)$, the graph representing the road network extended with those virtual links and with Bluetooth detectors. In particular we have $S \subset V$. Bluetooth detectors can now be considered as origin and destination points of the users and belong to the set of intersections V.

This change of graph definition appears after trajectories have been retrieved in Chapter 4. As a consequence, the virtual link from the first Bluetooth detector to the first intersection of each trajectory is added and similarly, the virtual link from the last detector to the last intersection is also added. By doing so, trajectories have for origin and destination a Bluetooth scanner.

We will present next how these two changes impact the definition of the objective function as presented in Chapter 5, Section 3.1.

1.2 Objective Function for the Real Case Study

1.2.1 Notations and Definitions

We propose here the estimation of detector-to-detector travel information, also referred to as *transcient* in [177]. The LOD matrix has for origins and destinations, points that are Bluetooth detectors. As a consequence, the LOD matrix $\underline{\underline{\underline{Q}}}$ and its sampled measure $\underline{\underline{\underline{\tilde{B}}}}$ are now of size $N_S \times N_S \times N_L$, where N_S is the number of Bluetooth detectors.

Let us denote by $\widetilde{L} \subseteq L$ the subset of links on which traffic counts are available. We define $\underline{\delta_{\widetilde{L}}}$ the Kronecker delta vector of size N_L such as

$$(\forall l \in L) \quad \left(\delta_{\widetilde{L}}\right)_l = \begin{cases} 1 & \text{if } l \in \widetilde{L}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

The traffic count variable $\underline{q}$ is of size N_L and, by definition

$$\underline{\tilde{q}} = \underline{\delta_{\widetilde{L}}} \circ \underline{\tilde{q}} \tag{6.2}$$

Last, the definitions of the incidence and of the *excidence* matrices remain unchanged (*cf.* Equation (5.1)).

1.2.2 Traffic Count Data Fidelity f_{TC}

With the above definitions, Equation (5.9) proposed in Chapter 5, Section 3.1.1 becomes:

$$f_{TC}(\underline{\underline{\underline{Q}}}) = \left\| \underline{\delta_{\widetilde{L}}} \circ \left(\underline{\tilde{q}} - \sum_{ij\bullet} \underline{\underline{\underline{Q}}} \right) \right\|^2 . \tag{6.3}$$

1.2.3 Poisson Bluetooth Sampling Data Fidelity f_P

Second, for the consistency with Bluetooth measures, we assume the sampling follows a Poisson model of parameter the origin-destination penetration rate. Yet, as seen in Chapter 5, Section 3.1.2, this information is not available and could be estimated with the link penetration rate $\underline{\tilde{\eta}}$ (*cf.* Equation (5.10)).

Here however, this link penetration rate can only be estimated on links for which traffic counts are available, therefore we propose instead the use of the global penetration rate $\overline{\eta}$ defined as

$$\overline{\eta} = \frac{\Sigma\left(\underline{\delta_{\widetilde{L}}} \circ \Sigma_{ij\bullet} \underline{\underline{\underline{\tilde{B}}}}\right)}{\Sigma\left(\underline{\delta_{\widetilde{L}}} \circ \underline{\tilde{q}}\right)} \tag{6.4}$$

Equation (5.11) becomes here:

$$f_P(\underline{\underline{\underline{Q}}}) = \sum_{ijl} \psi\left(B^l_{ij}, \overline{\eta} Q^l_{ij}\right) \tag{6.5}$$

1.2.4 Consistency Constraint f_C

Third, the term ensuring data consistency, presented in Chapter 5, Section 3.1.3, remains unchanged, but for the fact that the LOD matrix is now defined in S instead of V. Thus the convex set C in Equation (5.13) becomes:

$$C = \{\underline{\underline{Q}} = \left(Q^l_{ij}\right)_{(ijl)\in S\times S\times L} \in \mathbb{R}^{N_S\times N_S\times N_L} \mid Q^l_{ij} \geq B^l_{ij}\}, \tag{6.6}$$

with corresponding convex function, similarly to Equation (5.14), the indicator function ι_C:

$$f_C(\underline{\underline{Q}}) = \iota_C(\underline{\underline{Q}}) = \begin{cases} 0 & \text{if } \underline{\underline{Q}} \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{6.7}$$

1.2.5 Kirchhoff's Law f_K

Here, sources and destinations of the flows are Bluetooth detectors only, while traffic do not go through these detectors. This property permits to simplify the expression of Kirchhoff's law (Equation (5.15)), by considering intersections only (and not Bluetooth detectors) and all the links, including the virtual ones :

$$(\forall k \in V\backslash S, \quad \forall (i,j) \in S\times S) \quad \sum_l E^l_k Q^l_{ij} = \sum_l I^l_k Q^l_{ij} \tag{6.8}$$

where $\underline{\underline{I}}$ and $\underline{\underline{E}}$ are the incidence and *excidence* matrix of the graph (with Bluetooth detectors and virtual links). With this simplified Kirchhoff's law, Equation (5.19) becomes the following convex function:

$$f_K(\underline{\underline{Q}}) = \sum_{\substack{ijk \\ i,j\in S\times S \\ k\in V\backslash S}} \left(\sum_l \left(I^l_k - E^l_k\right) Q^l_{ij}\right)^2. \tag{6.9}$$

1.2.6 Total Variation f_{TV}

Finally, the total variation as implemented in Chapter 5, Section 3.1.5, involved adjacent origins and destinations. Here, this notion is not defined for the Bluetooth detectors: the virtual links added to the network connect Bluetooth detectors to intersections. Consequently, Bluetooth detectors cannot be adjacent.

We propose instead, to compute the shortest paths between all pairs of detectors. To limit the number of origin destination pairs involved in the total variation term, a threshold distance of $300m$ is used. Any pair of detectors whose shortest path is longer than this threshold is not taken into account in total variation penalisation.

Then, total variation can be expressed as:

$$f_{TV}(\underline{\underline{Q}}) = \sum_{\substack{i,i'\in S\times S \\ i\neq i' \\ d_{i,i'}<300m}} \sum_{\substack{j,l \\ j\neq i,i'}} \omega_{ii'} |Q^l_{ij} - Q^l_{i'j}| + \sum_{\substack{j,j'\in S\times S \\ j\neq j' \\ d_{j,j'}<300m}} \sum_{\substack{i,l \\ i\neq j,j'}} \omega_{jj'} |Q^l_{ij} - Q^l_{ij'}| \tag{6.10}$$

where d_{ij} is the length of the shortest path from detector i to detector j and ω_{ij} is a weight defined as

$$(\forall (i,j) \in S\times S; \quad i \neq j) \quad \omega_{ij} = e^{-\frac{d_{ij}}{d_0}}. \tag{6.11}$$

where d_0 is $300m$.

We define the matrix $\underline{\underline{J}}$ of size $N_P \times N_S$, where N_P is the number of detector pairs in P with shortest path shorter than $300m$ as,

$$(\forall (i,j,p) \in S \times S \times S^2;\ p = i \cdot N_S + j) \qquad \begin{aligned} J(p,i) &= \begin{cases} -\omega_{ij} & \text{if } d_{ij} \leq 300m \\ 0 & \text{otherwise,} \end{cases} \\ J(p,j) &= \begin{cases} +\omega_{ij} & \text{if } d_{ij} \leq 300m \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{6.12}$$

Total variation can hence be expressed as

$$f_{TV}(\underline{\underline{\underline{Q}}}) = \sum_l \|\underline{\underline{J}}^\top \underline{\underline{Q}}^l\|_1 + \sum_l \|\underline{\underline{J}}^\top (\underline{\underline{Q}}^l)^\top\|_1, \tag{6.13}$$

(*cf.* Equation (5.24)). $\underline{\underline{Q}}^l$ models the l-th extracted matrix from $\underline{\underline{\underline{Q}}}$. Its dimension is thus $N_S \times N_S$.

1.3 Algorithm

With the above definitions of the five terms f_{TC}, f_P, f_C, f_K and, f_{TV}, the optimisation problem is similar to that of Problem (5.25) and can therefore be solved by Algorithm 5.1.

2 LOD Matrix in Brisbane

2.1 Size of the Problem

Brisbane City on the 28th of July 2014 had 576 Bluetooth detectors. The simplified road network of the region covered by these detectors is composed of 8 900 intersections and 18 300 links. Restricting the study to the city centre closest 430 Bluetooth detectors, such figures drop down to 1 800 intersections and 3 950 links.

There are 3 030 identified counting sites composed of around 8 000 magnetic loops on the original road network. We have shown in Chapter 3, Section 1 how we have simplified the networks. During this simplification step, when links are concatenated, the resulting link get the average of the counts (average for links with traffic count information only). When two links have same origin and destination, one is removed and the remaining link get the sum of the counts. We end up with 1 430 links (36% of the links) with traffic count value.

Once virtual links and Bluetooth detectors are considered, one end up with a graph $\mathscr{G} = (V, L)$ of size: $N_V = 2\,230$, $N_L = 5\,370$, $N_S = 430$, $N_{\tilde{L}} = 1\,430$, and $N_P = 1\,190$ (pairs of Bluetooth detectors with path shorter than 300m).

For the 6 a.m. to 9 a.m. time interval, the traffic has the following characteristics:

- The Bluetooth LOD matrix is composed of 39 100 trajectories.
- The cumulated number of traffic counts is 3 252 172.
- $\overline{\eta}$, the Bluetooth OD penetration rate, computed as per Equation (6.4) is 0.21.
- The total number of cars is unknown.

Figure 6.1 illustrates the traffic count values for roads in $\widetilde{L}$ for the morning peak hours and Figure 6.2 presents, for one OD (Brisbane CBD to Moorooka), the road traffic as recovered from the Bluetooth data.

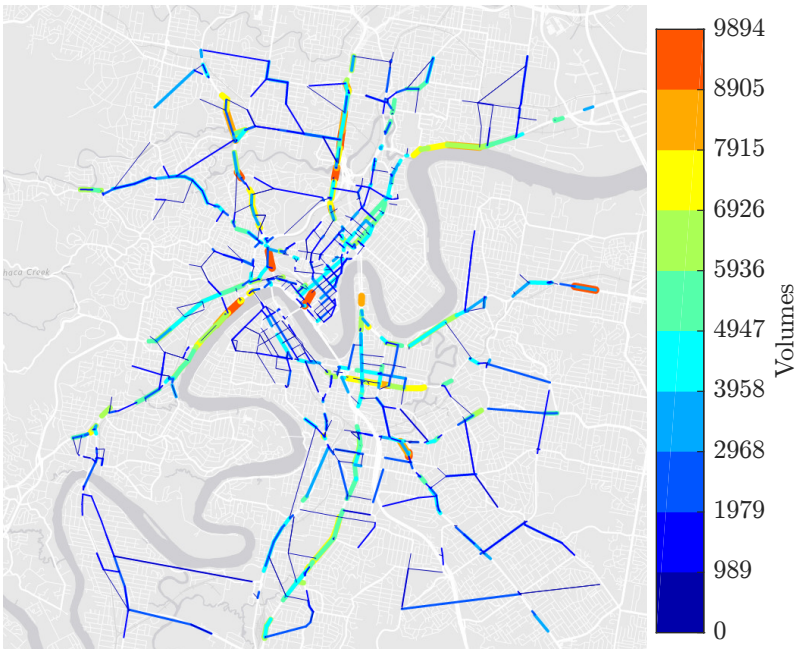


Figure 6.1: Brisbane measured traffic counts on the simplified network for the area of study. Only 36% of the links have non-zero values.

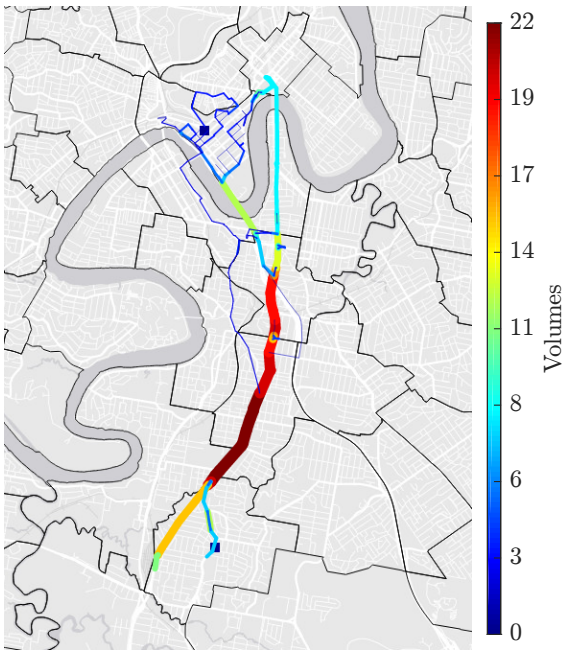


Figure 6.2: Bluetooth LOD flows for one specific OD pair (Brisbane CBD to Moorooka). Color and width of the roads are proportional to the volume of cars.

2.2 Estimation and Assessment of Brisbane LODM

For this case study, the ground truth is not available, therefore the assessment metrics RMSE and EMD used in Chapter 5 cannot be used. Yet, achieved solutions can be compared between each other using other indicators: the values of each term of the objective function are indicators of the solution consistency. In Table 6.1, we compare these indicators for the naive solution $\widehat{\underline{\underline{Q}}}^0$ (*cf.* Equation (5.40)) and two solutions with $(\gamma_\cdot/\beta_\cdot)$ values corresponding to the parameters obtained in Chapter 5, Section 4 minimising the RMSE for both simulated case study (with $N = 100\,000$ and $N = 10\,000$ users). The naive solution $\widehat{\underline{\underline{Q}}}^1$ (*cf.* Equation (5.41)) can not be computed in this real case study: it requires an estimation of the link penetration factor, which is not accessible as traffic count values are known on a subset of roads only. An estimation of $\widehat{\underline{\underline{Q}}}^1$, on this subset of roads only, would not be consistent from a traffic engineering perspective: in particular, the links of $\widetilde{L}$ are not always connected (*cf.* Figure 6.1) and the Kirchhoff's law would not be defined.

For the three estimates, we compare, in addition, the total number of users of the LOD matrix, computed thanks to the two relationships in Equation (5.3): the total number can be obtained either, by counting cars leaving from every origin or, by counting cars arriving at every destination. This leads to two estimations of the total number of cars $N_{(Or)}$ and $N_{(Dest)}$. If the Kirchhoff's law is perfectly satisfied, one should have $N_{(Or)} = N_{(Dest)}$.

Table 6.1: LOD Matrix Estimates Comparison. $\gamma_P = \gamma_C = 1$.

	γ_{TC}	γ_K	γ_{TV}	f_{TC}	f_K	f_P	f_{TV}	$N_{(Or)}$	$N_{(Dest)}$
$\widehat{\underline{\underline{Q}}}^0$				162	0	0.0025	19 050 400	186 181	186 181
$\widehat{\underline{\underline{Q}}}_1$	17.78	0.0128	0.0212	54	922	12201	13 904 421	162 090	165 654
$\widehat{\underline{\underline{Q}}}_2$	1.78	0.228	0.0377	134	53	76976	12 906 754	156 401	159 911

The solution $\widehat{\underline{\underline{Q}}}^0$ performs well compared to the other solutions on the indicators f_{TC}, f_K and f_P. This is to be expected as, by construction, it is based on measured Bluetooth trajectories (hence the perfect satisfaction of Kirchhoff's law), multiplied by the average penetration factor, hence the good performance on f_P. This global penetration factor is computed thanks to the measured traffic count hence the good performances on f_{TC}. Yet, we have shown in Chapter 5 that this solution is not satisfactory: it only estimates flows on roads where Bluetooth sample are accessible. Moreover, the estimated flows are, by construction, restricted to multiples of the global penetration rate only (*cf.* Figure 6.3a). The allowed values taken by traffic flows are limited and the estimation of low value OD flows is not efficient. As a consequence, the estimated number of users on the road networks is higher than for other solutions.

The other estimates, denoted by $\widehat{\underline{\underline{Q}}}_1$ and $\widehat{\underline{\underline{Q}}}_2$ represent trade-offs between the satisfaction of the four relaxed properties from which the objective function was built (the projection is always satisfied). Each weight $\gamma_\cdot$ represents the importance accorded to the satisfaction of the property it is linked to. For example $\widehat{\underline{\underline{Q}}}_2$ has a relatively higher γ_K value and satisfies thus the Kirchhoff's law better than $\widehat{\underline{\underline{Q}}}_1$. To the opposite, $\widehat{\underline{\underline{Q}}}_1$ is performing better on f_{TC} and f_P values.

As ground truth is not available here, it is not possible to assure which of the solutions is the best, yet we believe that estimating traffic flows on the whole network, rather than on sampled trajectories

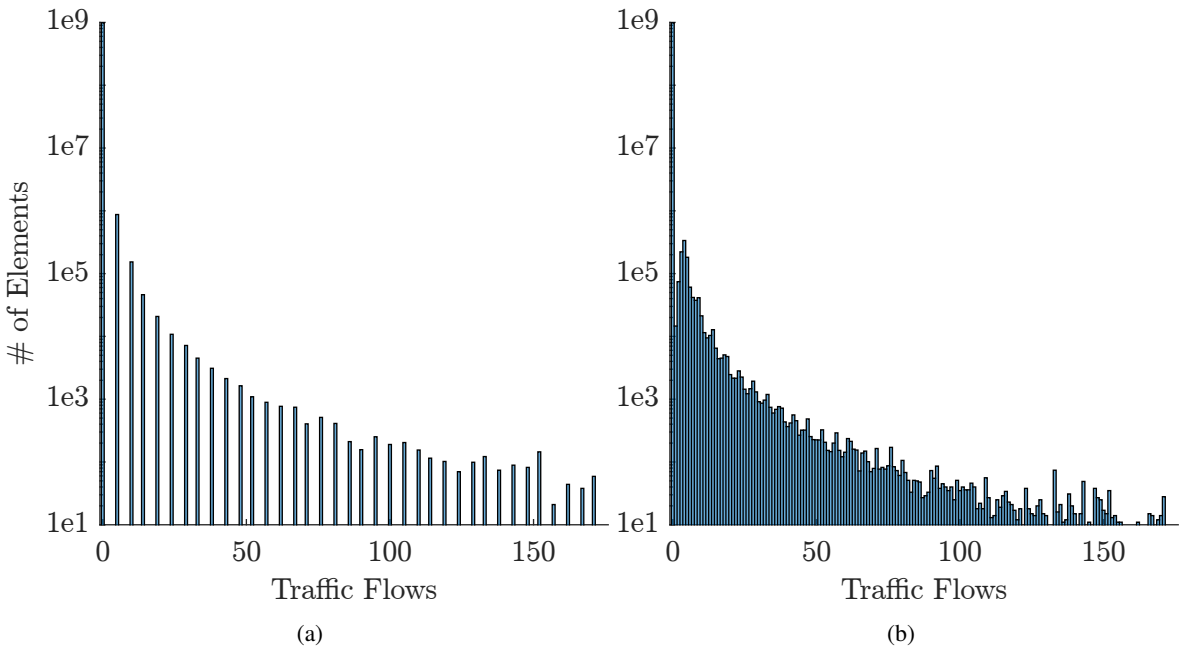


Figure 6.3: Distribution of traffic flows in the LODM, (a) for $\underline{\underline{\underline{\widehat{Q}}}}^0$, (b) for the estimated LOD matrix $\underline{\underline{\underline{\widehat{Q}}}}_1$. The Y-axis is logarithmic.

only, brings a more interesting insight on the traffic in the city. In the following, we will use the solution $\underline{\underline{\underline{\widehat{Q}}}}_1$ to illustrate traffic information inferred from the LOD matrix to Brisbane traffic.

2.3 Example of Traffic Map from LOD matrix

First, link volumes for the whole road network can be mapped, similarly to Figure 6.4. Compared to measured traffic counts in Figure 6.1, the flows are more continuous and consistent on adjacent roads.

For further illustration, one can choose a road segment and extract the corresponding OD matrix to get a better understanding of road usage. In Figure 6.5, the Bradfield Highway Bridge is selected (in magenta) and the the twenty most important OD flows have been represented. For a more readable map, OD flows are aggregated by Statistical Local Areas (SLA) of level 2[1]. This figure shows that an important fraction of cars uses the Bradfield Highway Bridge for crossing Brisbane: the OD flow leaving from the northernmost SLA region to the southernmost one, is the most important flow. A second identifiable behaviour is that an important amount of users in the eastern part of Brisbane use this bridge rather than the Gateway motorway bridge, at the far east of the river, even though it is not the fastest path. In fact, this bridge is the easternmost toll-free bridge and users might prefer it to non toll-free alternative.

Along another line, given one OD pair, it is possible to extract the corresponding road usage. In Figure 6.6, two SLA regions are selected (Brisbane CBD to Moorooka) and traffic volumes on each link, for this OD only, are represented proportionally with color and width. If this figure were plotted

[1]The Statistical Local Area (SLA) is an Australian Standard Geographical Classification (ASGC) defined area. http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2901.0Chapter23002011

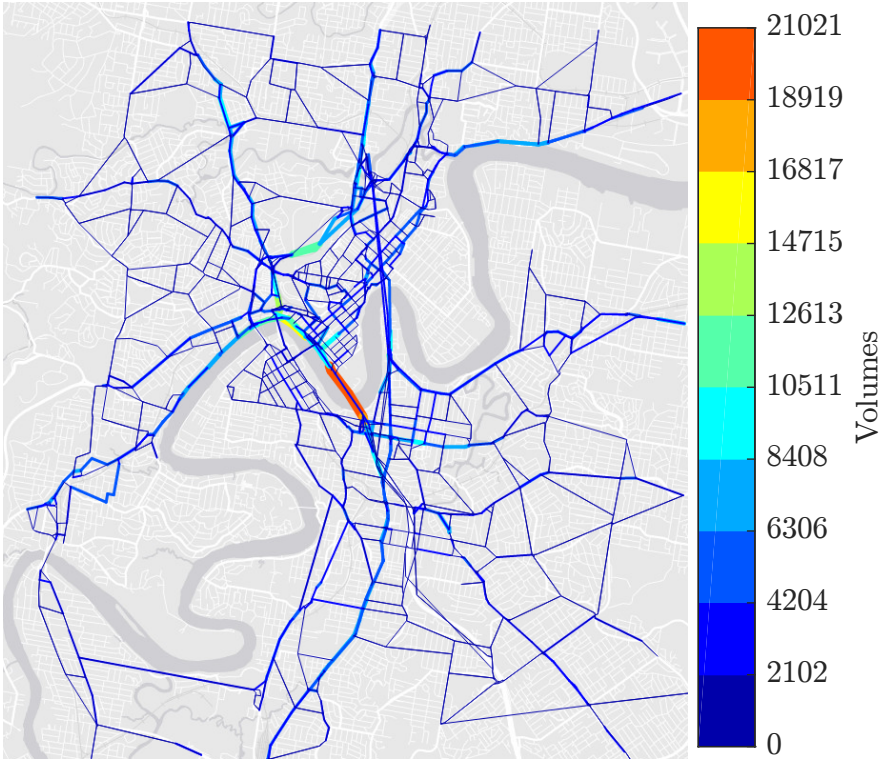


Figure 6.4: Brisbane Traffic Counts on the simplified network for the area of study.

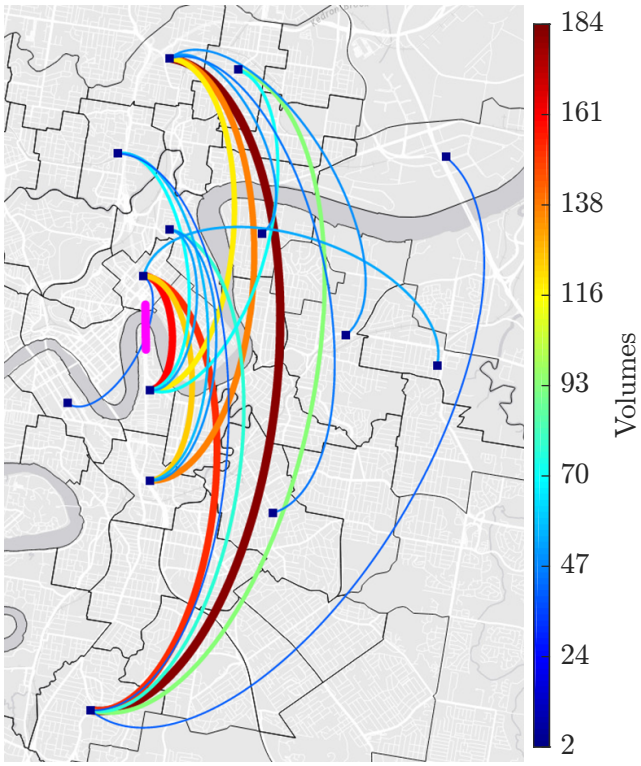


Figure 6.5: Origin-Destination flows of vehicles using the Bradfield Highway Bridge (North to South direction). The bridge is the road highlighted in magenta. Width and color of the semi-ellipses are representative of OD volumes. Only the twenty largest OD volumes are represented.

for $\widehat{\underline{\underline{Q}}}^0$, it would be exactly similar to Figure 6.2 with all the flows multiplied by the global penetration rate 4.76. In In Figure 6.6 however, the roads have been multiplied by different penetration factor. For example, the largest flows have been multiplied by less than 4 and the small traffic flows in South Bank (the SLA, west and across the river compared to the CBD) by a factor 3 only. This illustrates the penetration rate dependency to OD pairs and to chosen paths. Some artefacts also appear on the map: some isolated links have non-zero traffic flow. These artefacts stems from the term f_{TC} that will favour non zero flows on links where observed traffic counts are high. A stronger weight on the Kirchoff's law term in the objective function would remove such artefacts.

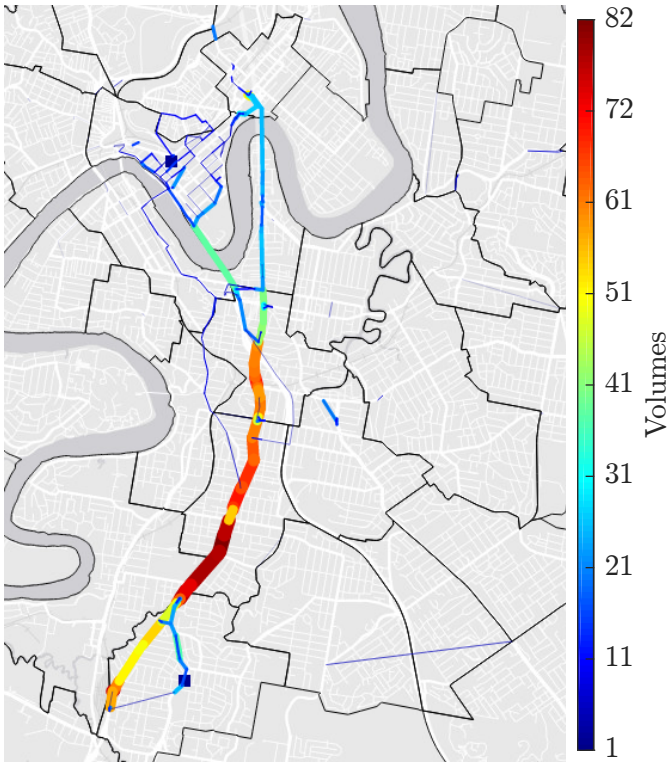


Figure 6.6: Road volumes for cars driving from CBD to Moorooka in Brisbane.

3 Conclusion

In this chapter, we have shown how to implement the method for LOD matrix estimation to a real large size case study. After a few modifications to the ideal model formulated in Chapter 5, the present chapter has shown that the method is suitable for application to large network ($430^2 \simeq 185\,000$ OD pairs, 5 370 links) and in urban context (here Brisbane City), filling thus the major research gaps identified at the beginning of this PhD.

This chapter has also illustrated some possible uses of LOD matrix for traffic analysis and has highlighted its potential: Without any additional model, OD flows and road usage can be jointly analysed and represented, giving thus a more detailed understanding of the traffic than the traditional OD matrix. We have presented here, some traffic information interpretation made possible by the access to the LOD matrix. Yet the aim has not been a complete assessment of such applications and many others could be developed: traffic evolution by comparing successive estimations, commuter traffic analysis by comparing traffic in the morning peak hours with the evening peak hours, ...

In addition, we tried to provide some insights, based on the analysis of the simulated case study of Chapter 5, on why the estimated solutions with Algorithm 5.1 is more satisfactory than the naive estimates. Yet the ground truth being unavailable, a more detailed analysis of the efficiency of the method when using real data would be a good future contribution. In particular, the question of selecting the most adapted $\gamma_{\cdot}$ values remains an important unanswered question.

CHAPTER 7

Conclusion and Perspectives

THE works presented in this manuscript detail the steps for using Bluetooth data as a complementary dataset for Transport Engineering problems. Starting from raw data, each chapter, capitalising on the previous ones, proposes new applications for the Bluetooth data. Along this manuscript, we have shown how the Bluetooth data can be used for basic analysis of traffic conditions, then, how to retrieve trajectory information from the Bluetooth data, and finally, how the Bluetooth data can be combined with other datasets to resolve an extended version of the classical OD matrix estimation problem. Thus, this thesis has also been the occasion to reformulate the classical OD matrix estimation problem by combining the opportunities offered by new technologies with recent advances in Transport Engineering and in Signal Processing.

In the continuity of the efforts from the research community to understand and to identify how to make the most of new technologies within Transport Engineering problems, one of the major contributions of this thesis is the development of a process to recover vehicle trajectories from Bluetooth point-measures. The numerous biases of the Bluetooth data make challenging to estimate the precise location both in time and space of detected devices: The large detection radius and the inquiry cycle duration lead to several possibilities for travel patterns. In addition, missed detections are not rare events and made the retrieval of travel pattern even more challenging. Last, Bluetooth devices are not specific to one mode of travel only and any user of Bluetooth equipped device might be detected independently of its usage of the network.

To solve this problem, we performed a thorough analysis of Bluetooth data. This led us to propose three algorithms: The first one for extracting trip information, consisting of origin, destination and intermediate detections for each detected device, the second one for proposing trajectories out of the recovered trip informations and the third one for discriminating motorised from non-motorised modes of travel. The second algorithm, the *spatially constrained algorithm*, has been developed as a new extension of traditional shortest path algorithms and computes the shortest path from first to last scanning area passing through all intermediate ones. We have illustrated its efficiency: tested on two case studies, the resulting trajectories were at 84% corresponding to the ground truth.

Probe trajectories are more and more used as primary transport datasets for traffic applications, e.g., traffic conditions analysis or route choice model calibration [211]. The access to sets of probe trajectories is therefore more and more critical to transport engineers. The works presented in this

thesis can be used as a detailed handbook for retrieving Bluetooth trajectories but the theoretical contributions actually extend to many other technologies collecting point-measures. This thesis provides thus valuable alternatives to the traditional source of trajectories: the GPS technology.

In light of the importance that trajectories are acquiring within the Transport field, further research on this subject would be valuable. Among the trails to develop the present contributions is the further development of the trajectory recovery process. For example, combining the present works with those of Feng, Sun, and Chen (2015) [106], more complex algorithms could be developed, taking into account constrained shortest paths, probabilities of detection, and mode of travel to assess the likelihood of a trajectory to correspond to the observations. Another interesting development would be, in tandem with BCC's engineers, to evaluate the impact of inquiry cycle duration change on the quality of the data collected.

A second major contribution of this thesis has been the extension of the concept of origin destination matrix to that of link dependent origin destination matrix.

While traditionally, demand estimation and traffic assignment have been treated as two separated problematics, this new concept of LOD matrix inherently combines both the demand information and the assignment of the users on the network. The definition of this new concept has been motivated by the access to probe trajectories, that directly sample LOD matrix. First we have demonstrated that, the dimension of the LOD matrix estimation problem is not larger than the dimension of the classical OD matrix estimation problem: The OD matrix estimation problem have some of its complexity hidden by the separation between the inverse problem aiming for the estimation of the OD matrix, with the assignment problem, which permits the comparison between OD matrix estimates and traffic observations. Moreover, we have shown that a set of probe trajectories could directly be interpreted as a sample of the LOD matrix and therefore that the LOD matrix estimation problem could be solved by means of nonsmooth convex optimisation. We formulated the problem as a regularised inverse problem, with some regularisation functions quantifying the fidelity between estimates and observed data and others favouring estimates satisfying transport properties on the network.

This minimisation problem has then been solved thanks to a proximal primal dual algorithm that we adapted for this problem. The impacts of each of the regularisation functions have been analysed and we have proven that the regularisations proposed, helps to improve the performance of the estimates. In particular, this analysis has highlighted the importance of probe trajectories for this estimation problem: the Poisson model assumed between observed trajectories and estimates has proven to be the most important of the proposed regularisation functions. This justify once more why trajectories are becoming more and more important in Transport Engineering problems. However important the trajectories are, best performances are achieved when all the proposed regularisations are jointly minimised. Notably, the traditional comparison to traffic counts is still an important part of this estimation problem, in the continuity with over forty years of research on OD matrix estimation.

The successive formulations of the LOD matrix estimation problem proposed in this manuscript have highlighted the importance of the regularisation function design. It is more than likely that other regularisation functions involving additional transport data would prove valuable. For example, if available, combining turning fractions at intersections with traffic counts could be a good development. Another research question on this problem is on the development of new algorithms for finding optimal estimates. These new algorithms could either aim for improved computation speed or for online estimation, which would be of great interest for traffic management. Last, the question of the dynamic LOD matrix has not much been discussed. By varying the number of users considered in our simulated case study, we have shown that the choice of the time granularity for our static approach

is flexible. However, the question of consistent estimates of successive LOD matrices has not been developed and could be the matter of future research. For example, one could involve an additional regularisation function favouring temporal correlations between successive estimates. Such dynamic LOD matrix would prove very valuable for traffic dynamic analysis and for traffic forecasts.

Last, an additional contribution of this work has been to illustrate the interest of all these methods by applying the whole expertise to the real case study of Brisbane. The methods for LOD matrix estimation had, before this contribution, only be tested on simulated case studies and the confrontation to real data is an important part of the research procedure. The ground truth in such a case study being not available, we proposed to check that the results were consistent with what can be expected in the city. We also illustrated how the LOD matrix can directly bring meaningful information on road usage and on travel pattern. It demonstrates that this process can actually be applied on real road networks, for large cities and should convince transport engineers they could benefit from the implementation of such tools.

This manuscript could not end without bringing some of my personal perspectives on a question that was raised frequently when I presented my work: The ethical question behind the detection and identification of personal Bluetooth devices. Of course, the easy answers I end up using, partly to avoid the debate maybe, were based on the arguments that, first, Brisbane City Council had never hidden the fact that they were installing Bluetooth scanners. Second, the MAC IDs are encrypted and the encryption tables are not accessible. Even if MAC IDs were not encrypted, it would be very hard to match names and MAC IDs: For such a match, one would first need to verify that the Bluetooth chips manufacturers have kept track of matches between, chips, devices and customers. Even so, the manufacturers would unlikely be willing to freely share this information with third parties. In addition, the very high number of manufacturers would make such requests unachievable anyway. Q.E.D., privacy and anonymity are not jeopardised.

In truth, the answer is not that simple. We have shown in Chapter 3, Section 3.2.5 that only five manufacturers represents 50% of the Bluetooth dataset. In addition, researches have shown that even without individual identification capabilities, meta-data can eventually bring sufficient information for *a posteriori* identification. de Montjoye, Hidalgo, Verleysen, and Blondel (2013) [212] have shown that from a database with half a million individuals located thanks to their mobile phones over 15 months, one had only to know four positions of a given person in order to identify that individual with a 95% probability. In comparison, it will soon be 10 years since the Bluetooth scanner network in Brisbane started to collect data on millions of users. Claiming that this data do not jeopardize the privacy and the anonymity of users would not be an honest assertion. It is were the critical element of this debate takes place: Independently of the effort to anonymise the data, the only way to ensure that the privacy and anonymity of the users is kept relies on the capacity to enforce that the uses of the data stay at an aggregated level.

This is why it is crucial, from my perspective, that data are collected, managed and analysed by a public stakeholder rather than by private companies. Regarding this specific point, Brisbane City Council made it clear that the data would not be freely distributed and the Smart Transport Research Centre could have access to this data only as a public research laboratory and with promises of applications at aggregated levels only: OD and LOD matrices, travel time, and turning fractions among others.

Finally, all the advances in Transport made possible by those data can have direct benefits to the Brisbane community (e.g., travel time along roads display, optimised transport infrastructure, traffic plans). To the opposite, private company, like Google, which manage to have access to the

location and movement of massive sets of users (e.g., by means of the location feature on Android) do not make the communities benefits from these collections of information. Instead of sharing their data, with public stakeholders for example, the data are monetised to partly finance the services they provide. Thus, as a last element of debate, the question is raised, whether one prefers to share personal data freely with public stakeholders for a relative certitude of anonymity and privacy and an expected return as benefits for the community, or whether personal data are more destined to become an alternative to money exchanges.

In any case, these ethical questions give an additional reason for supporting public and academic stakeholders, both for the data collection and for the data analysis. The ethical challenges can not be alleviated by technical solutions, and thus, the tools and expertise for gathering, managing and analysing personal data should be public and transparent.

APPENDIX A

Algorithm for Network Simplification

Algorithm A.1 Procedure for Simplifying the Road Network

```
Lmap ← 1 : |L*|
L ← L*
while a modification happened during previous iteration do
    for all v ∈ V*/{M_r}_{s∈S} do
        L_in ← find(L(:,2) = v)
        L_out ← find(L(:,1) = v)
        if L_in = ∅ or L_out = ∅ then
            L(L_in,:) = [−1,−1,−1]
            L(L_out,:) = [−1,−1,−1]
            Lmap((L_in) = −1
            Lmap((L_out) = −1
        else if |L_in| = 1 and |L_out| = 1 then
            L(L_in,:) = [L(L_in,1), L(L_out,2), L(L_in,3) + L(L_out,3)]
            L(L_out,:) = [−1,−1,−1]
            Lmap((L_out) = L_in
```

```
        else if |L_in| = 2 and |L_out| = 2 then
            if L(L_in(1),1) = L(L_out(1),2) and L(L_in(2),1) = L(L_out(2),2) then
                L(L_in(1),:) = [L(L_in,1), L(L_out(1),2), L(L_in(1),3) + L(L_out(1),3)]
                L(L_out(1),:) = [-1,-1,-1]
                L(L_in(2),:) = [L(L_in(2),1), L(L_out(2),2), L(L_in(2),3) + L(L_out(2),3)]
                L(L_out(2),:) = [-1,-1,-1]
                Lmap((L_out(1)) = L_in(1)
                Lmap((L_out(2)) = L_in(2)
            else if L(L_in(1),1) = L(L_out(2),2) and L(L_in(2),1) = L(L_out(1),2) then
                L(L_in(1),:) = [L(L_in,1), L(L_out(2),2), L(L_in(1),3) + L(L_out(2),3)]
                L(L_out(2),:) = [-1,-1,-1]
                L(L_in(2),:) = [L(L_in(2),1), L(L_out(1),2), L(L_in(2),3) + L(L_out(1),3)]
                L(L_out(1),:) = [-1,-1,-1]
                Lmap((L_out(2)) = L_in(1)
                Lmap((L_out(1)) = L_in(2)
        else if |L_in| = 2 and |L_out| = 1 then
            if L(L_in(1),1) = L(L_out,2) then
                L(L_in(1),:) = [L(L_in(1),1), L(L_out,2), L(L_in(1),3) + L(L_out,3)]
                L(L_out,:) = [-1,-1,-1]
                Lmap((L_out) = L_in(1)
            else if L(L_in(2),1) = L(L_out,2) then
                L(L_in(2),:) = [L(L_in(2),1), L(L_out,2), L(L_in(2),3) + L(L_out,3)]
                L(L_out,:) = [-1,-1,-1]
                Lmap((L_out) = L_in(2)
        else if |L_in| = 1 and |L_out| = 2 then
            if L(L_in,1) = L(L_out(1),2) then
                L(L_in,:) = [L(L_in,1), L(L_out(1),2), L(L_in,3) + L(L_out(1),3)]
                L(L_out(1),:) = [-1,-1,-1]
                Lmap((L_out(1)) = L_in
            else if L(L_in,1) = L(L_out(2),2) then
                L(L_in,:) = [L(L_in,1), L(L_out(2),2), L(L_in,3) + L(L_out(2),3)]
                L(L_out(2),:) = [-1,-1,-1]
                Lmap((L_out(2)) = L_in
    for all l ∈ L do
        if L(l,1) = L(l,2) then
            L(l,:) = [-1,-1,-1]
            Lmap(l) = -1
        else if ∃e ∈ L; L(l,1) = L(e,1) and L(l,2) = L(e,2) then
            if L(e,3) < L(l,3) then
                L(l,:) = [-1,-1,-1]
                Lmap(l) = e
            else
                L(e,:) = [-1,-1,-1]
                Lmap(e) = l
Remove from L lines [-1,-1,-1] and update Lmap accordingly
Create V from nodes involved in L and V*
```

APPENDIX B

Expectation Maximisation Estimation

Contents

1 Generalities on the EM algorithm **131**
2 Two Binomial Mixture Model **134**
3 Two Gaussian Mixture Model **135**

1 Generalities on the EM algorithm

Let us assume a distribution defined as the mixture of K distributions.
Let p be the probability associated with the distribution and $(p_k)_{(k\in K)}$ the probability of each K distribution respectively.
Let $\Pi = (\pi_k)_{(k\in K)}$ denotes the proportion of each class (mixture coefficients).
Let $\Theta = (\theta_k)_{(k\in K)}$ denotes the parameters characterising each distribution.
Let $X = (x_i)_{(i\in[1,N])}$ be a set of N independent and identically distributed observations.
Then

$$p(x_i|\Pi,\Theta) = \sum_{k=1}^{K} \pi_k p(x_i|\theta_k) \tag{B.1}$$

or

$$p(X|\Pi,\Theta) = \prod_{i=1}^{N}\sum_{k=1}^{K} \pi_k p(x_i|\theta_k) \tag{B.2}$$

Let $Z = (z_i)_{(i\in[1,N])}$ an indicator of the class observations in X belong to (z_i being the class of x_i).
Then:

$$p(x_i,z_i|\Pi,\Theta) = \pi_{z_i}\ p_{z_i}(x_i|\theta_{z_i}) \tag{B.3}$$

$$p(X,Z|\Pi,\Theta) = \prod_{i=1}^{N} \pi_{z_i} p_{z_i}(x_i|\theta_{z_i}) \tag{B.4}$$

Thus, the log-likelihood of this model can be written as

$$\mathscr{L}(\Theta,\Pi|X,Z) = \log\left(p(X,Z|\Pi,\Theta)\right) = \sum_{i=1}^{N} \log\left(\pi_{z_i}\ p_{z_i}(x_i|\theta_{z_i})\right) \tag{B.5}$$

At this point, note that Z is a random variable of unknown distribution.
Let $\mathscr{Z}$ be the set of possible realisation of Z.
Let us also define $\hat{\Pi} = (\hat{\pi}_k)_{(k \in K)}$ and $\hat{\Theta} = (\hat{\theta}_k)_{(k \in K)}$ estimates of Π and Θ.
Then, Bayes rules ensures that:

$$p(z_i|x_i, \hat{\Pi}, \hat{\Theta}) = \frac{\hat{\pi}_{z_i}\ p_{z_i}(x_i|\hat{\theta}_{z_i})}{p(x_i|\hat{\Pi}, \hat{\Theta})} \tag{B.6}$$

Moreover

$$p(Z|X, \hat{\Pi}, \hat{\Theta}) = \prod_{i=1}^{N} p(z_i|x_i, \hat{\Pi}, \hat{\Theta}) \tag{B.7}$$

The idea behind the Expectation Maximisation problem is to maximise the expected log likelihood of the model, given an observation X and some estimated parameters $\hat{\Theta}$. Indeed, as Z is unobserved, one can only compute the expected log likelihood of the model over the possible value of Z, $Z \in \mathscr{Z}$ for then to maximise this expectation.
But first, let us derive the expression of the expected log likelihood (usually called, expectation step):

Expectation Step:

$$\begin{aligned}
Q(\Theta, \Pi, \hat{\Theta}, \hat{\Pi}) &= \mathbb{E}_{Z|X,\hat{\Pi},\hat{\Theta}}\left[\mathscr{L}(\Theta, \Pi|X, Z)\right] \\
&= \mathbb{E}_{Z|X,\hat{\Pi},\hat{\Theta}}\left[\log\left(p(X, Z|\Pi, \Theta)\right)\right] \\
&= \sum_{Z \in \mathscr{Z}} \log\left(p(X, Z|\Pi, \Theta)\right) p(Z|X, \hat{\Pi}, \hat{\Theta}) \\
&= \sum_{Z \in \mathscr{Z}} \sum_{i=1}^{N} \log\left(\pi_{z_i}\ p_{z_i}(x_i|\theta_{z_i})\right) \prod_{j=1}^{N} p(z_j|x_j, \hat{\Pi}, \hat{\Theta}) \\
&= \left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} \sum_{i=1}^{N} \log\left(\pi_{z_i}\ p_{z_i}(x_i|\theta_{z_i})\right) \prod_{j=1}^{N} p(z_j|x_j, \hat{\Pi}, \hat{\Theta}) \\
&= \left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} \sum_{i=1}^{N} \sum_{z=1}^{K} \delta_{z,z_i} \log\left(\pi_z\ p_z(x_i|\theta_z)\right) \prod_{j=1}^{N} p(z_j|x_j, \hat{\Pi}, \hat{\Theta}) \\
&= \sum_{z=1}^{K} \sum_{i=1}^{N} \log\left(\pi_z\ p_z(x_i|\theta_z)\right) \underbrace{\left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} \delta_{z,z_i} \prod_{j=1}^{N} p(z_j|x_j, \hat{\Pi}, \hat{\Theta})}_{=A}
\end{aligned} \tag{B.8}$$

$$
\begin{aligned}
A &= \left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} \delta_{z,z_i} \prod_{j=1}^{N} p(z_j|x_j,\hat{\Pi},\hat{\Theta}) \\
&= \left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} \delta_{z,z_i} p(z_i|x_i,\hat{\Pi},\hat{\Theta}) \prod_{j=1,i\neq j}^{N} p(z_j|x_j,\hat{\Pi},\hat{\Theta}) \\
&= \left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} p(z|x_i,\hat{\Pi},\hat{\Theta}) \prod_{j=1,j\neq i}^{N} p(z_j|x_j,\hat{\Pi},\hat{\Theta}) \\
&= p(z|x_i,\hat{\Pi},\hat{\Theta}) \left(\sum_{z_m=1}^{K}\right)_{(m=1..N)} \prod_{j=1,j\neq i}^{N} p(z_j|x_j,\hat{\Pi},\hat{\Theta}) \\
&= p(z|x_i,\hat{\Pi},\hat{\Theta}) \prod_{j=1,j\neq i}^{N} \underbrace{\sum_{z_i=1}^{K} p(z_j|x_j,\hat{\Pi},\hat{\Theta})}_{=1} \\
&= p(z|x_i,\hat{\Pi},\hat{\Theta})
\end{aligned}
\tag{B.9}
$$

Finally,

$$
\begin{aligned}
Q(\Theta,\Pi,\hat{\Theta},\hat{\Pi}) &= \sum_{z=1}^{K}\sum_{i=1}^{N} \log\left(\pi_z\, p_z(x_i|\theta_n)\right) p(z|x_i,\hat{\Pi},\hat{\Theta}) \\
&= \sum_{z=1}^{K}\sum_{i=1}^{N} \log(\pi_z) p(z|x_i,\hat{\Pi},\hat{\Theta}) + \sum_{z=1}^{K}\sum_{i=1}^{N} \log\left(p_z(x_i|\theta_z)\right) p(z|x_i,\hat{\Pi},\hat{\Theta})
\end{aligned}
\tag{B.10}
$$

Note that, in this equation, the first term only depends on Π but not on Θ while it is the opposite for the second term.

Maximisation Step: Now that a simpler formulation of the expected log likelihood has been derived, one can proceed with the maximisation step as follow:

$$\tilde{\Pi} = \underset{\Pi}{\text{Argmax}}\, Q(\Theta,\Pi,\hat{\Theta},\hat{\Pi}) \tag{B.11}$$

$$\tilde{\Theta} = \underset{\Theta}{\text{Argmax}}\, Q(\Theta,\Pi,\hat{\Theta},\hat{\Pi}) \tag{B.12}$$

The solution of Equation (B.11) is derived by solving the following system:

$$
\begin{aligned}
&\frac{\partial Q(\Theta,\Pi,\hat{\Theta},\hat{\Pi})}{\partial \Pi} = 0 \\
&\text{s.t.} \quad \sum \Pi = 1
\end{aligned}
\tag{B.13}
$$

To do so, we use the Langrangian multiplier λ to enforce the constraint, and Equation (B.13) becomes:

$$\frac{\partial Q(\Theta,\Pi,\hat{\Theta},\hat{\Pi}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)}{\partial \pi_k} = 0 \tag{B.14}$$

$$\sum_{i=1}^{N} \frac{1}{\pi_k} p(z_i = k|x_i,\hat{\Pi},\hat{\Theta}) + \lambda = 0 \tag{B.15}$$

Which leads to:

$$\pi_k = \frac{-1}{\lambda} \sum_{i=1}^{N} p(z_i = k|x_i,\hat{\Pi},\hat{\Theta}) \tag{B.16}$$

To find the value of λ, remind that $\sum_{k=1}^{K} \pi_k = 1$. Thus, Equation (B.16) also writes as:

$$\begin{aligned} 1 &= \sum_{k=1}^{K} \pi_k \\ &= \frac{-1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{N} p(z_i = k | x_i, \hat{\Pi}, \hat{\Theta}) \\ &= \frac{-N}{\lambda} \end{aligned} \tag{B.17}$$

And thus

$$\lambda = -N \tag{B.18}$$

Finally we have,

$$\tilde{\pi}_k = \frac{1}{N} \sum_{i=1}^{N} p(z_i = k | x_i, \hat{\Pi}, \hat{\Theta}) \tag{B.19}$$

or, with Eq. (B.6)

$$\tilde{\pi}_k = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{\pi}_k \; p_k(x_i | \hat{\theta}_k)}{p(x_i | \hat{\Pi}, \hat{\Theta})} \tag{B.20}$$

Equation (B.12) is solved similarly depending on the assumed distribution and therefore on the corresponding parameters in Θ.

2 Two Binomial Mixture Model

In the case of the mixture of two binomials distribution, we have $K = 2$, $\pi_2 = 1 - \pi_1$ and $\Theta = \{\theta_1, \theta_2\}$ the probabilities involved in each binomial law. Moreover, each observation in $x_i \in X$ is composed of the pair (g_i, h_i) the number of detections which occurred and the number of detectors on the trajectory respectively. Thus:

$$\forall k \in [1,2] \quad p_k(x_i | \hat{\theta}_k) = \binom{h_i}{g_i} \hat{\theta}_k^{g_i} (1 - \hat{\theta}_k)^{(h_i - g_i)} \tag{B.21}$$

Thus Equation (B.20) of the EM algorithm becomes:

$$\tilde{\pi}_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{\pi}_1 \; p_1(x_i | \hat{\theta}_1)}{\hat{\pi}_1 \; p_1(x_i | \hat{\theta}_1) + (1 - \hat{\pi}_1) \; p_2(x_i | \hat{\theta}_2)} \tag{B.22}$$

Solving equation (B.12), is equivalent to solving the following system of equations:

$$\begin{cases} \frac{\partial Q(\Theta, \Pi, \hat{\Theta}, \hat{\Pi})}{\partial \theta_1} & = 0 \\ \frac{\partial Q(\Theta, \Pi, \hat{\Theta}, \hat{\Pi})}{\partial \theta_2} & = 0 \end{cases} \tag{B.23}$$

For θ_1, this leads to:

$$\begin{aligned} \frac{\partial Q(\Theta, \Pi, \hat{\Theta}, \hat{\Pi})}{\partial \theta_1} &= \frac{\partial \sum_{z=1}^{K} \sum_{i=1}^{N} \log\left(p_z(x_i | \theta_z)\right) p(z_i = z | x_i, \hat{\Pi}, \hat{\Theta})}{\partial \theta_1} \\ &= \frac{\partial \sum_{i=1}^{N} \log\left(p_1(x_i | \theta_1)\right) p(z_i = 1 | x_i, \hat{\Pi}, \hat{\Theta}) + \sum_{i=1}^{N} \log\left(p_2(x_i | \theta_2)\right) p(z_i = 2 | x_i, \hat{\Pi}, \hat{\Theta})}{\partial \theta_1} \\ &= \sum_{i=1}^{N} p(z_i = 1 | x_i, \hat{\Pi}, \hat{\Theta}) \frac{\partial \log\left(p_1(x_i | \theta_1)\right)}{\partial \theta_1} \\ &= \sum_{i=1}^{N} p(z_i = 1 | x_i, \hat{\Pi}, \hat{\Theta}) \left(\frac{g_1}{\theta_1} - \frac{(h_i - g_i)}{(1 - \theta_1)} \right) \end{aligned} \tag{B.24}$$

And thus,

$$\tilde{\theta}_1 = \frac{\sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})\ g_i}{\sum_{i=1}^{N} p(1|x_i, \hat{\Pi}, \hat{\Theta})\ h_i} \tag{B.25}$$

Similarly, we have

$$\begin{aligned}\tilde{\theta}_2 &= \frac{\sum_{i=1}^{N} p(z_i = 2|x_i, \hat{\Pi}, \hat{\Theta})\ g_i}{\sum_{i=1}^{N} p(z_i = 2|x_i, \hat{\Pi}, \hat{\Theta})\ h_i} \\ &= \frac{\sum_{i=1}^{N} \left(1 - p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})\right) g_i}{\sum_{i=1}^{N} \left(1 - p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})\right) h_i}\end{aligned} \tag{B.26}$$

3 Two Gaussian Mixture Model

In the case of the mixture of two Gaussian distributions, we still have $K = 2$, $\pi_2 = 1 - \pi_1$. Now we define $\Theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2\}$ the means and variances necessary to define each Gaussian distribution. Moreover, each observation in $x_i \in X$ is now considered as $x_i = g_i/h_i$ where g_i is the number of detections which occured and h_i the number of detectors on the trajectory respectively. Thus: Thus:

$$\forall k \in [1,2] \quad p_k(x_i|\hat{\mu}_k, \hat{\sigma}_k) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}} \tag{B.27}$$

With similar calculations than for the binomial case lead us to the following results:
Equation (B.20) is now:

$$\tilde{\pi}_1 = \frac{1}{N}\sum_{i=1}^{N} \frac{\hat{\pi}_1\ p_1(x_i|\hat{\mu}_1, \hat{\sigma}_1)}{\hat{\pi}_1\ p_1(x_i|\hat{\mu}_1, \hat{\sigma}_1) + (1 - \hat{\pi}_1)\ p_2(x_i|\hat{\mu}_2, \hat{\sigma}_2)} \tag{B.28}$$

while for Θ Equation (B.12) is solved for (μ_1, μ_2) respectively as:

$$\begin{aligned}\frac{\partial Q(\Theta, \Pi, \hat{\Theta}, \hat{\Pi})}{\partial \mu_1} &= \sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta}) \frac{\partial \log\left(p_1(x_i|\mu_1, \sigma_1)\right)}{\partial \mu_1} \\ &= \sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta}) \left(\frac{(x_i - \mu_1)}{\sigma_1^2}\right)\end{aligned} \tag{B.29}$$

Leading to:

$$\tilde{\mu}_1 = \frac{\sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})\ x_i}{\sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})} \tag{B.30}$$

$$\begin{aligned}\tilde{\mu}_2 &= \frac{\sum_{i=1}^{N} p(z_i = 2|x_i, \hat{\Pi}, \hat{\Theta})\ x_i}{\sum_{i=1}^{N} p(z_i = 2|x_i, \hat{\Pi}, \hat{\Theta})} \\ &= \frac{\sum_{i=1}^{N} \left(1 - p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})\right) x_i}{\sum_{i=1}^{N} \left(1 - p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})\right)}\end{aligned} \tag{B.31}$$

Last, Equation (B.12) is solved for (σ_1, σ_2):

$$\begin{aligned}\frac{\partial Q(\Theta, \Pi, \hat{\Theta}, \hat{\Pi})}{\partial \sigma_1} &= \sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta}) \frac{\partial \log\left(p_1(x_i|\mu_1, \sigma_1)\right)}{\partial \sigma_1} \\ &= \sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta}) \left(-\frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3}\right)\end{aligned} \tag{B.32}$$

And finally:

$$\tilde{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})(x_i - \mu_1)^2}{\sum_{i=1}^{N} p(z_i = 1|x_i, \hat{\Pi}, \hat{\Theta})}} \tag{B.33}$$

$$\tilde{\sigma}_2 = \sqrt{\frac{\sum_{i=1}^{N} p(z_i = 2|x_i, \hat{\Pi}, \hat{\Theta})(x_i - \mu_2)^2}{\sum_{i=1}^{N} p(z_i = 2|x_i, \hat{\Pi}, \hat{\Theta})}} \tag{B.34}$$

APPENDIX C

Optimisation on Bluetooth Penetration Factors

Contents

1 Objective Function **138**

1.1 Consistency with Traffic Counts 138

1.2 Consistency with Bluetooth Sampling 138

1.3 Definition Domain 138

1.4 Kirchhoff's Law 139

2 Algorithm **139**

3 Simulated Case Study **140**

Work published and presented at :
G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 19–24, 2015, pp. 5480–5484. DOI: 10.1109/ICASSP.2015.7179019

IN this preliminary work, the idea is to directly estimate the penetration factor (inverse of the penetration rate) per link, origin and destination. In theory, for a given OD, the penetration factor should be constant on each path binding the OD pair. However, one link might belong to several of such paths and therefore the penetration rate, per OD and per link cannot be considered as constant. Therefore, a supplementary variable $\underline{\underline{\underline{\alpha}}}$, of size $N_V \times N_V \times N_L$ is introduced , such that :

$$\underline{\underline{\underline{Q}}} = \underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}}. \tag{C.1}$$

1 Objective Function

Very similarly to Chapter 5, Section 3.1, the objective functions can be derived from similar properties and still be classified in two types: The first type, composed of the functions presented in Section 1.1, Section 1.2 and Section 1.3, is aiming for consistency between the measures and the estimate. The second type, with the function in Section 1.4, stems from the topology of the network.

1.1 Consistency with Traffic Counts

Similarly to Chapter 5, Section 3.1.1, a first function aims to ensure the consistency between the estimates and the measured flows on the links. It is thus a constraint on the edges.

The assumption of a Gaussian noise $\underline{\varepsilon}$ on link counts is kept and thus, and reminding that,

$$\underline{\tilde{q}} = \underline{q}^{\star} + \underline{\varepsilon} \tag{5.5}$$

and

$$\underline{q}^{\star} = \sum_{ij\bullet} \underline{\underline{\underline{Q}}}^{\star} \tag{5.6}$$

a natural choice for such a function is:

$$f_{TC}(\underline{\underline{\underline{\alpha}}}) = \|\underline{\tilde{q}} - \sum_{ij\bullet} \underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}}\|^2. \tag{C.2}$$

1.2 Consistency with Bluetooth Sampling

A sound assumption is that the penetration factor, of Bluetooth among the vehicles in use on the road network is not varying much. Measures have shown that in Brisbane the Bluetooth penetration rate varies between 15 and 30% (see Chapter 3, Section 3.2.3). Thus, solutions with limited variability of the Bluetooth penetration factor are more likely.

The variability of the variable $\underline{\underline{\underline{\alpha}}}$ can be quantified through the following function, which is all the more interesting as being strongly convex, it insures the uniqueness of the solution :

$$f_{BS}(\alpha) = \|\underline{\underline{\underline{\alpha}}} - \tilde{\alpha}\|^2. \tag{C.3}$$

where $\tilde{\alpha}$ is the a priori average sampling ratio, computed as

$$\tilde{\alpha} = \frac{\sum \underline{\tilde{q}}}{\sum \underline{\underline{\underline{\tilde{B}}}}} \tag{C.4}$$

1.3 Definition Domain

As the total flow is at least greater or equal to the flow of Bluetooth enabled vehicles, it is further imposed that α belongs to the following convex constraint set:

$$C = \left\{ \alpha = \left(\alpha_{ijl}\right)_{(ijl)\in V\times V\times L} \in \mathbb{R}^{N_V\times N_V\times N_L} \mid \alpha_{ijl} \geq 1 \right\} \tag{C.5}$$

In the criterion, this constraint appears through an indicator function $f_C(\underline{\underline{\underline{\alpha}}}) = \iota_C(\underline{\underline{\underline{\alpha}}})$, equals to 0 if $\underline{\underline{\underline{\alpha}}} \in C$ and $+\infty$ otherwise.

1.4 Kirchhoff's Law

A last function comes from the balance of the flows on each node. It can be written using the classical ODM $\underline{\underline{T}}$. Remind Equation (5.3):

$$\underline{\underline{T}} = \sum_{\bullet\bullet l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{Q}}} = \sum_{\bullet\bullet l} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{Q}}} \tag{5.3}$$

The balance requires that, at every node, the flow having for destination this node ($\underline{D}$), minus the flow originating this same node ($\underline{O}$) should equal the flow going through it. This is written as:

$$\underline{D} - \underline{O} = (\underline{\underline{I}} - \underline{\underline{E}})^{\top} \top_{13} \underline{q} \tag{C.6}$$

where $\top_{13}$ operates the transposition operation between the first and third dimension ($\underline{q}$ is of size $1 \times 1 \times N_L$).

Using variable $\underline{\underline{\underline{\alpha}}}$ and data $\underline{\underline{\underline{\tilde{B}}}}$ and $\underline{\tilde{q}}$ with Equations (5.3) and (5.4), it reads as

$$\left(\sum_{i\bullet l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}} \right)^{\top} - \sum_{\bullet jl} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}} = (\underline{\underline{I}} - \underline{\underline{E}})^{\top} \top_{13} \underline{q}. \tag{C.7}$$

The natural function stemming from the above constraint relaxed, considering Equation (5.5) is

$$f_K(\underline{\underline{\underline{\alpha}}}) = \| \left(\sum_{i\bullet l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{\tilde{B}}}} \circ \underline{\underline{\underline{\alpha}}} \right)^{\top} - \sum_{\bullet jl} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{\tilde{B}}}} \circ \underline{\underline{\underline{\alpha}}} - (\underline{\underline{I}} - \underline{\underline{E}})^{\top} \top_{13} \underline{\tilde{q}} \|^2. \tag{C.8}$$

2 Algorithm

The criterion to obtain a relevant *transport* solution, designed using the topology of the networks and the data available, then reads:

$$\underline{\underline{\underline{\hat{\alpha}}}} \in \underset{\underline{\underline{\underline{\alpha}}}}{\text{Argmin}} \left\{ \gamma_{TC} f_{TC}(\underline{\underline{\underline{\alpha}}}) + \gamma_{BS} f_{BS}(\underline{\underline{\underline{\alpha}}}) + \gamma_K f_K(\underline{\underline{\underline{\alpha}}}) + \gamma_C f_C(\underline{\underline{\underline{\alpha}}}) \right\} \tag{C.9}$$

with $\gamma_{\cdot} \geq 0$, the weight of each constraint.

The functions involved in Problem (C.9) are convex, lower semi-continuous and proper. Moreover, $\gamma_{TC} f_{TC} + \gamma_{BS} f_{BS} + \gamma_K f_K$ is differentiable with a β-Lipschitz gradient where the value of β depends on the norm of the matrices involved in each function. The gradients are:

$$\nabla f_{TC}(\underline{\underline{\underline{\alpha}}}) = \left(\left(-2\, B^l_{ij} \left(\tilde{q}^l - \sum_{k,m} \alpha^l_{km} B^l_{km} \right) \right)^l_{ij} \right)_{(ijl) \in V \times V \times L} \tag{C.10}$$

$$\nabla f_{BS}(\underline{\underline{\underline{\alpha}}}) = 2 \left(\left(\alpha^l_{ij} - \tilde{\alpha} \right)^l_{ij} \right)_{(ijl) \in V \times V \times L} \tag{C.11}$$

$$\nabla f_K(\underline{\underline{\underline{\alpha}}}) = \left(\left(2B^l_{ij} \sum_m (\delta_{jm} I^l_j - \delta_{im} E^l_i) \left(\sum_{ke} I^e_k B^e_{km} \alpha^e_{km} - E^e_k B^e_{mk} \alpha^e_{mk} - (I^e_m - E^e_m) \tilde{q}^e \right) \right)^l_{ij} \right)_{(ijl) \in V \times V \times L} \tag{C.12}$$

where δ_{km} is the Kronecker delta:

$$\delta_{km} = \begin{cases} 1 & \text{if } k = m \\ 0 & \text{otherwise} \end{cases} \tag{C.13}$$

We denote by β_{TC}, β_{BS}, and β_K the Lipschitz constants of f_{TC}, f_{BS}, and f_K respectively and thus by $\beta = \gamma_{TC}\beta_{TC} + \gamma_{BS}\beta_{BS} + \gamma_K\beta_K$ the Lipschitz constant of $\gamma_{TC}f_{TC} + \gamma_{BS}f_{BS} + \gamma_K f_K$

The function $f_C = \iota_C$ is non-differentiable but it has a closed form expression for its projection [198]. To find $\widehat{\underline{\underline{\alpha}}}$, we used the forward-backward algorithm, adapted from [10], [196], [203], described as follow in Algorithm C.1.

Algorithm C.1 Forward Backward Algorithm for LOD Bluetooth Penetration Factor Estimation

Input: $\gamma_{TC} \geq 0, \quad \gamma_{BS} \geq 0, \quad \gamma_K \geq 0, \quad \gamma_D \in [0,1]$
Compute: β
1: $\tau \leftarrow \frac{1.99}{\beta}$
2: $\underline{\underline{\alpha}}^{[0]} \leftarrow 0$
3: **for all** $n \in [0,1,...]$ **do**
4: $\quad \underline{\underline{\alpha}}^{[n+\frac{1}{2}]} = \underline{\underline{\alpha}}^{[n]} - \tau\left(\gamma_{TC}\nabla f_{TC} + \gamma_{BS}\nabla f_{BS} + \gamma_K \nabla f_K\right)\left(\underline{\underline{\alpha}}^{[n]}\right)$
5: $\quad \underline{\underline{\alpha}}^{[n+1]} = \max\left\{\underline{\underline{\alpha}}^{[n+\frac{1}{2}]}, 1\right\}$
6: $\quad$ **if** $\frac{\|\underline{\underline{\alpha}}^{[n+1]} - \underline{\underline{\alpha}}^{[n]}\|_2}{\|\underline{\underline{\alpha}}^{[n+1]}\|_2} < 10^{-6}$ **or** $n > 10^5$ **then return** $\underline{\underline{\alpha}}^{[n+1]}$

According to Combettes and Pesquet (2010) [10], the sequence $(\underline{\underline{\alpha}}^{[n]})_{n\in\mathcal{N}}$ converges to $\widehat{\underline{\underline{\alpha}}}$. Moreover, its convergence rate has been described in [213]. In practice, we consider the convergence is achieved when the relative error between two iterates is such that $\frac{\|\alpha^{[n]} - \alpha^{[n-1]}\|^2}{\|\alpha^{[n]}\|^2} \leq 10^{-6}$ or if there has been over 10^5 iteration steps.

3 Simulated Case Study

This algorithm has been tested on the same case study as presented in Chapter 5, Section 4. Tables comparing the best results achieved for the three formulations of the LOD matrix estimation problem is presented in Appendix D.

APPENDIX D

A simple Forward-Backward approach

Contents

1 Objective Function **142**
1.1 Consistency with Traffic Counts 142
1.2 Definition Domain 142
1.3 Kirchhoff's Law 142
1.4 Poisson Bluetooth Sampling Data Fidelity 143
2 Algorithm **143**
3 Case Study and Results **144**

Work published and presented at :
G. Michau, P. Borgnat, N. Pustelnik, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", presented at the XXV GRETSI, Lyon, France, Sep. 8, 2015. [Online]. Available: http://eprints.qut.edu.au/86449/

AS a transitional work between the one presented in Appendix C and the one presented in Chapter 5, we proposed as a second preliminary work, the direct estimation of the LDOM, that is, to reinterpret Problem (C.9) in Appendix C as a problem similar to Problem (5.8).

Remind that:

$$\underline{\underline{\underline{\widehat{\alpha}}}} \in \underset{\underline{\underline{\underline{\alpha}}}}{\text{Argmin}} \left\{ \gamma_{TC} f_{TC}(\underline{\underline{\underline{\alpha}}}) + \gamma_{BS} f_{BS}(\underline{\underline{\underline{\alpha}}}) + \gamma_K f_K(\underline{\underline{\underline{\alpha}}}) + \gamma_C f_C(\underline{\underline{\underline{\alpha}}}) \right\} \tag{C.9}$$

and

$$\underline{\underline{\underline{\widehat{Q}}}} \in \underset{\underline{\underline{\underline{Q}}}}{\text{Argmin}} \left\{ \gamma_1 \mathscr{D}_1(\underline{\underline{\underline{\tilde{B}}}}, \underline{\underline{\underline{\eta_o}}}\, \underline{\underline{\underline{Q}}}) + \gamma_2 \mathscr{D}_2(\underline{\tilde{q}}, \sum_{ij\bullet} \underline{\underline{\underline{Q}}}) + \sum_{k\geq 2} \gamma_k f_k(\underline{\underline{\underline{Q}}}) \right\} \tag{5.8}$$

Indeed, in Problem (C.9), all the four terms but f_{BS} can be seen as functions of $\underline{\underline{\underline{\alpha}}}$ but also of $\underline{\underline{\underline{\alpha}}} \circ \underline{\underline{\underline{\tilde{B}}}}$, that is, of the LODM $\underline{\underline{\underline{Q}}}$. Therefore, the problem can easily be adapted for the direct estimation of the LODM.

Moreover, this change of variable is useful for two reasons: First, it enables the estimation of the elements of $\underline{\underline{\underline{Q}}}$ where the corresponding element in $\underline{\underline{\underline{\tilde{B}}}}$ is zeros. This, accounts for the possibility of a trajectory not being represented by a Bluetooth sample. Second, it gives the possibility of assuming that the variables $\underline{\underline{\underline{Q}}}$ and $\underline{\underline{\underline{\tilde{B}}}}$ are related by a Poisson distribution of type $\underline{\underline{\underline{\tilde{B}}}} \sim \mathscr{P}\left(\eta \underline{\underline{\underline{Q}}}\right)$ where η can be related to the Bluetooth penetration rate.

1 Objective Function

1.1 Consistency with Traffic Counts

The consistency between volumes on links computed from the estimate and observed traffic counts can be expressed very similarly for both cases when $\underline{\underline{\underline{\alpha}}}$ or $\underline{\underline{\underline{Q}}}$ are the variable of interest. The case of $\underline{\underline{\underline{\alpha}}}$ has been developed in Appendix C, Section 1.1 while the one of $\underline{\underline{\underline{Q}}}$ has been presented in Chapter 5, Section 3.1.1. The function is, in this case, the same as in Chapter 5, Section 3.1.1:

$$f_{TC}(\underline{\underline{\underline{Q}}}) = \|\underline{\tilde{q}} - \sum_{ij\bullet} \underline{\underline{\underline{Q}}}\|^2. \tag{5.9}$$

1.2 Definition Domain

Similarly, this term has already been presented for both cases: as a function of $\underline{\underline{\underline{\alpha}}}$ in Appendix C, Section 1.3 and as a function of $\underline{\underline{\underline{Q}}}$ in Chapter 5, Section 3.1.3. Thus, here we have

$$f_C(\underline{\underline{\underline{Q}}}) = \iota_C(\underline{\underline{\underline{Q}}}) = \begin{cases} 0 & \text{if } \underline{\underline{\underline{Q}}} \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.14}$$

where

$$C = \left\{\underline{\underline{\underline{Q}}} = \left(Q^l_{ij}\right)_{(ijl)\in V\times V\times L} \in \mathbb{R}^{N_V\times N_V\times N_L} \mid Q^l_{ij} \geq B^l_{ij}\right\}. \tag{5.13}$$

1.3 Kirchhoff's Law

The Kirchhoff law as presented in Appendix C, Section 1.4 can also be adapted for $\underline{\underline{\underline{Q}}}$ as the variable. We propose however a slight change:

Here, we also start from Equation (C.6) reminded here,

$$\underline{D} - \underline{O} = (\underline{\underline{I}} - \underline{\underline{E}})^\top \top_{13} \underline{q} \tag{C.6}$$

but instead of using the observed traffic counts $\tilde{q}$ in the left-hand side of this equation, we propose here to use the estimate:

$$\underline{D} - \underline{O} = (\underline{\underline{I}} - \underline{\underline{E}})^\top \top_{13} \sum_{ij\bullet} \underline{\underline{\underline{Q}}}. \tag{D.1}$$

With Equation (5.3), this leads to

$$\left(\sum_{i\bullet l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{Q}}}\right)^\top - \sum_{\bullet jl} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{Q}}} = (\underline{\underline{I}} - \underline{\underline{E}})^\top \top_{13} \sum_{ij\bullet} \underline{\underline{\underline{Q}}} \tag{D.2}$$

The function resulting from this relationship is thus

$$f_K(\underline{\underline{\underline{Q}}}) = \| \left(\sum_{i\bullet l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{Q}}} \right)^\top - \sum_{\bullet jl} \underline{\underline{\underline{E}}} \circ \underline{\underline{\underline{Q}}} - (\underline{\underline{I}} - \underline{\underline{E}})^\top \top_{13} \sum_{ij\bullet} \underline{\underline{\underline{Q}}} \|^2. \tag{D.3}$$

Note that neither is it the same function as presented in Chapter 5, Section 3.1.2 which was a Kirchhoff law applied to each node of the network and for each OD. This version of the Kirchhoff's law describes the balance of the flows at each node independently of the OD. A demonstration on how the Kirchhoff law applied to each node and OD also balance the flow at each node is found in Appendix E, Section 2.2 with other considerations on flow conversation.

1.4 Poisson Bluetooth Sampling Data Fidelity

Similarly to Chapter 5, Section 3.1.2, we propose here to assume that the Bluetooth trajectories are sampled from the total traffic according to a Poisson distribution and therefore to consider the following function:

$$f_P(\underline{\underline{\underline{Q}}}) = \sum_{ijl} \psi \left(B^l_{ij}, \eta^l Q^l_{ij} \right) \tag{5.11}$$

2 Algorithm

To sum up, the objective is to find an estimate $\widehat{\underline{\underline{\underline{Q}}}}$ of $\underline{\underline{\underline{Q}}}^\star$ satisfying

$$\widehat{\underline{\underline{\underline{Q}}}} \in \underset{\underline{\underline{\underline{Q}}}}{\text{Argmin}} \left\{ \gamma_{TC} f_{TC}(\underline{\underline{\underline{Q}}}) + \gamma_P f_P(\underline{\underline{\underline{Q}}}) + \gamma_C f_C(\underline{\underline{\underline{Q}}}) + \gamma_K f_K(\underline{\underline{\underline{Q}}}) \right\} \tag{D.4}$$

where $\gamma_\cdot$ are positive weights applied to the objectives and model their relative importance within the global objective.

All the four functions involved in Equation (D.4) are convex, lower-semicontinuous (l.s.c.) and proper. Moreover, both the functions f_{TC} and f_K are differentiable and their gradients are:

$$\nabla f_{TC}(\underline{\underline{\underline{Q}}}) = \left(\left(-2 \left(\tilde{q}^l - \sum_{k,m} Q^l_{km} \right) \right)^l_{ij} \right)_{(ijl) \in V \times V \times L} \tag{D.5}$$

$$\nabla f_K(\underline{\underline{\underline{Q}}}) = \left(\left(2 \sum_m (\delta_{jm} I^l_j - \delta_{im} E^l_i - (I^l_m - E^l_m)) \left(\sum_{ke} I^e_k Q^e_{km} - E^e_k Q^e_{mk} - (I^e_m - E^e_m) \sum_{np} Q^e_{np} \right) \right)^l_{ij} \right)_{(ijl) \in V \times V \times L} \tag{D.6}$$

The Lipschitz constants of f_{TC} and f_K are denoted β_{TC} and β_K respectively.

Similarly to Chapter 5, Section 3.2, the proximal operator of the sum of f_P and f_C satisfies the composition property [199]:

$$\text{prox}_{\gamma_C f_C + \gamma_P f_P}(\underline{\underline{\underline{Q}}}) = P_C(\text{prox}_{\gamma_P f_P}(\underline{\underline{\underline{Q}}})). \tag{D.7}$$

and, Problem (D.4) has been solved, similarly to Appendix C, Section 2, using the Forward-Backward Algorithm D.1.

Algorithm D.1 Forward Backward Algorithm for LOD Matrix Estimation

Input: $\gamma_{TC} \geq 0, \quad \gamma_K \geq 0, \quad \gamma_P \geq 0, \quad \gamma_D \in [0,1]$
Compute: β

1: $\tau \leftarrow \frac{1.99}{\beta}$
2: $\underline{\underline{Q}}^{[0]} \leftarrow 0$
3: **for all** $n \in [0,1,...]$ **do**
4: $\quad \underline{\underline{Q}}^{[n+\frac{1}{2}]} = \underline{\underline{Q}}^{[n]} - \tau\left(\gamma_{TC}\nabla f_{TC} + \gamma_K \nabla f_K\right)\left(\underline{\underline{Q}}^{[n]}\right)$
5: $\quad \underline{\underline{Q}}^{[n+1]} = \max\left\{\text{prox}_{\gamma_P f_P}\left(\underline{\underline{Q}}^{[n+\frac{1}{2}]}\right), \underline{\underline{\tilde{B}}}\right\}$
6:
7: $\quad$ **if** $\frac{\|\underline{\underline{Q}}^{[n+1]} - \underline{\underline{Q}}^{[n]}\|_2}{\|\underline{\underline{Q}}^{[n+1]}\|_2} < 10^{-6}$ **or** $n > 10^5$ **then return** $\underline{\underline{Q}}^{[n+1]}$

3 Case Study and Results

The three approaches have been tested on the same case study, as presented in Chapter 5, Section 4. Table D.1 and Table D.2 summarise the results of the three approaches presented in this manuscript (Table D.1 for $N = 100\,000$ users and Table D.2 for $N = 10\,000$). For comparison purposes, performances of the naive solutions are also reminded in these tables (*cf.* Equation (5.40) and (5.41)).

For simplifying the comparison, estimates minimising the RMSE indicator only are presented here. Yet the following discussion also applies for estimates optimal with respect to the EMD indicator. Moreover, the values of the γ. providing the optimal performances are not presented here as, depending on the approach, they would not refer to the same regularisation function (different version of the Kirchhoff's law, optimisation over the penetration factor or over the LOD matrix directly).

Table D.1: Results for LOD matrix estimates with $N = 100\,000$. Optimal Solutions for RMSE only.

	RMSE	EMD	f_K	f_{TC}
Naive Solution $\underline{\underline{\widehat{Q}}}^0$	0.320	0.086	**0**	55
Naive Solution $\underline{\underline{\widehat{Q}}}^1$	0.307	0.069	1142	84
Alg. C.1 (App. C)	0.306	0.078	701	36
Alg. D.1 (App. D)	0.279	0.068	598	41
Prox. Alg. (Chap.5)	**0.239**	**0.047**	289	**1**

These two tables illustrates that the successive refinements of the problem formulation led, at each step, to improved performances:

First, the three estimation problems have solutions outperforming the naive solutions. Yet, the estimate of the penetration factors (*Alg. C.1 (App. C)*) gives results close to the naive solutions. This is the consequence of the optimisation process that estimates values of the LOD penetration factor only for non-zero elements of $\underline{\underline{\tilde{B}}}$. Similarly, the naive solutions have zero values wherever $\underline{\underline{\tilde{B}}}$ has zero elements. Moreover, one of the regularisation function favour estimates with penetration factors close to the global average $\tilde{\alpha}$. By definition, we have $\underline{\underline{\widehat{Q}}}^0 = \tilde{\alpha}\underline{\underline{\tilde{B}}}$. It is therefore expected to have a solution similar to the naive ones.

Table D.2: Results for LOD matrix estimates with $N = 10\,000$. Optimal Solutions for RMSE only.

	RMSE	EMD	f_K	f_{TC}
Naive Solution $\underline{\underline{\underline{\widehat{Q}}}}^0$	0.398	0.021	**0**	13.3
Naive Solution $\underline{\underline{\underline{\widehat{Q}}}}^1$	0.396	0.017	128	**0.1**
Alg. C.1 (App. C)	0.394	0.020	45	22
Alg. D.1 (App. D)	0.384	0.016	79	7
Prox. Alg. (Chap.5)	**0.341**	**0.013**	10.7	9.8

As a consequence, the direct estimation of the LOD matrix, as in Chapter 5 and in Appendix D led to improved performances compared both to the naive solutions but also to the solution of Algorithm C.1.

Second, the implementation of the local Kirchhoff's law, in Chapter 5, led, in particular, to improved performances with respect to the function f_K, a important indicator of the consistency of the assignment information contained by the LOD matrix.

For more detailed presentation of the behaviour and performances of the two preliminary approaches, the reader is invited to refer to both papers [17] and [20].

APPENDIX E

Flow Conservation and Kirchhoff's law

Contents

1 No Cycle Hypothesis **147**
1.1 Outgoing Flows at Destination and Incoming Flows at Origins 147
1.2 No Cycle Hypothesis : From LOD to OD Matrix 147
2 Proposition: Local Kirchhoff's law induces other relationships **148**
2.1 Local Kirchhoff's law induces Equation (5.3) 148
2.2 Local Kirchhoff's law induces the Global Kirchhoff's law 149

1 No Cycle Hypothesis

1.1 Outgoing Flows at Destination and Incoming Flows at Origins

If the *no cycles* hypothesis hold, then both the outgoing flows at the destination nodes and the incoming flows at the origin nodes should be null:

$$\forall (i,j,l) \in (V \times V \times L)$$

$$I_i^l \; Q_{ij}^l = 0 \tag{E.1}$$

$$E_j^l \; Q_{ij}^l = 0 \tag{E.2}$$

which also corresponds to

$$\underline{\underline{\underline{E_2}}} \circ \underline{\underline{\underline{Q}}} = 0 \tag{E.3}$$

$$\underline{\underline{\underline{I_1}}} \circ \underline{\underline{\underline{Q}}} = 0 \tag{E.4}$$

1.2 No Cycle Hypothesis : From LOD to OD Matrix

Under the *no cycles* assumption we have the property that any cut of the road network between a given origin node i and a given destination node j should have the same flow going though, that corresponds

to the OD flows T_{ij}. That can be more formally expressed as:

$$\forall (i,j) \in (V \times V),\ \forall p \in |(Cut_{i/j})| \\ T_{ij} = \sum_{l \in Cut^p_{i/j}} Q^l_{ij} \tag{E.5}$$

where $Cut_{i/j}$ is the set of link-sets cutting the graph between i and j and $Cut^p_{i/j}$ is the set of links of the p-th possible cut.

Two obvious cuts are around originating node i and destination nodes j thus

$$\forall (i,j) \in (V \times V) \quad T_{ij} = \begin{cases} \sum_{l;E^l_i=1} Q^l_{ij} \\ \sum_{l;I^l_j=1} Q^l_{ij} \end{cases} \tag{E.6}$$

which also corresponds to Equation (5.3)

$$\underline{\underline{T}} = \begin{cases} \sum_{\bullet\bullet l} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{Q}}}\ , \\ \sum_{\bullet,\bullet,l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{Q}}}\ . \end{cases} \tag{5.3}$$

2 Proposition: Local Kirchhoff's law induces other relationships

2.1 Local Kirchhoff's law induces Equation (5.3)

Equation (5.15):

$$\forall (i,j,k) \in (V \times V \times V) \quad \sum_l E^l_k\, Q^l_{ij} - \delta_{ik}\, T_{ij} = \sum_l I^l_k\, Q^l_{ij} - \delta_{jk}\, T_{ij}. \tag{5.15}$$

Lets consider the case where $i = k$:

$$\forall (i,j) \in (V \times V) \quad \sum_l E^l_i\, Q^l_{ij} - \sum_l I^l_i\, Q^l_{ij} = T_{ij} - \delta_{kj}\, T_{ij} \tag{E.7}$$

Under the assumption that itineraries do not have cycles (fair assumption in transport networks), then

$$\forall (i,j,l) \in (V \times V \times L) \quad I^l_i\, Q^l_{ij} = 0 \tag{E.8}$$

Moreover, under the same assumption we have

$$\forall (i,l) \in (V \times L) \quad Q^l_{ii} = 0 \quad (\text{in particular } T_{ii} = 0) \tag{E.9}$$

Thus:

$$\forall (i,j) \in (V \times V) \quad \sum_l E^l_i\, Q^l_{ij} = T_{ij} \tag{E.10}$$

Similarly for $k = j$ we have

$$\forall (i,j) \in (V \times V) \quad \sum_l I^l_j\, Q^l_{ij} = T_{ij} \tag{E.11}$$

and

$$\left(\forall (i,j,l) \in (V \times V \times L) \quad E^l_j\, Q^l_{ij} = 0\right) \tag{E.12}$$

Finally we have Equation (5.3)

$$\underline{\underline{T}} = \sum_{\bullet\bullet l} \underline{\underline{\underline{E_1}}} \circ \underline{\underline{\underline{Q}}} = \sum_{\bullet,\bullet,l} \underline{\underline{\underline{I_2}}} \circ \underline{\underline{\underline{Q}}}. \tag{5.3}$$

2.2 Local Kirchhoff's law induces the Global Kirchhoff's law

We demonstrate here that the Kirchhoff's law, applied per OD, as in Chapter 5, Section 3.1.4 has for consequence that the global Kirchhoff's law, as in Appendix D, Section 1.3 is also satisfied.

Starting from the Kirchhoff's law applied per OD, Equation (5.15):

$$\forall (i,j,k) \in (V \times V \times V) \quad \sum_l E_k^l \, Q_{ij}^l - \delta_{ik} \, T_{ij} = \sum_l I_k^l \, Q_{ij}^l - \delta_{jk} \, T_{ij}. \tag{5.15}$$

Thus, summing this relationships over every origin i and every destination j, we have

$$\sum_l E_k^l \sum_{ij} Q_{ij}^l - \sum_l I_k^l \sum_{ij} Q_{ij}^l = \sum_i T_{ik} - \sum_j T_{kj} \tag{E.13}$$

or, using matrix notations

$$\left(\underline{\underline{E}} - \underline{\underline{I}}\right)^\top \cdot \underline{\underline{Q}} = \underline{O} - \underline{D}. \tag{E.14}$$

Bibliography

[1] M. Coates, A. Hero, R. Nowak, and B. Yu, "Internet tomography", *IEEE Signal Processing Magazine*, vol. 19, no. 3, pp. 47–65, May 2002, ISSN: 1053-5888. DOI: 10.1109/79.998081 (cited pp. 20, 170).

[2] A. Girard, B. Sansò, and F. Vazquez-Abad, *Performance Evaluation and Planning Methods for the next Generation Internet*. Springer, 2006, vol. 6 (cited pp. 20, 170).

[3] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (extended version)", *Networking, IEEE/ACM Transactions on*, vol. 20, pp. 662–676, 2012. DOI: 10.1109/tnet.2011.2169424 (cited pp. 20, 170).

[4] M. Mardani and G. Giannakis, "Estimating traffic and anomaly maps via network tomography", *IEEE/ACM Transactions on Networking*, no. 99, pp. 1–15, 2015, ISSN: 1063-6692. DOI: 10.1109/TNET.2015.2417809 (cited pp. 20, 170).

[5] A. Pirayre, C. Couprie, F. Bidard, L. Duval, and J.-C. Pesquet, "BRANE Cut: Biologically-related a priori network enhancement with graph cuts for gene regulatory network inference", *BMC Bioinformatics*, vol. 16, p. 369, 2015, ISSN: 1471-2105. DOI: 10.1186/s12859-015-0754-2 (cited pp. 20, 171).

[6] A. Pirayre, C. Couprie, L. Duval, and J. C. Pesquet, "Fast convex optimization for connectivity enforcement in gene regulatory network inference", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 1002–1006. DOI: 10.1109/ICASSP.2015.7178120 (cited pp. 20, 171).

[7] C. Couprie, L. Grady, L. Najman, J.-C. Pesquet, and H. Talbot, "Dual constrained TV-based regularization on graphs", *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1246–1273, 2013. DOI: 10.1137/120895068 (cited pp. 20, 101, 171).

[8] A. Cauchy, "Méthode générale pour la résolution des systemes d'équations simultanées", *Comp. Rend. Sci. Paris*, vol. 25, no. 1847, pp. 536–538, 1847. DOI: 10.1017/cbo9780511702396.063 (cited pp. 21, 171).

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2009, ISBN: 978-0-521-83378-3 (cited pp. 21, 171).

[10] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing", in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds., New-York: Springer-Verlag, 2010, pp. 185–212 (cited pp. 21, 102, 140, 171).

[11] G. Michau, *Opérateur Dérivé Va À La Montagne*, PhyLab Editions. Oct. 2015, vol. 1, 14 pp. [Online]. Available: http://perso.ens-lyon.fr/gabriel.michau/Papers/MICHAU_2015_OperateurDeriveMontagne.pdf (cited pp. 21, 171).

[12] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms", *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, Aug. 1, 2013, ISSN: 0022-3239, 1573-2878. DOI: 10.1007/s10957-012-0245-9 (cited pp. 21, 102, 171).

[13] P. L. Combettes and J.-C. Pesquet, "Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators", *Set-Valued and Variational Analysis*, vol. 20, no. 2, pp. 307–330, Aug. 27, 2011, ISSN: 1877-0533, 1877-0541. DOI: 10.1007/s11228-011-0191-y (cited pp. 21, 102, 171).

[14] G. Michau, A. Nantes, A. Bhaskar, E. Chung, P. Borgnat, and P. Abry, "Bluetooth data in urban context: Retrieving vehicles trajectories", *Submitted in IEEE Transaction on Intelligent Transport Systems*, 2016 (cited pp. 22, 46, 70, 95, 167).

[15] G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, A. Bhaskar, and E. Chung, "A primal-dual algorithm for link dependent origin destination matrix estimation", *Submitted in IEEE Transactions on Signal and Information Processing Over Networks*, 2016 (cited pp. 22, 94, 167).

[16] G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving dynamic origin-destination matrices from Bluetooth data", in *Transportation Research Board, 93rd Annual Meeting*, Washington DC, Jan. 12–16, 2014. [Online]. Available: http://eprints.qut.edu.au/66511/ (cited pp. 22, 46, 70, 167).

[17] G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 19–24, 2015, pp. 5480–5484. DOI: 10.1109/ICASSP.2015.7179019 (cited pp. 23, 94, 95, 100, 109, 137, 145, 167).

[18] G. Michau, A. Nantes, and E. Chung, "Towards the retrieval of accurate OD matrices from Bluetooth data: Lessons learned from 2 years of data", in *36th Australasian Transport Research Forum (ATRF)*, QUT, Brisbane, Australia, Oct. 4, 2013. [Online]. Available: http://eprints.qut.edu.au/62727/ (cited pp. 23, 46, 167).

[19] G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving trip information from a discrete detectors network: The case of Brisbane Bluetooth detectors", in *CAITR*, Sydney, Feb. 17–18, 2014. [Online]. Available: http://eprints.qut.edu.au/83110/ (cited pp. 23, 46, 70, 83, 168).

[20] G. Michau, P. Borgnat, N. Pustelnik, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", presented at the XXV GRETSI, Lyon, France, Sep. 8, 2015. [Online]. Available: http://eprints.qut.edu.au/86449/ (cited pp. 23, 94, 95, 100, 109, 110, 141, 145, 168).

[21] G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Graph Signal Processing Workshop*, Philadelphia, May 25–27, 2016 (cited pp. 23, 94, 168).

[22] ——, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Complex Networks*, Marseilles, France, Jul. 11–13, 2016 (cited pp. 23, 94, 168).

[23] D. McFadden, A. Talvitie, S. Cosslett, I. Hasan, M. Johnson, F. Reid, and K. Train, *Demand Model Estimation and Validation*. Institute of Transportation Studies, 1977, vol. 5 (cited p. 27).

[24] J. Bates, "Stated preference technique for the analysis of transportation behavior", in *Proceedings of World Conference of Transportation Research*, 1982, pp. 252–265 (cited p. 27).

[25] J. J. Louviere, "Conjoint analysis modelling of stated preferences: A review of theory, methods, recent developments and external validity", *Journal of transport economics and policy*, pp. 93–119, 1988, ISSN: 0022-5258 (cited p. 27).

[26] D. A. Hensher, "Stated preference analysis of travel choices: The state of practice", *Transportation*, vol. 21, pp. 107–133, 1994, 2, ISSN: 0049-4488. DOI: 10.1007/bf01098788 (cited p. 27).

[27] S. Fujii and T. Gärling, "Application of attitude theory for improved predictive accuracy of stated preference methods in travel demand analysis", *Transportation Research Part A: Policy and Practice*, vol. 37, pp. 389–402, 2003, ISSN: 0965-8564. DOI: 10.1016/s0965-8564(02)00032-0 (cited p. 27).

[28] T. Veitch, M. Paech, and J. Eaton, "What's missing from australian household travel surveys? Off-peak travel!", in *Australasian Transport Research Forum 2013*, Brisbane, Oct. 2–4, 2013 (cited p. 27).

[29] W. J. Reilly, *Methods for the Study of Retail Relationships*. Bureau of Business Research, University of Texas, 1959, 50 pp. (cited pp. 28, 33).

[30] P. Robillard, "Estimating the O-D matrix from observed link volumes", *Transportation Research*, vol. 9, no. 2–3, pp. 123–128, Jul. 1975, ISSN: 0041-1647. DOI: 10.1016/0041-1647(75)90049-0 (cited pp. 30, 34).

[31] L. G. Willumsen, "Estimation of an OD matrix from traffic counts: A review", *Institute of Transport Studies, Universities of Leed*, 1978 (cited pp. 31, 33, 34, 37).

[32] J. D. Ortuzar and L. G. Willumsen, *Modelling Transport*, 4th. John Wiley & Sons, Ltd, 2011, 608 pp., ISBN: 978-0-470-76039-0 (cited p. 31).

[33] A. Peterson, "The origin-destination matrix estimation problem: Analysis and computations", Linköping University, Department of Science and Technology, 2007 (cited pp. 31, 38).

[34] H. J. Casey, "Applications to traffic engineering of the law of retail gravitation", *Traffic Quarterly*, vol. 9, no. 1, pp. 23–35, 1955 (cited p. 33).

[35] D. Low, "A new approach to transportation systems modelling", *Traffic Quartely*, vol. 26, pp. 391–404, 1972 (cited pp. 33, 34).

[36] OCED, *Urban Traffic Models: Possibilities for Simplification : A Report*. Paris: Organisation for Economic Co-operation and Development, 1974, 136 pp. (cited p. 34).

[37] L. Lamarre, *Une méthode linéaire simple d'estimation de la matrice des origines et des destinations à partir de comptages sur les liens d'un réseau: Une application au réseau routier du québec*. Université de Montréal, Centre de recherche sur les transports, 1977, 190 pp. (cited p. 34).

[38] J. Symons, R. Wilson, and J. Paterson, "A model of inter city motor travel estimated by link volumes", in *Australian Road Research Board (ARRB) Conference, 8th, 1976, Perth*, vol. 8, 1976 (cited p. 34).

[39] S. Erlander and N. F. Stewart, *The Gravity Model in Transportation Analysis: Theory and Extensions*. Vsp, 1990, vol. 3 (cited p. 34).

[40] M. West, "Statistical inference for gravity models in transportation flow forecasting", *Discussion Paper, Institue of Statistics and Decision Sciences, Duke University, Durham*, 1994 (cited p. 34).

[41] I. Ekowicaksono, F. Bukhari, and A. Aman, "Estimating origin-destination matrix of bogor city using gravity model", *IOP Conference Series: Earth and Environmental Science*, vol. 31, no. 1, p. 012 021, 2016, ISSN: 1755-1315. DOI: 10.1088/1755-1315/31/1/012021 (cited p. 34).

[42] H. D. Sherali, R. Sivanandan, and A. G. Hobeika, "A linear programming approach for synthesizing origin-destination trip tables from link traffic volumes", *Transportation Research Part B: Methodological*, vol. 28, no. 3, pp. 213–233, Jun. 1994, ISSN: 0191-2615. DOI: 10.1016/0191-2615(94)90008-6 (cited pp. 34, 37).

[43] A. G. Wilson, *Entropy in Urban and Regional Modelling*. Pion Ltd, 1970, 166 pp., ISBN: 0-85086-021-0 (cited pp. 34, 35).

[44] A. G. Wilson, *The Use of Entropy Maximising Models in the Theory of Trip Distribution, Mode Split and Route Split*. London: Centre for Environmental Studies, 1968 (cited pp. 34, 35).

[45] A. Wilson, "Entropy in urban and regional modelling: Retrospect and prospect.", *Geographical Analysis*, vol. 42, no. 4, pp. 364–394, 2010, ISSN: 1538-4632. DOI: 10.1111/j.1538-4632.2010.00799.x (cited p. 34).

[46] H. Van Zuylen, "A method to estimate a trip matrix from traffic volume counts", in *PTRC SUMMER ANNUAL MEETING*, University of Warwick, London, Jul. 1978 (cited p. 35).

[47] C. S. Fisk, "On combining maximum entropy trip matrix estimation with user optimal assignment", *Transportation Research Part B: Methodological*, vol. 22, no. 1, pp. 69–73, Feb. 1, 1988, ISSN: 0191-2615. DOI: 10.1016/0191-2615(88)90035-5 (cited p. 36).

[48] H. J. Van Zuylen and L. G. Willumsen, "The most likely trip matrix estimated from traffic counts", *Transportation Research Part B: Methodological*, vol. 14, no. 3, pp. 281–293, 1980. DOI: 10.1016/0191-2615(80)90008-9 (cited p. 36).

[49] W. H. K. Lam and H. P. Lo, "Estimation of origin-destination matrix from traffic counts: A comparison of entropy maximizing and information minimizing models", *Transportation Planning and Technology*, vol. 16, no. 2, pp. 85–104, 1991, ISSN: 0308-1060. DOI: 10.1080/03081069108717474 (cited p. 36).

[50] G. Minty, "Statistical estimation of flows in networks", *IEEE Transactions on Circuit Theory*, vol. 10, no. 2, pp. 310–311, 1963, ISSN: 0018-9324. DOI: 10.1109/TCT.1963.1082139 (cited p. 36).

[51] H. J. van Zuylen and D. M. Branston, "Consistent link flow estimation from counts", *Transportation Research Part B: Methodological*, vol. 16, no. 6, pp. 473–476, Dec. 1982, ISSN: 0191-2615. DOI: 10.1016/0191-2615(82)90006-6 (cited p. 36).

[52] H. Spiess, "A maximum likelihood model for estimating origin-destination matrices", *Transportation Research Part B: Methodological*, vol. 21, no. 5, pp. 395–412, Oct. 1987, ISSN: 0191-2615. DOI: 10.1016/0191-2615(87)90037-3 (cited p. 36).

[53] M. E. Ben-Akiva, "Methods to combine different data sources and estimate origin-destination matrices", *Transportation and traffic theory*, 1987 (cited p. 36).

[54] R. J. Vanderbei and J. Iannone, "An EM approach to OD matrix estimation", *Technical Report SOR-94-04*, 1994 (cited p. 36).

[55] H. Lo, N. Zhang, and W. Lam, "Decomposition algorithm for statistical estimation of OD matrix with random link choice proportions from traffic counts", *Transportation Research Part B: Methodological*, vol. 33, no. 5, pp. 369–385, Jun. 1999, ISSN: 0191-2615. DOI: 10.1016/S0191-2615(98)00042-3 (cited pp. 36, 37).

[56] M. Maher, "Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach", *Transportation Research Part B: Methodological*, vol. 17, no. 6, pp. 435–447, Dec. 1983, ISSN: 0191-2615. DOI: 10.1016/0191-2615(83)90030-9 (cited pp. 36, 37).

[57] B. Li, "Bayesian inference for origin-destination matrices of transport networks using the EM algorithm", *Technometrics*, vol. 47, no. 4, pp. 399–408, Nov. 2005, ISSN: 00401706. DOI: 10.1198/004017005000000283 (cited p. 36).

[58] M. L. Hazelton, "Bayesian inference for network-based models with a linear inverse structure", *Transportation Research Part B: Methodological*, Bayesian Methods, vol. 44, no. 5, pp. 674–685, Jun. 2010, ISSN: 0191-2615. DOI: 10.1016/j.trb.2010.01.006 (cited p. 36).

[59] E. Castillo, P. Jiménez, J. M. Menéndez, and M. Nogal, "A Bayesian method for estimating traffic flows based on plate scanning", *Transportation*, vol. 40, no. 1, pp. 173–201, Jan. 1, 2013, ISSN: 0049-4488, 1572-9435. DOI: 10.1007/s11116-012-9443-4 (cited pp. 36, 39).

[60] K. Perrakis, D. Karlis, M. Cools, and D. Janssens, "Bayesian inference for transportation origin-destination matrices: The Poisson-inverse Gaussian and other Poisson mixtures", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 178, no. 1, pp. 271–296, 2015, ISSN: 1467-985X. DOI: 10.1111/rssa.12057 (cited pp. 36, 103).

[61] M. Carey, C. Hendrickson, and K. Siddharthan, "A method for direct estimation of origin/destination trip matrices", *Transportation Science*, vol. 15, no. 1, pp. 32–49, 1981. DOI: 10.1287/trsc.15.1.32 (cited p. 37).

[62] E. Cascetta, "Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator", *Transportation Research Part B*, vol. 18, no. 4-5, pp. 289–299, 1984, ISSN: 01912615. DOI: 10.1016/0191-2615(84)90012-2 (cited p. 37).

[63] M. G. Bell, "The estimation of origin-destination matrices by constrained generalised least squares", *Transportation Research Part B: Methodological*, vol. 25, pp. 13–22, 1991, ISSN: 0191-2615. DOI: 10.1016/0191-2615(91)90010-g (cited p. 37).

[64] H. D. Sherali and T. Park, "Estimation of dynamic origin-destination trip tables for a general network", *Transportation Research Part B: Methodological*, vol. 35, no. 3, pp. 217–235, 2001. DOI: 10.1016/s0191-2615(99)00048-x (cited pp. 37, 38).

[65] Y. Nie, H. M. Zhang, and W. W. Recker, "Inferring origin-destination trip matrices with a decoupled GLS path flow estimator", *Transportation Research Part B: Methodological*, vol. 39, no. 6, pp. 497–518, Jul. 2005, ISSN: 0191-2615. DOI: 10.1016/j.trb.2004.07.002 (cited p. 37).

[66] E. Codina and J. Barceló, "Adjustment of O-D trip matrices from observed volumes: An algorithmic approach based on conjugate directions", *European Journal of Operational Research*, Traffic and Transportation Systems Analysis, vol. 155, no. 3, pp. 535–557, Jun. 16, 2004, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2003.08.004 (cited p. 37).

[67] C. Xie, K. M. Kockelman, and S. T. Waller, "A maximum entropy-least squares estimator for elastic origin-destination trip matrix estimation", *Transportation Research Part B: Methodological*, Select Papers from the 19th ISTTT, vol. 45, no. 9, pp. 1465–1482, Nov. 2011, ISSN: 0191-2615. DOI: 10.1016/j.trb.2011.05.018 (cited p. 37).

[68] K. Parry and M. L. Hazelton, "Estimation of origin-destination matrices from link counts and sporadic routing data", *Transportation Research Part B: Methodological*, vol. 46, no. 1, pp. 175–188, Jan. 2012, ISSN: 0191-2615. DOI: 10.1016/j.trb.2011.09.009 (cited pp. 37, 39).

[69] R. B. Smock, *An Iterative Assignment Approach to Capacity Restraint on Arterial Networks*. Wayne State University, 1962, 40 pp. (cited p. 37).

[70] J. Holm, T. Jensen, S. Nielsen, A. Christensen, B. Johnsen, and G. Ronby, "Calibrating traffic models on traffic census results only", *Traffic Engineering and Control*, vol. 17, Apr. 1976 (cited p. 37).

[71] M. Yousefikia, A. R. Mamdoohi, and M. Noruzoliaee, "Iterative update of route choice proportions in OD estimation", *Proceedings of the Institution of Civil Engineers - Transport*, vol. 169, no. 1, pp. 53–60, Jan. 13, 2016, ISSN: 0965-092X. DOI: 10.1680/jtran.12.00071 (cited p. 37).

[72] H. Yang, T. Sasaki, Y. Iida, and Y. Asakura, "Estimation of origin-destination matrices from link traffic counts on congested networks", *Transportation Research Part B: Methodological*, vol. 26, no. 6, pp. 417–434, Dec. 1992, ISSN: 0191-2615. DOI: 10.1016/0191-2615(92)90008-K (cited p. 37).

[73] Y. Sheffi, "Urban transportation networks: Equilibrium analysis with mathematical programming methods", *Transportation Research Part A: General*, vol. 21, no. 6, pp. 481–484, 1985. DOI: 10.1016/0191-2607(87)90038-0 (cited p. 37).

[74] B. Colson, P. Marcotte, and G. Savard, "Bilevel programming: A survey", *4OR*, vol. 3, no. 2, pp. 87–107, Jun. 2005, ISSN: 1619-4500, 1614-2411. DOI: 10.1007/s10288-005-0071-0 (cited p. 37).

[75] H. Yang, Q. Meng, and M. G. Bell, "Simultaneous estimation of the origin-destination matrices and travel-cost coefficient for congested networks in a stochastic user equilibrium", *Transportation Science*, vol. 35, no. 2, pp. 107–123, 2001. DOI: 10.1287/trsc.35.2.107.10133 (cited p. 37).

[76] M. J. Maher, X. Zhang, and D. Van Vliet, "A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows", *Transportation Research Part B: Methodological*, vol. 35, no. 1, pp. 23–40, 2001. DOI: 10.1016/s0191-2615(00)00017-5 (cited p. 37).

[77] S. Erlander, S. Nguyen, and N. Stewart, "On the calibration of the combined distribution-assignment model", *Transportation Research Part B: Methodological*, vol. 13, no. 3, pp. 259–267, Sep. 1979, ISSN: 0191-2615. DOI: 10.1016/0191-2615(79)90017-1 (cited p. 38).

[78] C. Fisk and D. Boyce, "A note on trip matrix estimation from link traffic count data", *Transportation Research Part B: Methodological*, vol. 17, no. 3, pp. 245–250, Jun. 1983, ISSN: 0191-2615. DOI: 10.1016/0191-2615(83)90018-8 (cited p. 38).

[79] C. Fisk, "Trip matrix estimation from link traffic counts: The congested network case", *Transportation Research Part B: Methodological*, vol. 23, no. 5, pp. 331–336, Oct. 1989, ISSN: 0191-2615. DOI: 10.1016/0191-2615(89)90009-X (cited p. 38).

[80] L. G. Willumsen, "Estimating time-dependent trip matrices from traffic counts", in *Ninth International Symposium on Transportation and Traffic Theory, VNU Science Press*, 1984, pp. 397–411 (cited p. 38).

[81] J. Wu, "A real-time origin-destination matrix updating algorithm for on-line applications", *Transportation Research Part B: Methodological*, vol. 31, no. 5, pp. 381–396, Oct. 1997, ISSN: 0191-2615. DOI: 10.1016/S0191-2615(97)00001-5 (cited p. 38).

[82] E. Cascetta and G. Marquis, "Dynamic estimators of origin-destination matrices using traffic counts", *Transportation Science*, vol. 27, no. 4, pp. 363–373, 1993, ISSN: 00411655. DOI: 10.1287/trsc.27.4.363 (cited p. 38).

[83] H. Tavana, "Internally-consistent estimation of dynamic network origin-destination flows from intelligent transportation systems data using bi-level optimization", 2001. [Online]. Available: http://repositories.lib.utexas.edu/bitstream/handle/2152/1662/tavanah66033.pdf (visited on 04/04/2016) (cited p. 38).

[84] K. Ashok and M. E. Ben-Akiva, "Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows", *Transportation Science*, vol. 36, no. 2, pp. 184–198, 2002. DOI: 10.1287/trsc.36.2.184.563 (cited p. 38).

[85] M. L. Hazelton, "Estimation of origin-destination matrices from link flows on uncongested networks", *Transportation Research: Part B: Methodological*, vol. 34, no. 7, pp. 549–566, 2000, ISSN: 0191-2615. DOI: 10.1016/s0191-2615(99)00037-5 (cited p. 38).

[86] M. L. Hazelton, "Some comments on origin-destination matrix estimation", *Transportation Research Part A: Policy and Practice*, vol. 37, pp. 811–822, 2003, 10, ISSN: 0965-8564. DOI: 10.1016/s0965-8564(03)00044-2 (cited p. 38).

[87] C. D. R. Lindveld, "Dynamic OD matrix estimation: A behavioural approach", TU Delft, Delft University of Technology, 2003 (cited p. 38).

[88] C.-C. Lu, X. Zhou, and K. Zhang, "Dynamic origin-destination demand flow estimation under congested traffic conditions", *Transportation Research Part C: Emerging Technologies*, vol. 34, pp. 16–37, Sep. 2013, ISSN: 0968-090X. DOI: 10.1016/j.trc.2013.05.006 (cited p. 38).

[89] D. C. Gazis and C. H. Knapp, "On-line estimation of traffic densities from time-series of flow and speed data", *Transportation Science*, vol. 5, no. 3, pp. 283–301, Aug. 1971, ISSN: 0041-1655, 1526-5447. DOI: 10.1287/trsc.5.3.283 (cited p. 38).

[90] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory", *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, Feb. 1984, ISSN: 0191-2615. DOI: 10.1016/0191-2615(84)90002-X (cited p. 38).

[91] M. Cremer and H. Keller, "A new class of dynamic methods for the identification of origin-destination flows", *Transportation Research Part B: Methodological*, vol. 21, no. 2, pp. 117–132, Apr. 1987, ISSN: 0191-2615. DOI: 10.1016/0191-2615(87)90011-7 (cited pp. 38, 111).

[92] K. Ashok and M. E. Ben-Akiva, "Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems", in *International Symposium on the Theory of Traffic Flow and Transportation (12th: 1993: BERKELEY, Calif.). Transportation and Traffic Theory*, 1993 (cited p. 38).

[93] ——, "Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows", *Transportation Science*, vol. 34, no. 1, pp. 21–36, 2000. DOI: 10.1287/trsc.34.1.21.12282 (cited p. 38).

[94] B. Li and B. De Moor, "Dynamic identification of origin–destination matrices in the presence of incomplete observations", *Transportation Research Part B: Methodological*, vol. 36, no. 1, pp. 37–57, 2002. DOI: 10.1016/s0191-2615(00)00037-0 (cited p. 38).

[95] R. Frederix, F. Viti, and C. M. Tampère, "A hierarchical approach for dynamic origin-destination matrix estimation on large-scale congested networks", in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference On*, IEEE, 2011, pp. 1543–1548, ISBN: 1-4577-2198-8. DOI: 10.1109/itsc.2011.6082901 (cited p. 38).

[96] J. Barceló and L. Montero, "A robust framework for the estimation of dynamic od trip matrices for reliable traffic management", *Transportation Research Procedia*, 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands, vol. 10, pp. 134–144, 2015, ISSN: 2352-1465. DOI: 10.1016/j.trpro.2015.09.063 (cited p. 38).

[97] G. G. Makowski and K. C. Sinha, "A statistical procedure to analyze partial license plate numbers", *Transportation Research*, vol. 10, no. 2, pp. 131–132, Apr. 1976, ISSN: 0041-1647. DOI: 10.1016/0041-1647(76)90049-6 (cited p. 38).

[98] N. J. Van Der Zijpp, "Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1607, pp. 87–94, 1997, ISSN: 0361-1981. DOI: 10.3141/1607-13 (cited p. 38).

[99] D. P. Watling and M. J. Maher, "A statistical procedure for estimating a mean origin-destination matrix from a partial registration plate survey", *Transportation Research Part B: Methodological*, vol. 26, pp. 171–193, 1992, 3, ISSN: 0191-2615. DOI: 10.1016/0191-2615(92)90023-p (cited p. 38).

[100] D. P. Watling, "Maximum likelihood estimation of an origin-destination matrix from a partial registration plate survey", *Transportation Research Part B: Methodological*, vol. 28, pp. 289–314, 1994, 4, ISSN: 0191-2615. DOI: 10.1016/0191-2615(94)90003-5 (cited p. 38).

[101] Y. Asakura, E. Hato, and M. Kashiwadani, "Origin-destination matrices estimation model using automatic vehicle identification data and its application to the Han-Shin expressway network", *Transportation*, vol. 27, no. 4, pp. 419–438, 2000 (cited p. 38).

[102] M. P. Dixon and L. R. Rilett, "Real-time OD estimation using automatic vehicle identification and traffic count data", *Computer-Aided Civil and Infrastructure Engineering*, vol. 17, no. 1, pp. 7–21, 2002, ISSN: 1467-8667. DOI: 10.1111/1467-8667.00248 (cited p. 39).

[103] X. Zhou and H. Mahmassani, "Dynamic origin-destination demand estimation using automatic vehicle identification data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 105–114, 2006, ISSN: 1524-9050. DOI: 10.1109/TITS.2006.869629 (cited p. 39).

[104] K. Teknomo and P. Fernandez, "A theoretical foundation for the relationship between generalized origin-destination matrix and flow matrix based on ordinal graph trajectories", *Journal of Advanced Transportation*, vol. 48, no. 6, pp. 608–626, Oct. 1, 2014, ISSN: 2042-3195. DOI: 10.1002/atr.1214 (cited p. 39).

[105] ——, (Oct. 2012). A theoretical foundation for the relationship between generalized origin-destination matrix and flow matrix based on ordinal graph trajectories: OD and flow matrices from trajectories, Journal of Advanced Transportation, [Online]. Available: http://people.revoledu.com/kardi/research/trajectory/od/index.html (visited on 09/10/2014) (cited p. 39).

[106] Y. Feng, J. Sun, and P. Chen, "Vehicle trajectory reconstruction using automatic vehicle identification and traffic count data", *Journal of Advanced Transportation*, vol. 49, no. 2, pp. 174–194, Mar. 1, 2015, ISSN: 2042-3195. DOI: 10.1002/atr.1260 (cited pp. 39, 43, 71, 95, 126, 174).

[107] E. Castillo, J. M. Menéndez, and P. Jiménez, "Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations", *Transportation Research Part B: Methodological*, vol. 42, no. 5, pp. 455–481, Jun. 2008, ISSN: 0191-2615. DOI: 10.1016/j.trb.2007.09.004 (cited pp. 39, 43, 71).

[108] J. Kwon and P. Varaiya, "Real-time estimation of origin-destination matrices with partial trajectories from electronic toll collection tag data", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1923, no. 1, pp. 119–126, 2005. DOI: 10.3141/1923-13 (cited p. 39).

[109] Tiratanapakhom Tawin, "Analyses on travel demand variation considering users' choice behaviors on urban expressway using ETC data", THE UNIVERSITY OF TOKYO, Sep. 2013, 162 pp. (cited p. 40).

[110] J. Kim, F. Kurauchi, N. Uno, T. Hagihara, and T. Daito, "Using electronic toll collection data to understand traffic demand", *Journal of Intelligent Transportation Systems*, vol. 18, no. 2, pp. 190–203, 2014, ISSN: 1547-2450, 1547-2442. DOI: 10.1080/15472450.2013.806858 (cited p. 40).

[111] H. ZHAO, L. YU, J. GUO, N. ZHAO, H. WEN, and L. ZHU, "Estimation of time-varying OD demands incorporating FCD and RTMS data", *Journal of Transportation Systems Engineering and Information Technology*, vol. 10, pp. 72–80, 2010, 1, ISSN: 1570-6672. DOI: 10.1016/s1570-6672(09)60024-6 (cited p. 40).

[112] J. C. Herrera, D. B. Work, R. Herring, X. (Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment", *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568–583, Aug. 2010, ISSN: 0968-090X. DOI: 10.1016/j.trc.2009.10.006 (cited pp. 40, 95).

[113] S. N. Hadjidimitriou, M. Dell'Amico, G. Cantelmo, and F. Viti, "Assessing the consistency between observed and modelled route choices through GPS data", in *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Jun. 2015, pp. 216–222. DOI: 10.1109/MTITS.2015.7223259 (cited p. 40).

[114] S. Liu, C. Liu, Q. Luo, L. Ni, and R. Krishnan, "Calibrating large scale vehicle trajectory data", in *2012 IEEE 13th International Conference on Mobile Data Management (MDM)*, Jul. 2012, pp. 222–231. DOI: 10.1109/MDM.2012.15 (cited p. 40).

[115] J.-L. Ygnace, C. Drane, Y. B. Yim, and R. De Lacvivier, "Travel time estimation on the san francisco bay area network using cellular phones as probes", *California Partners for Advanced Transit and Highways (PATH)*, 2000 (cited p. 40).

[116] Y. B. Yim and R. Cayford, "Investigation of vehicles as probes using global positioning system and cellular phone tracking: Field operational test", *California Partners for Advanced Transit and Highways (PATH)*, 2001 (cited p. 40).

[117] A. Alessandri, R. Bolla, and M. Repetto, "Estimation of freeway traffic variables using information from mobile phones", in *American Control Conference, 2003. Proceedings of the 2003*, vol. 5, Jun. 2003, 4089–4094 vol.5. DOI: 10.1109/ACC.2003.1240476 (cited p. 40).

[118] E. Mellegard, S. Moritz, and M. Zahoor, "Origin/destination-estimation using cellular network data", in *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, Dec. 2011, pp. 891–896. DOI: 10.1109/ICDMW.2011.132 (cited p. 40).

[119] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome", *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 141–151, 2011. DOI: 10.1109/tits.2010.2074196 (cited p. 40).

[120] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data", *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 36–44, Apr. 2011, ISSN: 1536-1268. DOI: 10.1109/MPRV.2011.41 (cited p. 40).

[121] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr., and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example", *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, Jan. 2013, ISSN: 0968-090X. DOI: 10.1016/j.trc.2012.09.009 (cited p. 40).

[122] W. Jing, W. Dianhai, S. Xianmin, and S. Di, "Dynamic OD expansion method based on mobile phone location", in *2011 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, vol. 1, Mar. 2011, pp. 788–791. DOI: 10.1109/ICICTA.2011.204 (cited p. 40).

[123] J. Ma, H. Li, F. Yuan, and T. Bauer, "Deriving operational origin-destination matrices from large scale mobile phone data", *International Journal of Transportation Science and Technology*, vol. 2, no. 3, pp. 183–204, Sep. 1–3, 2013. DOI: 10.1260/2046-0430.2.3.183 (cited p. 40).

[124] A. N. Larijani, A.-M. Olteanu-Raimond, J. Perret, M. Brédif, and C. Ziemlicki, "Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region", *Transportation Research Procedia*, 4th International Symposium of Transport Simulation (ISTS'14) Selected Proceedings, Ajaccio, France, 1-4 June 2014, vol. 6, pp. 64–78, 2015, ISSN: 2352-1465. DOI: 10.1016/j.trpro.2015.03.006 (cited p. 40).

[125] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin-destination matrices using mobile phone call data", *Transportation Research Part C: Emerging Technologies*, vol. 40, pp. 63–74, Mar. 2014, ISSN: 0968-090X. DOI: 10.1016/j.trc.2014.01.002 (cited pp. 40, 41, 73).

[126] J. Barceló, L. Montero, L. Marqués, and C. Carmona, "A Kalman-filter approach for dynamic od estimation in corridors based on Bluetooth and wifi data collection", in *Proceedings 12th World Conf. on Transportation Research*, 2010, pp. 11–15 (cited pp. 40, 42).

[127] M. Blogg, C. Semler, M. Hingorani, and R. Troutbeck, "Travel time and origin-destination data collection using Bluetooth MAC address readers", in *Australasian Transport Research Forum (ATRF), 33rd, 2010, Canberra, ACT, Australia*, 2010 (cited pp. 40, 42).

[128] R. M. Reiff, "Determination of origin-destination using Bluetooth technology", in *Institute of Transportation Engineers Annual Meeting and Exhibit 2012, August 12, 2012 - August 15, 2012*, ser. Institute of Transportation Engineers Annual Meeting and Exhibit 2012, Institute of Transportation Engineers, 2012, pp. 23–30 (cited p. 40).

[129] C. Carpenter, M. Fowler, and T. J. Adler, "Generating route-specific origin-destination tables using Bluetooth technology", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2308, pp. 96–102, 2012, ISSN: 0361-1981. DOI: 10.3141/2308-10 (cited pp. 40, 43, 71, 73, 105).

[130] J. Barceló, L. Montero, M. Bullejos, O. Serch, and C. Carmona, "Dynamic OD matrix estimation exploiting Bluetooth data in urban networks", in *Proceedings of the 14th International Conference on Automatic Control, Modelling & Simulation, and Proceedings of the 11th International Conference on Microelectronics, Nanoelectronics, Optoelectronics*, World Scientific and Engineering Academy and Society (WSEAS), 2012, pp. 116–121, ISBN: 1-61804-080-4 (cited p. 40).

[131] S. Yucel, H. Tuydes-Yaman, O. Altintasi, and M. Ozen, "Determination of vehicular travel patterns in an urban location using Bluetooth technology", in *ITS AMERICA ANNUAL MEETING and Expo, Nashville, Tenn*, 2013 (cited p. 40).

[132] Y. Canon-Lozano, A. Melo-Castillo, K. Banse, and L. Felipe Herrera-Quintero, "Automatic generation of O/D matrix for mass transportation systems using an its approach", in *Intelligent Transportation Systems Symposium (CITSS), 2012 IEEE Colombian*, IEEE, 2012, pp. 1–6, ISBN: 1-4673-4360-9. DOI: 10.1109/citss.2012.6336681 (cited p. 40).

[133] L. Montero Mercadé, J. Barceló Bugeda, and E. Codina Sancho, "Adapting a dynamic OD matrix estimation approach for private traffic based on Bluetooth data to passenger OD matrices", 2012 (cited p. 40).

[134] N. Abedi, A. Bhaskar, and E. Chung, "Tracking spatio-temporal movement of human in terms of space utilization using media-access-control address data", *Applied Geography*, vol. 51, pp. 72–81, Jul. 2014, ISSN: 0143-6228. DOI: 10.1016/j.apgeog.2014.04.001 (cited p. 40).

[135] N. Abedi, A. Bhaskar, E. Chung, and M. Miska, "Assessment of antenna characteristic effects on pedestrian and cyclists travel-time estimation based on Bluetooth and wifi MAC addresses", *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 124–141, Nov. 2015, ISSN: 0968-090X. DOI: 10.1016/j.trc.2015.08.010 (cited pp. 40, 58, 84).

[136] L. Bianco, G. Confessore, and M. Gentili, "Combinatorial aspects of the sensor location problem", *Annals of Operations Research*, vol. 144, no. 1, pp. 201–234, Apr. 1, 2006, ISSN: 0254-5330, 1572-9338. DOI: 10.1007/s10479-006-0016-9 (cited p. 40).

[137] T. M. Brennan Jr, J. M. Ernst, C. M. Day, D. M. Bullock, J. V. Krogmeier, and M. Martchouk, "Influence of vertical sensor placement on data collection efficiency from Bluetooth MAC address collection devices", *Journal of Transportation Engineering*, vol. 136, pp. 1104–1109, 2010, 12, ISSN: 0733-947X. DOI: 10.1061/(asce)te.1943-5436.0000178 (cited pp. 40, 42, 43, 66).

[138] F. Simonelli, V. Marzano, A. Papola, and I. Vitiello, "A network sensor location procedure accounting for O-D matrix estimate variability", *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1624–1638, Dec. 2012, ISSN: 0191-2615. DOI: 10.1016/j.trb.2012.08.007 (cited p. 40).

[139] S.-R. Hu and H.-T. Liou, "A generalized sensor location model for the estimation of network origin-destination matrices", *Transportation Research Part C: Emerging Technologies*, vol. 40, pp. 93–110, Mar. 2014, ISSN: 0968-090X. DOI: 10.1016/j.trc.2014.01.004 (cited p. 40).

[140] R. Nusser and R. Pelz, "Bluetooth-based wireless connectivity in an automotive environment", in *Vehicular Technology Conference, 2000. IEEE-VTS Fall VTC 2000. 52nd*, vol. 4, 2000, 1935–1942 vol.4. DOI: 10.1109/VETECF.2000.886152 (cited p. 42).

[141] P. Murphy, E. Welsh, and J. Frantz, "Using Bluetooth for short-term ad hoc connections between moving vehicles: A feasibility study", in *Vehicular Technology Conference, 2002. VTC Spring 2002. IEEE 55th*, vol. 1, 2002, 414–418 vol.1. DOI: 10.1109/VTC.2002.1002746 (cited p. 42).

[142] H. Sawant, J. Tan, Q. Yang, and Q. Wang, "Using Bluetooth and sensor networks for intelligent transportation systems", in *The 7th International IEEE Conference on Intelligent Transportation Systems, 2004. Proceedings*, Oct. 2004, pp. 767–772. DOI: 10.1109/ITSC.2004.1398999 (cited p. 42).

[143] B. Potter, "Wireless-based location tracking", *Network Security*, vol. 2003, no. 11, pp. 4–5, Nov. 2003, ISSN: 1353-4858. DOI: 10.1016/S1353-4858(03)01105-X (cited p. 42).

[144] J. S. Wasson, J. R. Sturdevant, and D. M. Bullock, "Real-time travel time estimates using media access control address matching", *Institute of Transportation Engineers. ITE Journal*, vol. 78, no. 6, pp. 20–23, Jun. 2008, ISSN: 01628178 (cited p. 42).

[145] D. D. Puckett and M. J. Vickich, "Bluetooth-based travel time/speed measuring systems development", 2010 (cited p. 42).

[146] A. Haghani, M. Hamedi, K. Sadabadi, S. Young, and P. Tarnoff, "Data collection of freeway travel time ground truth with Bluetooth sensors", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2160, pp. 60–68, Sep. 10, 2010, ISSN: 0361-1981. DOI: 10.3141/2160-07 (cited p. 42).

[147] S. Quayle, P. Koonce, D. DePencier, and D. Bullock, "Arterial performance measures with media access control readers", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2192, pp. 185–193, Dec. 1, 2010, ISSN: 0361-1981. DOI: 10.3141/2192-18 (cited p. 42).

[148] Y. Malinovskiy, U.-K. Lee, Y.-J. Wu, and Y. Wang, "Investigation of Bluetooth-based travel time estimation error on a short corridor", in *Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, DC*, 2011 (cited p. 42).

[149] A. Hainen, J. Wasson, S. Hubbard, S. Remias, G. Farnsworth, and D. Bullock, "Estimating route choice and travel time reliability with field observations of Bluetooth probe vehicles", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2256, pp. 43–50, Dec. 1, 2011. DOI: 10.3141/2256-06 (cited pp. 42, 43, 71, 95).

[150] B. N. Araghi, R. Krishnan, and H. Lahrmann, "Application of Bluetooth technology for mode-specific travel time estimation on arterial roads: Potentials and challenges", *Trafikdage pa Aalborg Universitet*, 2012, ISSN: 1603-9696 (cited p. 42).

[151] B. N. Araghi, K. S. Pedersen, L. T. Christensen, R. Krishnan, and H. Lahrmann, "Accuracy of travel time estimation using Bluetooth technology: Case study limfjord tunnel aalborg", in *19th ITS World Congress*, Vienna, Austria, 2012 (cited p. 42).

[152] E. Mitsakis, J.-M. S. Grau, E. Chrysohoou, and G. Aifadopoulou, "A robust method for real time estimation of travel times for dense urban road networks using point-to-point detectors", in *Transportation Research Board 92nd Annual Meeting*, 13-3654, 2013. DOI: 10.3846/16484142.2015.1078845 (cited p. 42).

[153] J. Cox, "Development of a permanent system to record and analyse Bluetooth travel time and SCATS lane count data", in *AITPM 2013 National Conference*, Perth, Western Australia, 2013 (cited p. 42).

[154] A. Bhaskar and E. Chung, "Fundamental understanding on the use of Bluetooth scanner as a complementary transport data", *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 42–72, Dec. 2013, ISSN: 0968090X. DOI: 10.1016/j.trc.2013.09.013 (cited p. 42).

[155] B. N. Araghi, J. H. Olesen, R. Krishnan, L. T. Christensen, and H. Lahrmann, "Reliability of Bluetooth technology for travel time estimation", *Journal of Intelligent Transportation Systems*, vol. 19, no. 3, 2015, ISSN: 1547-2450. DOI: 10.1080/15472450.2013.856727 (cited p. 42).

[156] A. Bhaskar, M. Qu, and E. Chung, "Bluetooth vehicle trajectory by fusing Bluetooth and loops: Motorway travel time statistics", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 113–122, Feb. 2015, ISSN: 1524-9050. DOI: 10.1109/TITS.2014.2328373 (cited pp. 42, 43).

[157] T. Tsubota, A. Bhaskar, E. Chung, and R. Billot, "Arterial traffic congestion analysis using Bluetooth duration data", in *Australasian Transport Research Forum 2011,*, Adelaide, South Australia, Sep. 28–30, 2011 (cited pp. 42, 57).

[158] M. Delafontaine, M. Versichele, T. Neutens, and N. Van de Weghe, "Analysing spatiotemporal sequences in Bluetooth tracking data", *Applied Geography*, vol. 34, pp. 659–668, 2012, ISSN: 0143-6228. DOI: 10.1016/j.apgeog.2012.04.003 (cited p. 42).

[159] P.-A. Laharotte, R. Billot, E. Come, L. Oukhellou, A. Nantes, and N.-E. El Faouzi, "Spatiotemporal analysis of Bluetooth data: Application to a large urban network", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1439–1448, Jun. 2015, ISSN: 1524-9050. DOI: 10.1109/TITS.2014.2367165 (cited pp. 42, 60, 104).

[160] M. R. Friesen and R. D. McLeod, "Bluetooth in intelligent transportation systems: A survey", *International Journal of Intelligent Transportation Systems Research*, vol. 13, no. 3, pp. 143–153, May 29, 2014, ISSN: 1348-8503, 1868-8659. DOI: 10.1007/s13177-014-0092-1 (cited p. 42).

[161] B. S. Peterson, R. O. Baldwin, and J. P. Kharoufeh, "A specification-compatible Bluetooth inquiry simplification", in *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, IEEE, Jan. 5, 2004, ISBN: 0-7695-2056-1. DOI: 10.1109/HICSS.2004.1265720 (cited pp. 42, 55).

[162] "Technique for automated MAC address cloning", US 7,342,925 B2, Mar. 11, 2008 (cited pp. 43, 62).

[163] A. Bhaskar, L. M. Kieu, M. Qu, A. Nantes, M. Miska, and E. Chung, "Is bus overrepresented in Bluetooth MAC scanner data? Is MAC-ID really unique?", *International Journal of Intelligent Transportation Systems Research*, vol. 13, no. 2, pp. 119–130, May 2015, ISSN: 1348-8503, 1868-8659. DOI: 10.1007/s13177-014-0089-9 (cited p. 43).

[164] S. Hay and R. Harle, "Bluetooth tracking without discoverability", in *Location and Context Awareness*, Springer, 2009, pp. 120–137, ISBN: 3-642-01720-7 (cited p. 43).

[165] A. Franssens, "Impact of multiple inquires on the Bluetooth discovery process: And its application to localization", Master Thesis, Jul. 2010 (cited pp. 43, 66).

[166] J. D. Porter, D. S. Kim, M. E. Magaña, P. Poocharoen, and C. A. G. Arriaga, "Antenna characterization for Bluetooth-based travel time data collection", *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 17, no. 2, pp. 142–151, 2013, ISSN: 1547-2450. DOI: 10.1080/15472450.2012.696452 (cited pp. 43, 66).

[167] B. Coifman, "Estimating travel times and vehicle trajectories on freeways using dual loop detectors", *Transportation Research Part A: Policy and Practice*, vol. 36, no. 4, pp. 351–364, May 2002, ISSN: 0965-8564. DOI: 10.1016/S0965-8564(01)00007-6 (cited p. 43).

[168] P. Davies, "Vehicle detection and classification", *Information Technology Applications in Transport*, pp. 11–40, 1987 (cited p. 50).

[169] Federal Highway Administration, "Chapter 2. SENSOR TECHNOLOGY", in *Traffic Detector Handbook: THIRD EDITION- FHWA-HRT-06-108*, vol. Volume I, May 2006 (cited p. 50).

[170] SIG. (2002). IEEE SA - 802.15.1-2002 - IEEE Standard for Telecommunications and Information Exchange Between Systems, [Online]. Available: http://standards.ieee.org/findstds/standard/802.15.1-2002.html (visited on 02/09/2016) (cited p. 51).

[171] Australian Government Geoscience. (May 15, 2014). Australian Map Grid, Australian Governement - Geoscience Australia (cited p. 53).

[172] SIG. (2016). Bluetooth core specification 4.2: Adopted specifications, [Online]. Available: https://www.bluetooth.com/specifications/adopted-specifications (visited on 02/10/2016) (cited p. 55).

[173] P.-A. Laharotte, R. Billot, E. Côme, L. Oukhellou, A. Nantes, and N.-E. El Faouzi, "Spatio temporal clustering analysis of Bluetooth OD (B-OD) matrices: Application to a large urban network", in *93rd Transport Research Board*, Washington DC, 2014 (cited p. 60).

[174] B. Coifman, "Using dual loop speed traps to identify detector errors", *Transportation Research Record: Journal of the Transportation Research Board*, no. 1683, pp. 47–58, 1999, ISSN: 0361-1981. DOI: 10.3141/1683-07 (cited pp. 60, 104).

[175] Courtney Trenwith. (Oct. 3, 2010). City peak-hour speeds down to 35km/h, says RACQ, Brisbane Times (cited pp. 60, 82, 83).

[176] S. Hirai, J. Xing, R. Horiguchi, T. Shiraishi, and M. Kobayashi, "Development of a network traffic simulator for the entire inter-urban expressway network in Japan", *Transportation Research Procedia*, 4th International Symposium of Transport Simulation (ISTS'14) Selected Proceedings, Ajaccio, France, 1-4 June 2014, vol. 6, pp. 285–296, 2015, ISSN: 2352-1465. DOI: 10.1016/j.trpro.2015.03.022 (cited p. 71).

[177] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, "Understanding road usage patterns in urban areas", *Scientific Reports*, vol. 2, Dec. 20, 2012. DOI: 10.1038/srep01001 (cited pp. 73, 116).

[178] L. Fu and L. R. Rilett, "Expected shortest paths in dynamic and stochastic traffic networks", *Transportation Research Part B: Methodological*, vol. 32, no. 7, pp. 499–516, Sep. 1998, ISSN: 0191-2615. DOI: 10.1016/S0191-2615(98)00016-2 (cited p. 75).

[179] L. Fu, D. Sun, and L. R. Rilett, "Heuristic shortest path algorithms for transportation applications: State of the art", *Computers & Operations Research*, Part Special Issue: Operations Research and Data Mining, vol. 33, no. 11, pp. 3324–3343, Nov. 2006, ISSN: 0305-0548. DOI: 10.1016/j.cor.2005.03.027 (cited p. 75).

[180] C. G. Prato, "Route choice modeling: Past, present and future research directions", *Journal of Choice Modelling*, vol. 2, no. 1, pp. 65–100, 2009, ISSN: 1755-5345. DOI: 10.1016/S1755-5345(13)70005-8 (cited p. 75).

[181] C. Moon, J. Kim, G. Choi, and Y. Seo, "An efficient genetic algorithm for the traveling salesman problem with precedence constraints", *European Journal of Operational Research*, vol. 140, no. 3, pp. 606–617, Aug. 1, 2002, ISSN: 0377-2217. DOI: 10.1016/S0377-2217(01)00227-2 (cited p. 75).

[182] E. Bartolini, L. Bodin, and A. Mingozzi, "The traveling salesman problem with pickup, delivery, and ride-time constraints", *Networks*, vol. 67, no. 2, pp. 95–110, Mar. 2016, WOS:000370140100001, ISSN: 0028-3045. DOI: 10.1002/net.21663 (cited p. 75).

[183] T. Sander and P. Wehner, "On A* search with stopover areas", *International Journal of Geographical Information Science*, vol. 23, no. 6, pp. 799–819, 2009. DOI: 10.1080/13658810902885375 (cited p. 75).

[184] (2016). TSS-Transport Simulation Systems, [Online]. Available: https://www.aimsun.com/wp/ (visited on 01/20/2016) (cited p. 77).

[185] Queensland Department of Transport and Main Roads, "Brisbane Strategic Transport Model", *Queensland Government, Brisbane*, 2008 (cited p. 77).

[186] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977. JSTOR: 2984875 (cited p. 86).

[187] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998 (cited p. 86).

[188] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis", *The computer journal*, vol. 41, no. 8, pp. 578–588, 1998. DOI: 10.1093/comjnl/41.8.578 (cited p. 86).

[189] P. Gómez, M. Menéndez, and E. Mérida-Casermeiro, "Evaluation of trade-offs between two data sources for the accurate estimation of origin-destination matrices", *Transportmetrica B: Transport Dynamics*, pp. 1–24, Mar. 26, 2015, ISSN: 2168-0566. DOI: 10.1080/21680566.2015.1025892 (cited p. 95).

[190] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery", *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, 2007. DOI: 10.1109/jstsp.2007.910264 (cited pp. 99, 102).

[191] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, "An introduction to total variation for image analysis", *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, pp. 263–340, 2010 (cited p. 101).

[192] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, and J.-C. Pesquet, "Wavelet-based image deconvolution and reconstruction", *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–34, 2016. DOI: 10.1002/047134608x.w8294 (cited pp. 101, 102).

[193] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators", *Advances in Computational Mathematics*, vol. 38, no. 3, pp. 667–681, Nov. 29, 2011, ISSN: 1019-7168, 1572-9044. DOI: 10.1007/s10444-011-9254-8 (cited p. 102).

[194] N. Komodakis and J.-C. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems", *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, Nov. 2015, ISSN: 1053-5888. DOI: 10.1109/MSP.2014.2377273 (cited p. 102).

[195] J.-J. Moreau, "Proximité et dualité dans un espace Hilbertien", *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965 (cited p. 102).

[196] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, ser. CMS Books in Mathematics Ser. Springer, 2011, ISBN: 978-1-4419-9466-0 (cited pp. 102, 140).

[197] N. Parikh and S. Boyd, "Proximal algorithms", *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 123–231, 2013. DOI: 10.1561/2400000003 (cited p. 102).

[198] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections", *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 97–123, 2011, ISSN: 1053-5888. DOI: 10.1109/MSP.2010.938752 (cited pp. 102, 140).

[199] C. Chaux, J.-C. Pesquet, and N. Pustelnik, "Nested iterative algorithms for convex constrained image recovery problems", *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, p. 33, 2009, ISSN: 19364954. DOI: 10.1137/080727749 (cited pp. 102, 143).

[200] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems", *Inverse Problems*, vol. 23, no. 4, pp. 1495–1518, 2007. DOI: 10.1088/0266-5611/23/4/008 (cited p. 102).

[201] N. Pustelnik, C. Chaux, and J.-C. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization – extended version", *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2450–2462, Nov. 8, 2009. DOI: 10.1109/tip.2011.2128335 (cited p. 102).

[202] D. L. Donoho, "De-noising by soft-thresholding", *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 613–627, 1995. DOI: 10.1109/18.382009 (cited p. 102).

[203] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting", *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005. DOI: 10.1137/050626090 (cited pp. 102, 140).

[204] T. Djukic, J. Barcelò, M. Bullejos, L. Montero, E. Cipriani, H. van Lint, and S. Hoogendoorn, "Advanced traffic data for dynamic origin-destination demand estimation: State of the art and benchmark study", presented at the Transportation Research Board 94th Annual Meeting, 2015 (cited p. 103).

[205] S. Porta, P. Crucitti, and V. Latora, "The network analysis of urban streets: A dual approach", *Physica A: Statistical Mechanics and its Applications*, vol. 369, no. 2, pp. 853–866, Sep. 15, 2006, ISSN: 0378-4371. DOI: 10.1016/j.physa.2005.12.063 (cited p. 103).

[206] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "The algorithms of Kruskal and Prim", in *Introduction to Algorithms*, The MIT Press, 2009, pp. 631–638, ISBN: 978-0-262-03384-8 (cited p. 103).

[207] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's distance as a metric for image retrieval", *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, Nov. 2000, ISSN: 0920-5691, 1573-1405. DOI: 10.1023/A:1026543900054 (cited p. 105).

[208] J. Barceló, L. Montero, M. Bullejos, M. Linares, and O. Serch, *Robustness and Computational Efficiency of Kalman Filter Estimator of Time-Dependent Origin-Destination Matrices*. 2013, 31 pp. (cited p. 111).

[209] Z. Lu, W. Rao, Y.-J. Wu, L. Guo, and J. Xia, "A Kalman filter approach to dynamic od flow estimation for urban road networks using multi-sensor data", *Journal of Advanced Transportation*, vol. 49, no. 2, pp. 210–227, Mar. 1, 2015, ISSN: 2042-3195. DOI: 10.1002/atr.1292 (cited p. 111).

[210] P. L. Combettes and J.-C. Pesquet, "Stochastic approximations and perturbations in forward-backward splitting for monotone operators", Jul. 25, 2015. arXiv: 1507.07095 (cited p. 111).

[211] H. Rehborn, B. S. Kerner, and R.-P. Schäfer, "Traffic jam warning messages from measured vehicle data with the use of three-phase traffic theory", in *Advanced Microsystems for Automotive Applications 2012*, G. Meyer, Ed., Springer Berlin Heidelberg, 2012, pp. 241–250, ISBN: 978-3-642-29672-7 978-3-642-29673-4. DOI: 10.1007/978-3-642-29673-4_22 (cited pp. 125, 174).

[212] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility", *Scientific Reports*, vol. 3, Mar. 25, 2013, ISSN: 2045-2322. DOI: 10.1038/srep01376 (cited pp. 127, 176).

[213] G. H.-G. Chen and R. T. Rockafellar, "Convergence rates in forward-backward splitting", *SIAM Journal on Optimization*, vol. 7, no. 2, p. 24, May 1997, ISSN: 10526234. DOI: 10.1137/S1052623495290179 (cited p. 140).

List of Publications

Journal Papers

1. G. Michau, A. Nantes, A. Bhaskar, E. Chung, P. Borgnat, and P. Abry, "Bluetooth data in urban context: Retrieving vehicles trajectories", *Submitted in IEEE Transaction on Intelligent Transport Systems*, 2016

2. G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, A. Bhaskar, and E. Chung, "A primal-dual algorithm for link dependent origin destination matrix estimation", *Submitted in IEEE Transactions on Signal and Information Processing Over Networks*, 2016

International Conferences

1. G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving dynamic origin-destination matrices from Bluetooth data", in *Transportation Research Board, 93rd Annual Meeting*, Washington DC, Jan. 12–16, 2014. [Online]. Available: http://eprints.qut.edu.au/66511/

2. G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 19–24, 2015, pp. 5480–5484. DOI: 10.1109/ICASSP.2015.7179019

National Conferences and workshops

1. G. Michau, A. Nantes, and E. Chung, "Towards the retrieval of accurate OD matrices from Bluetooth data: Lessons learned from 2 years of data", in *36th Australasian Transport Research Forum (ATRF)*, QUT, Brisbane, Australia, Oct. 4, 2013. [Online]. Available: http://eprints.qut.edu.au/62727/

2. G. Michau, A. Nantes, E. Chung, P. Abry, and P. Borgnat, "Retrieving trip information from a discrete detectors network: The case of Brisbane Bluetooth detectors", in *CAITR*, Sydney, Feb. 17–18, 2014. [Online]. Available: http://eprints.qut.edu.au/83110/

3. G. Michau, P. Borgnat, N. Pustelnik, P. Abry, A. Nantes, and E. Chung, "Estimating link-dependent origin-destination matrices from sample trajectories and traffic counts", presented at the XXV GRETSI, Lyon, France, Sep. 8, 2015. [Online]. Available: http://eprints.qut.edu.au/86449/

4. G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Graph Signal Processing Workshop*, Philadelphia, May 25–27, 2016

5. G. Michau, P. Abry, P. Borgnat, N. Pustelnik, A. Nantes, and E. Chung, "Estimation of link-dependent origin-destination matrix for traffic on road networks", in *Complex Networks*, Marseilles, France, Jul. 11–13, 2016

Résumé Long du Manuscrit

1 Introduction

Le sujet de recherche ici traité, a émergé en 2013 au Smart Transport Research Centre, (Queensland University of Technology), à Brisbane où depuis plusieurs mois déjà, un réseau important de détecteurs Bluetooth avait été installé pour la collection de données relatives au trafic routier. Si l'objectif primaire de cette installation par la mairie de Brisbane avait pour vocation l'estimation des temps de trajets, il nous est apparu très vite qu'un tel jeu de données avait un fort potentiel d'applications pour de nombreux domaines de l'ingénierie du trafic. De manière générale, l'ingénierie du transport, tout comme de nombreuses autres disciplines, bénéficient aujourd'hui de l'impact des nouvelles technologies : elles permettent la collection d'énormes quantités de données, donnant ainsi son nom à l'ère actuelle : *the Big Data era*.

1.1 Ingénierie du Trafic

Côté transport, l'étude du trafic se base traditionnellement sur les données issues des enquêtes et des comptages de trafic, éventuellement automatisés par l'installation de boucles magnétiques sous les chaussées. Aujourd'hui, l'utilisation des nouvelles technologies vient changer la donne : les GPS, la reconnaissance vidéo, la détection des badges électroniques (e.g., télépéage) et la technologie LiDar, parmi bien d'autres, permettent une identification directe et automatique des véhicules. D'autres technologies, telles que la téléphonie mobile, le Bluetooth et le WiFi, permettent de détecter les appareils électroniques à l'intérieur des véhicules. Toutes ces technologies permettent d'identifier un véhicule (ou un appareil qui lui est lié) tout au long de son mouvement. Elles rentrent donc dans la catégorie dites des *systèmes d'Identification Automatique de Véhicules* (IAV).

Ces nouvelles technologies sont installées à des échelles toujours plus grande : Pour donner un aperçu de cette croissance, le réseau de détecteurs Bluetooth à Brisbane, a commencé en 2007 avec l'installation d'un unique détecteur pour un projet pilote, contenait plus de 600 détecteurs en 2014 et en possède plus de 900 aujourd'hui.

Souvent, l'objectif premier, lors de l'utilisation de ces technologies, est la mesure précise des conditions de trafic. Néanmoins, en permettant l'identification des véhicules, elles apportent aussi des informations importantes pour d'autres applications.

L'Ingénierie du Transport se subdivise en deux sous catégories : D'un côté se trouvent les questions relatives aux *conditions de trafic* qui ont pour objectif d'étudier l'usage du réseau routier par l'estimation de plusieurs indicateurs tels que la vitesse, les temps de trajet, la densité et les volumes

de véhicules, et en établissant les modèles qui relient ces informations. De l'autre côté, l'étude de la *demande*, a pour objectif de comprendre quels facteurs influent sur la mobilité et de quelles façons. Les deux principales approches pour modéliser et représenter la demande sont l'estimation des matrices origines destinations (matrices OD) d'une part et l'analyse des trajectoires d'autre part. Les matrices OD recensent le nombre de véhicules en déplacement ayant la même zone (ou le même quartier) d'origine et de destination. L'étude des trajectoires, elle, vise à expliquer quels itinéraires sont empruntés. Le processus permettant de relier l'*offre de transport*, c'est-à-dire le réseau, avec la *demande*, consiste en l'attribution de trajectoires aux voitures recensées par la matrice OD. Ce processus est appelé *affectation*. La combinaison de ces deux problèmes, l'estimation de la demande et son affectation, est requise pour pouvoir établir une comparaison avec les *conditions de trafic*. Cela est d'ailleurs le défi majeur des problèmes d'estimation de la demande : ils ne peuvent se dissocier des problèmes d'affectation et ces deux étapes sont généralement calibrées par les mêmes observations. La fiabilité de l'estimation de la matrice OD est donc fortement dépendante de la fiabilité de l'étape d'affectation.

Les nouvelles technologies, en permettant l'identification des véhicules, invitent à revisiter cette question et à remettre en cause cette séparation entre les problèmes de *demande* et d'*affectation*. Ces nouvelles données ont le potentiel de fournir à la fois des informations sur la *demande* (origine, destination et trajectoire du véhicule) tout en permettant aussi la mesure des *conditions de trafic* (vitesse, temps de trajet). En particulier, l'accès aux trajectoires permet de reformuler les deux problèmes d'estimation de la demande de de l'affectation en un unique problème. Cette approche est celle proposée dans ce manuscrit : nous proposons l'estimation directe d'un nouvel outil de représentation du trafic : la *matrice origine destination lien*. Cette approche se base sur la disponibilité de jeux de trajectoires, cependant, à l'exception de la technologie GPS, la plupart des nouvelles technologies ne fournissent que des mesures ponctuelles à proximité des détecteurs. La reconstruction des trajectoires à partir de telles données est un deuxième défi auquel cette thèse propose des réponses.

1.2 Traitement du Signal sur Réseaux

L'estimation de la demande de trafic consiste à estimer des quantités inconnues à partir de mesures partielles ou agrégées (par exemple les volumes). Cette catégorie de problèmes est généralement appelée, en Traitement du Signal, *problèmes inverses*. Ce problème étant contraint par une infrastructure sous-jacente, à savoir le réseau routier, il appartient plus précisément à la sous-catégorie des *problèmes inverses sur graphes*. Il est en effet assez direct d'interpréter le réseau routier comme un graphe orienté où les intersections correspondent aux nœuds du graphe et où les routes, reliant les intersections entre elles, correspondent aux liens, orientés selon les directions de conduites autorisées.

L'estimation des matrices OD consiste alors à estimer les volumes entre chaque paires de nœuds du graphe, soit en se basant sur les données issues des enquêtes, qui ne fournissent qu'une information sur une petite fraction de la population, soit en se basant sur des données collectées automatiquement. Jusqu'à récemment, les comptages de voitures étaient la principale source automatisée de collecte de données. Le nombre d'observations ainsi disponibles s'avère faible devant la quantité d'inconnues à estimer et la quantité de paramètres impliqués dans les étapes successives de modélisation et ce problème est donc sous-déterminé. En conséquence, sa résolution nécessite régularisation à partir d'informations complémentaires, la loi de distribution des volumes par exemple.

La formulation de problèmes réels en tant que des problèmes d'estimation sur réseau n'est pas nouvelle. Par exemple, l'analyse des flux de données sur Internet est un problème très similaire et à déjà été traité comme un problème sur réseau dans la littérature [1]–[4]. Un autre exemple, en biologie cette fois, est l'observation et l'analyse de l'expression des gènes afin d'inférer les mécanismes de

régulation [5], [6]. Une des applications les plus célèbres des *problèmes inverses* est peut-être la restauration des images certains pixels bruités ou manquants sont estimés à partir d'informations sur les pixels voisins [7]. Une analogie directe avec le problème de l'estimation du trafic permet d'interpréter les pixels manquants comme les utilisateurs pour lesquels l'information de mobilité est inconnue et les autres pixels comme les utilisateurs détectés par les nouvelles technologies. Ainsi l'image reconstruite correspond à la matrice origine destination lien.

Les avancées récentes en Traitement du Signal ont permis le développement d'algorithmes efficaces généralisant l'algorithme traditionnel de descente de gradient [8], [9], et permettant de traiter des fonctions convexes mais non nécessairement dérivables. Pour ces fonctions, la notion de dérivée est généralisée à l'aide de l'opérateur proximal [10], [11]. Ces méthodes donnent ainsi une plus grande liberté dans le choix et la construction des fonctions objectives à optimiser.

1.3 Objectifs

Cette thèse a pour objectif de tirer profit des nouvelles technologies pour l'estimation du trafic. Elle présente d'abord une méthode de reconstruction des trajectoires à partir de données ponctuelles avec identification des véhicules. Elle propose ensuite une nouvelle formulation du problème d'estimation de la demande de trafic par l'extension du concept de matrice OD à celui de matrice ODL (origine-destination-lien). Le défi de la régularisation, posé par l'estimation de ces matrices ODL, est traité ensuite et deux types de fonctions sont proposées : des fonctions permettant de mesurer la fidélité avec les mesures de l'estimée et des fonctions quantifiant la pertinence de la solution proposée. Ce faisant, l'estimation de la matrice ODL est formulée comme un problème d'estimation convexe non lisse et est résolue à l'aide d'un algorithme proximal primal-dual, basé sur les avancées les plus récentes en Traitement du Signal [12], [13].

1.4 Contributions

Cette thèse vise-t-à démontrer comment les nouvelles technologies, telles que le Bluetooth, permettent de revisiter et de reformuler les deux problèmes de l'estimation des matrices OD et de l'affectation comme un unique problème d'estimation de matrices ODL.

1. Cette thèse garde un lien très fort avec les données réelles issues de la ville de Brisbane. La question du traitement de ces données, issues de la technologie Bluetooth, est donc naturellement un axe fort. La première contribution est de caractériser et d'évaluer les propriétés de tels jeux de données.

2. Une deuxième contribution est le développement d'une méthode afin d'extraire les informations sur les trajets effectués par les véhicules équipés de la technologie Bluetooth : origines, destinations, et trajectoires. L'obtention des trajectoires permet ensuite d'analyser d'avantage la qualité et les propriétés de la technologie Bluetooth.

3. On propose ensuite de reformuler le problème d'estimation des matrices OD et de l'affectation comme un unique problème : celui de l'estimation de la matrice ODL.

4. Le problème de l'estimation des matrices ODL est formulé comme un problème d'optimisation convexe non lisse visant à minimiser une fonction objective construite à partir de plusieurs propriétés importantes du problème : fidélité aux mesures et pertinence de l'estimation avec le problème en question (par exemple, conservation du nombre de voiture).

5. Une fois ce problème formulé, un algorithme proximal primal dual est proposé pour sa résolution.

6. La méthode ainsi formulée est ensuite testée sur un réseau simulé afin de démontrer la réalisabilité du concept.

7. Enfin, l'intégralité des contributions théoriques sont appliquées au cas de Brisbane et des exemples d'applications sont proposés.

2 Résumé des Chapitres

Le manuscrit est donc construit selon la structure suivante. Il est divisé en 6 chapitres dont le contenu est résumé ci-après.

2.1 Chapitre 2

Ce chapitre présente les approches classiques du problème de l'estimation du trafic : estimation des matrices OD et affectation. La revue de littérature présente, dans un premier temps, les diverses approches pour l'estimation des matrices OD et comment ces méthodes peuvent s'interpréter en tant que problèmes inverses. Les méthodes d'affectation utilisées sont présentées ensuite, puis, enfin, la place prise récemment par les nouvelles technologies dans l'ingénierie du trafic.

2.2 Chapitre 3

Ce chapitre présente les jeux de données utilisés dans cette thèse : le réseau routier issu de fichiers GIS, les comptages trafic fournis par la mairie de Brisbane, les données Bluetooth et les données Taxi mises à la disposition du STRC par l'entreprise *Black & White Cabs*.

Le jeu de données Bluetooth est traité plus en détails et son analyse est une des contributions de cette thèse. Ses caractéristiques et son bruit sont mis en évidence, notamment les incertitudes spatio-temporelles dues au large rayon de détection et au protocole de connexion propre à la technologie Bluetooth (d'une durée de 18 secondes à Brisbane). Par ailleurs, un fort taux de détections manquantes est mis en valeur. Toutes ces sources d'incertitudes sont loin d'être négligeables et vont devoir être prises en considération dans l'extraction d'information sur le trafic à partir de ces données.

2.3 Chapitre 4

Alors que les caractéristiques des jeux de données ont été identifiées au chapitre précédent, une première utilisation des données Bluetooth est proposée dans ce chapitre : la reconstruction de trajectoires. Pour cette reconstruction, une méthode en trois étapes est proposée : En premier lieu, l'identification des trajets Bluetooth, composés d'une origine, une destination et des détections intermédiaires. Cela permet en particulier de mesurer les matrices OD Bluetooth. Dans un deuxièmes temps, la reconstruction des trajectoires à partir de ces trajets. Pour ce faire un algorithme de plus court chemin est modifié afin que le chemin final soit contraint par certaines zones géographiques dans le graphe (en l'occurrence ici, les zones de détections intermédiaires). Cet algorithme est testé sur un scénario simulé et sur les données Taxi, pour lesquels une correspondance avec un appareil

Bluetooth a pu être identifiée. La méthode donne des résultats satisfaisants avec 84% des trajectoires correctement reconstruites.

A partir de ces trajectoires, ce chapitre montre d'abord que la discrimination du mode de transport est rendue possible (au moins entre modes motorisés ou non). Enfin, les trajectoires permettent aussi une analyse plus fine des caractéristiques de la technologie Bluetooth, notamment en ce qui concerne les détections manquées. Nous démontrons en particulier qu'il est possible d'interpréter la probabilité de manquer une détection comme une mixture de deux Gaussiennes, correspondant probablement à deux classes de détections.

Dans son ensemble, ce chapitre prouve donc qu'il est possible d'utiliser des technologies, comme le Bluetooth, avec des fortes incertitudes sur les temps et lieux de détection, pour extraire des trajectoires précises. Ces technologies ont généralement un taux de pénétration plus élevé que d'autres technologies plus précises, comme par exemple le GPS, et apparaissent finalement comme une bonne alternative

2.4 Chapitre 5

Ce chapitre introduit le concept de matrice origine destination par lien dont l'estimation permet le traitement conjoint des deux problèmes de l'estimation des matrices OD et de l'estimation de l'affectation. L'estimation de la matrice ODL est un problème difficile et son estimation ne peut se baser uniquement sur les données traditionnelles (comptages et enquêtes). Mais l'accès à des échantillons de trajectoires maintenant rendu possible par les nouvelles technologies (en particulier le Bluetooth comme démontré au chapitre précèdent) permet de traiter maintenant ce problème. L'estimation de la matrice ODL est présentée comme un problème inverse en grande dimension, nécessitant ainsi l'implémentation de fonctions objectives complexes rendant compte de l'ensemble des propriétés que l'estimée doit satisfaire. Dans ce chapitre, la fonction objective proposée est une combinaison linéaire de cinq termes. Deux de ces termes ont pour objectifs d'assurer que l'estimée est cohérente avec les mesures de trafic (trajectoires et comptages routiers), deux autres assurent que deux propriétés essentielles sont satisfaites : d'une part que le nombre de voiture total soit plus élevé que le nombre de trajectoires échantillonnées, d'autre part que le nombre de voiture soit conservé aux intersections. Enfin, le cinquième terme mesure la variation de volume pour les flux de trafic ayant soit des origines, soit destinations proches. Il favorise ainsi les estimées pour lesquels ces trafics sont similaires (en terme d'utilisation du réseau).

Afin de minimiser cette fonction objective complexe dont certains termes ne sont pas dérivables, un algorithme proximal primal dual est adapté à ce problème. La méthode est finalement testée sur un petit réseau simulé. La matrice ODL estimée est comparée à deux autres solutions naïves, toutes trois évaluées par comparaison avec la vraie matrice ODL simulée. Les résultats mettent en valeur la solution estimée par la méthode d'optimisation et démontrent ainsi l'intérêt de la méthode.

2.5 Chapitre 6

Ce dernier chapitre illustre l'ensemble des méthodes proposées dans les chapitres précédents au cas particulier des données obtenues pour la ville de Brisbane. A partir des trajectoires reconstruites grâce à la méthode du Chapitre 4, des comptages de trafic et du réseau routier obtenu par OpenStreetMap, la matrice ODL de Brisbane est estimée selon une méthode similaire au Chapitre 5. En effet certains des termes impliqués dans la fonction objective doivent être légèrement modifiés afin de prendre en compte les réalités propres aux vraies données. La matrice ODL de Brisbane est finalement estimée

pour le mardi 28 juillet 2014 entre 6 et 9 heures du matin. Une comparaison à la solution naïve montre des comportements similaires à ceux observés dans le scénario simulé, laissant ainsi présager la validité de la méthode. Finalement, ce chapitre se clôture sûr des exemples qui mettent en valeur l'utilité et les applications des matrices ODL.

3 Conclusion

Le travail présenté dans ce manuscrit détaille les étapes nécessaires à l'utilisation des données Bluetooth comme un jeu de données complémentaire pour l'ingénierie du trafic. En partant des données brutes, les chapitres successifs se basent sur les résultats des précédents afin de proposer de nouvelles utilisations de ces données. Ce manuscrit détaille successivement comment les données Bluetooth peuvent être utilisées d'abord pour des analyses générales des conditions de trafic, puis pour une lecture plus fine, notamment en permettant la reconstruction de trajectoires. Enfin, ces données combinées avec les données traditionnelles permettent de traiter une extension du problème d'estimation des matrices OD, celui de l'estimation des matrices ODL.

Dans la continuité des autres travaux récents sur l'utilisation des nouvelles technologies en Ingénierie du Transport, ce travail contribue au domaine en développant une méthode de reconstruction des trajectoires et de classification par mode de transport à partir des données Bluetooth. Les nombreuses incertitudes inhérentes à ce jeu de données rendent l'identification du chemin suivi par les véhicules difficiles. De plus, rien dans les données brutes ne permet à priori de distinguer le mode de transport des utilisateurs.

Afin de résoudre ces problèmes, une analyse détaillée des données Bluetooth est proposée et a finalement mené à trois algorithmes : le premier pour extraire les informations relatives aux trajets des véhicules Bluetooth, le deuxième, afin d'interpréter ces trajets en terme de trajectoires et le troisième afin de différencier les modes de transport des utilisateurs. Le deuxième algorithme est une nouvelle extension de l'algorithme classique de calcul des plus courts chemins en contraignant le chemin calculé à passer par certaines zones du graphe. Son efficacité est démontrée grâce à deux scénarii qui montrent que les trajectoires peuvent être ainsi reconstruites à 84%.

Les échantillons de trajectoires sont de plus en plus utilisés comme sources de données primaires (notamment pour l'étude des conditions de trafic ou pour calibrer les modèles de choix de trajectoires [211]). Un accès à de tels échantillons est donc important pour les ingénieurs du transport. Le travail présenté dans ce manuscrit peut être utilisé comme un manuel expliquant les différentes étapes vers l'obtention d'un échantillon de trajectoires à partir de données Bluetooth brutes. Il est par ailleurs important de noter que la technologie utilisée importe peu et que les contributions théoriques proposées ici peuvent s'appliquer à bien d'autre technologies similaires (par exemple le WiFi ou la détection des téléphones mobiles). Ce travail propose une alternative solide aux sources traditionnelles de trajectoires comme les données GPS.

A la lumière de l'importance qu'ont acquise les trajectoires au sein de l'Ingénierie du Transport, plusieurs pistes de développement de ces travaux restent à approfondir. Parmi ces pistes, il est surement possible de modifier l'algorithme de reconstruction de trajectoires pour prendre en compte des paramètres supplémentaires, notamment les détections manquantes, similairement aux travaux proposés par Feng, Sun, and Chen (2015) [106]. Une autre piste pourrait être l'étude de l'impact du temps de connexion du protocole Bluetooth sur la qualité des données collectées, en partenariat avec les ingénieurs de la maire de Brisbane.

Une seconde contribution majeure de cette thèse a été l'extension du concept de matrice OD à celui de matrice ODL.

Alors que traditionnellement, l'estimation de la demande et de l'affectation avaient toujours été traitées comme deux problèmes séparés, ce nouveau concept de matrice LOD combine ces deux aspects. Proposer ce nouvel outil a été motivé par la disponibilité d'un échantillon de trajectoires, alors que tout laisse-t-à penser que l'accès à de tels échantillons va se généraliser à l'avenir. Les trajectoires peuvent alors directement être interprétées comme un échantillon de la matrice LOD. Nous avons démontré dans un premier temps que ce problème, s'il parait de plus grande dimension que le problème classique ne l'est en fait que si la dimension du problème d'affectation est ignorée. Nous avons montré ensuite qu'un jeu de trajectoires pouvait directement s'interpréter comme un échantillonnage de la matrice ODL et que son estimation pouvait alors s'opérer par la formulation d'une fonction objective convexe à minimiser. Ce problème a donc été formulé comme un problème inverse régularisé avec deux termes favorisant la cohérence entre estimées et données et trois termes favorisant le respect de contraintes soit inhérentes au problème, soit importantes en ingénierie du transport.

Ce problème de minimisation d'une fonction objective convexe mais non lisse, donc non différentiable, est résolu par l'adaptation d'un algorithme proximal primal dual à ce problème. Nous avons alors analysé l'impact de chaque terme de la fonction objective, démontrant ainsi leur importance dans le processus. En particulier, cette analyse a souligné le rôle critique du jeu de trajectoires pour la qualité de l'estimation. Le terme associé à l'hypothèse d'un modèle de Poisson entre les données et le trafic total, est celui ayant le plus d'impact sur la performance de l'estimée finale. Cela justifie une fois de plus l'importance actuelle et à venir du rôle des jeux de trajectoires en Ingénierie du Transport. Ces trajectoires sont importantes mais nous avons montré que ce n'est tout de même que lorsque l'ensemble des contraintes développées dans cette thèse sont minimisées conjointement, que la performance de l'estimée est maximale. En particulier, les comptages de véhicules jouent toujours un rôle important dans ce processus.

Lors de cette thèse, nous avons proposé plusieurs formulations successives du problème de l'estimation des matrices ODL. Cela a en particulier mis en valeur l'importance des fonctions de régularisation sur la performance des résultats. Il apparait alors clair que d'autres fonctions pourraient s'avérer adaptées et plus efficaces que celles proposées ici. Débusquer de telles fonction serait une bonne suite à ce travail. Par exemple, la répartition des véhicules par mouvement possible aux intersections pourrait aider à mieux contraindre l'estimée et à améliorer les performances. De même, d'autres algorithmes pourraient répondre à ce même problème avec diverses améliorations : augmenter la vitesse de convergence ou permettre une estimation en temps réel seraient deux améliorations d'intérêt pour les ingénieurs du transport. Enfin la question de l'estimation de matrices ODL dynamiques (qui dépendent du temps) n'a pas été particulièrement développée ici. Nous avons fait varier le nombre d'utilisateurs considérés dans les scénarii simulés, ce qui revient à changer le pas de temps sur lequel les données sont récoltées, et nous avons montré une certaine résilience des résultats face à de telles variations. Néanmoins la question de la cohérence des estimations successives ou de l'estimation de l'évolution des matrices ODL n'a pas été abordée et serait surement d'un grand intérêt. Une piste relativement simple pourrait être l'addition d'un terme dans la fonction objective favorisant la continuité entre deux estimations temporellement successives. Il est certain que l'accès à des matrices ODL dynamiques aurait beaucoup d'applications pour les prédictions de trafic.

Enfin, une dernière contribution de ce travail a été d'illustrer l'intérêt d'estimer une matrice ODL en appliquant l'ensemble de l'expertise au cas de Brisbane. L'estimation des matrices ODL jusque-là n'avait été effectuée que sur des scénarii simulés et la confrontation avec des données réelles a été une partie importante de cette recherche. Il a d'une part fallu modifier la fonction objective afin de prendre

en compte les caractéristiques propres aux jeux de données réels donc complexes et d'autre part, la réalité à laquelle comparer les estimations n'étant pas disponible, seule une étude de cohérence des résultats a été possible. Cela a permis de démontrer l'avantage des matrices ODL sur les matrices OD : elles apportent une information supplémentaire et directe sur l'usage du réseau routier et sur les déplacements. C'est de plus une preuve que les méthodes proposées ici peuvent sortir du cadre théorique et être appliquées directement à de vrais réseaux.

Ce manuscrit de thèse ne saurait se terminer sans apporter quelques réflexions sur une question très fréquemment soulevée lors de la présentation de mon travail : la question éthique de détecter des appareils Bluetooth personnels. Une réponse facile, qui permet souvent d'éviter le débat, est basée sur les arguments suivants : d'une part, la mairie de Brisbane n'a jamais caché l'installation des détecteurs Bluetooth, d'autre part, l'association entre les adresses MAC et le nom des utilisateurs s'avérerait très difficile : Premièrement, les adresses sont encryptées et la table d'encryptage n'est pas accessible. Deuxièmement, l'adresse MAC n'est généralement pas associée au nom des clients par les vendeurs d'appareils Bluetooth et dans tous les cas ; il serait peu probable que les vendeurs acceptent de partager cette information. Enfin, au vu du nombre de fabricants et vendeurs Bluetooth, obtenir cette information pour les millions d'adresses détectées serait un travail titanesque. Bref, CQFD, les questions de vie privée et d'anonymat ne sont pas un problème ici.

En réalité, ce n'est pas si simple. Nous avons montré dans la partie 3.2.5 du Chapitre 3 que 5 fabricants représentent à eux seuls 50 % des appareils détectés. De plus, des études ont déjà démontré que même sans possibilité d'identification, les métadonnées finissent toujours par apporter suffisamment d'informations pour que l'identification à posteriori soit possible. de Montjoye, Hidalgo, Verleysen, and Blondel (2013) [212] ont démontré que dans une base de données avec les détections des téléphones mobiles d'un demi-million d'utilisateurs pendant 15 mois, il suffisait de connaitre la position d'un utilisateur à 4 moments différents pour pouvoir l'identifier avec une certitude de 95%. Pour comparaison, cela va bientôt faire 10 ans que les données Bluetooth sont collectées à Brisbane pour plusieurs millions d'appareils. Ainsi, affirmer que la vie privée des utilisateurs est et sera toujours respectée ne serait pas une réponse parfaitement honnête. Et c'est ici que se situe en fait le point clef de ce débat : l'unique façon d'assurer que les données ne remettent pas en cause la vie privée et l'anonymat des utilisateurs est d'être en mesure d'imposer que les données ne servent qu'à des études aux niveaux agrégés.

C'est pour cela qu'il me semble important que les données collectées, le soit par un pouvoir public plutôt que par des entreprises privées. Et sur ce point, la mairie de Brisbane a toujours été claire : les données ne seront pas accessibles à tout le monde et si le STRC en a obtenu l'accès, c'est contre la promesse d'études du trafic à un niveau agrégé : temps de trajets, vitesses, matrices OD et ODL, mouvements aux intersections.

Finalement, toutes les applications permises par ces nouvelles données sont au bénéfice de la communauté de Brisbane (affichage dynamique des temps de parcours, optimisations des plans de feu, ...). Ceci est à l'opposé des tendances actuelles prises par les entreprises qui collectent des quantités de données importantes sur leurs clients (par exemple par la localisation automatique des téléphones mobiles), données qui seront d'abord monétisées, par exemple par la publicité ciblée, bien avant d'avoir des retombées plus directe pour la communauté. Cela amène donc à dernier point de débat, à savoir si l'on préfère que les données soit gérées par les pouvoirs publics à qui elles vont couter de l'argent pour la collecte, le stockage et l'analyse, mais contre une plus grande visibilité de leur utilisation réelle, ou si les données ne sont finalement vouée qu'à devenir une monnaie d'échange pour des services fournis par les entreprises.

Dans tous les cas, ces questions éthiques soulignent l'importance de soutenir un effort public et académique pour garder la maitrise et l'expertise dans la collecte et l'analyse des données. Les questions éthiques ne pouvant être résolue par des solutions techniques, les outils et l'expertise de collecte, traitement, et analyse des données devraient rester publics et transparents.

Résumé

L' estimation des matrices origine-destination (OD) est un sujet de recherche important depuis les années 1950. En effet, ces tableaux à deux entrées recensent la demande de transport d'une zone géographique donnée et sont de ce fait un élément clé de l'ingénierie du trafic. Historiquement, les seules données disponibles pour leur estimation par les statistiques étaient les comptages de véhicules par les boucles magnétiques. Ce travail s'inscrit alors dans le contexte de l'installation à Brisbane de plus de 600 détecteurs Bluetooth qui ont la capacité de détecter et d'identifier les appareils électroniques équipés de cette technologie.

Dans un premier temps, il explore la possibilité offerte par ces détecteurs pour les applications en ingénierie du transport en caractérisant ces données et leurs bruits. Ce projet aboutit, à l'issue de cette étude, à une méthode de reconstruction des trajectoires des véhicules équipés du Bluetooth à partir de ces seules données. Dans un second temps, en partant de l'hypothèse que l'accès à des échantillons importants de trajectoires va se démocratiser, cette thèse propose d'étendre la notion de matrice OD à celle de matrice OD par lien afin de combiner la description de la demande avec celle de l'utilisation du réseau. Reposant sur les derniers outils méthodologies développés en optimisation convexe, nous proposons une méthode d'estimation de ces matrices à partir des trajectoires inférées par Bluetooth et des comptages routiers.

À partir de peu d'hypothèses, il est possible d'inférer ces nouvelles matrices pour l'ensemble des utilisateurs d'un réseau routier (indépendamment de leur équipement en nouvelles technologies). Ce travail se distingue ainsi des méthodes traditionnelles d'estimation qui reposaient sur des étapes successives et indépendantes d'inférence et de modélisation.

Mots Clés: matrice origines-destinations, matrice origines-destinations par lien, optimisation convexe non lisse, algorithme Proximal Primal-Dual, comptages trafic, bluetooth, identifiaction automatique des véhicules, trajectoires.

Abstract

ORIGIN Destination matrix estimation is a critical problem of the Transportation field since the fifties. OD matrix is a two-entry table taking census of the zone-to-zone traffic of a geographic area. This traffic description tools is therefore paramount for traffic engineering applications. Traditionally, the OD matrix estimation has solely been based on traffic counts collected by networks of magnetic loops. This thesis takes place in a context with over 600 Bluetooth detectors installed in the City of Brisbane. These detectors permit in-car Bluetooth device detection and thus vehicle identification.

This manuscript explores first, the potentialities of Bluetooth detectors for Transport Engineering applications by characterising the data, their noises and biases. This leads to propose a new methodology for Bluetooth equipped vehicle trajectory reconstruction. In a second step, based on the idea that probe trajectories will become more and more available by means of new technologies, this thesis proposes to extend the concept of OD matrix to the one of link dependent origin destination matrix that describes simultaneously both the traffic demand and the usage of the network. The problem of LOD matrix estimation is formulated as a minimisation problem based on probe trajectories and traffic counts and is then solved thanks to the latest advances in nonsmooth convex optimisation.

This thesis demonstrates that, with few hypothesis, it is possible to retrieve the LOD matrix for the whole set of users in a road network. It is thus different from traditional OD matrix estimation approaches that relied on successive steps of modelling and of statistical inferences.

Keywords: Origin-Destination Matrix, Nonsmooth Convex Optimisation, Proximal Primal-Dual Algorithm, Traffic Flows, Bluetooth, Automated Vehicle identification, Trajectories.