

Contents

Acknowledgments	1
1 Introduction	3
1.1 Motivation	4
1.2 Data-driven transport modelling	5
1.2.1 Application to transit	6
1.3 Research methodology	7
1.4 Thesis structure	8
2 System design	10
2.1 Overview	10
2.1.1 Assumptions driving modelling focus	11
2.2 MATSim	12
2.3 Data elements	13
2.3.1 The Contactless EPurse Application System (CEPAS)	14
2.3.2 Road network description	16
2.3.3 Transit schedule	16
2.3.4 Vehicle fleet description	17
2.4 Process elements	18
2.4.1 Data synthesis and demand generation	18
2.4.2 Dwell time modelling	19

2.4.3	Network modification	19
2.4.4	Trajectory reconstruction	21
2.4.5	Speed regression modelling and modified transit simulation	22
2.4.6	Distributed re-planning and pseudo-simulation	22
2.4.7	Post-processing and analysis	23
3	Trajectory reconstruction	25
3.1	Introduction	25
3.2	Challenges	26
3.3	Method	28
3.4	Applications	32
3.4.1	Average waiting times by time of day	32
3.4.2	Waiting times related to vehicle trajectories for a given bus service . . .	35
3.5	Conclusion	35
4	Localised prediction of bus speeds from reconstructed trajectories	36
4.1	Overview	36
4.2	Deriving topological and dynamic variables	37
4.2.1	Variable listing	38
4.3	Data preparation	40
4.3.1	Multivariate outlier detection	40
4.3.2	Centring and scaling	41
4.4	Correlation and multicollinearity	43
4.4.1	Trading off Akaike Information Criterion (AIC) vs Variance Inflation Factor (VIF)	45
4.5	Comparison of instantaneous vs. hourly, speed vs. travel time prediction	48
4.6	Interactions	51

4.7	Random forest regression	52
4.7.1	Training	53
4.7.2	Results	53
4.8	Modelling stochasticity	55
4.8.1	OLS regression results	55
4.8.2	RF regression results	55
4.8.3	Comparison of simulated results versus actual	57
4.9	Temporal autocorrelation	60
4.10	Spatial autocorrelation	61
4.10.1	Background	61
4.10.2	Results	67
4.11	Conclusion	70
5	Transit simulation	72
5.1	Key elements of the data-driven transit simulation	72
5.1.1	Demand generation	72
5.1.2	Link dynamics	73
5.1.3	Dwell time model	75
5.2	Validation and performance	75
5.2.1	Headways, dwell times and bus bunching	75
5.2.2	Passenger travel time measures	79
5.2.3	Computation times of simplified simulation	81
5.3	Application	82
5.3.1	Impact on bus bunching	83
5.3.2	Excess waiting times	83
5.4	Conclusion and future work	85

6	Improving simulation performance	86
6.1	Development context	86
6.1.1	Transit	87
6.2	Motivation	88
6.2.1	Harnessing parallelisation	88
6.2.2	The expanding scope of MATSim applications	89
6.3	Background	90
6.3.1	Mutation approaches	90
6.3.2	Best-response vs. random-response replanning	91
6.3.3	Simulation-based optimisation using surrogate models	91
6.3.4	Feedback and learning	92
6.4	Design	93
6.4.1	MATSim events	93
6.4.2	Pseudo-Simulation (PSim) operation	94
6.5	Experimental setup	95
6.5.1	Simulation scenario	96
6.6	Results	96
6.6.1	Characterizing solution state	96
6.6.2	Varying QSim:PSim ratio	98
6.6.3	Performance test	98
6.6.4	Solution state	99
6.7	Conclusion	103
6.7.1	Performance	103
6.7.2	Solution state	104
6.7.3	Application to data-driven simulation	104

7	A simple framework for distributed simulation	105
7.1	Motivation and background	106
7.1.1	Transit simulation sample size	106
7.1.2	Agent memory and diversification	107
7.2	Design	107
7.2.1	Master/slave configuration	108
7.2.2	Serialisation	108
7.2.3	Serial vs. parallel operation	111
7.2.4	Distributed simulation as a replanning strategy	111
7.3	Method	112
7.4	Scenario	113
7.4.1	Testing overhead impact of increasing number of nodes	113
7.4.2	Experimental design: finding optimal parameters	114
7.5	Results	118
7.6	Conclusion and outlook	121
7.6.1	Next step: ensemble simulations	122
8	Surrogate data synthesis	123
8.1	Motivation	123
8.1.1	Aggregates	124
8.2	Overview	124
8.3	Context	127
8.3.1	Quantifying privacy	127
8.3.2	Privacy protection specialization	128
8.3.3	The special case of trajectories	128
8.3.4	Continuous trajectory privacy protection	129

8.3.5	Published trajectory privacy protection	129
8.3.6	Generalisation-based anonymisation	130
8.3.7	Surrogate data synthesis in agent-based transport demand modelling . .	132
8.3.8	Histogram matching	133
8.3.9	Principal Component Analysis (PCA)	134
8.4	Design idea: iterative multiple histogram matching	134
8.4.1	Two-dimensional histogram matching	136
8.4.2	Using PCA to improve restoration of joint distribution	137
8.5	Method	137
8.5.1	Encoding	137
8.5.2	Synthesis	138
8.6	Results	140
8.7	Conclusion and Outlook	144
8.7.1	‘Painting’ scenarios	144
8.7.2	Higher-dimensional reconstruction, performance	144
8.7.3	Alternative encoding method: grand tour sinograms	145
9	Discussion and outlook	147
	Curriculum Vitae	159