POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	A Methodological Framework to Develop Optimal Sets of Driving Cycles for a Region
Auteur: Author:	Asad Yarahmadi
Date:	2023
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Yarahmadi, A. (2023). A Methodological Framework to Develop Optimal Sets of Driving Cycles for a Region [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/57092/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/57092/
Directeurs de recherche: Advisors:	Catherine Morency, & Martin Trépanier
Programme: Program:	Génie civil

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

A Methodological Framework to Develop Optimal Sets of Driving Cycles for a Region

ASAD YARAHMADI

Département des génies civil, géologique et des mines

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie civil

Décembre 2023

© Asad Yarahmadi, 2023.

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

A Methodological Framework to Develop "Optimal" Sets of Driving Cycles for a Region

présentée par **Asad YARAHMADI**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment accepté par le jury d'examen constitué de :

SAUNIER, Nicolas, président

MORENCY, Catherine, membre et directeur de recherche

TREPANIER, Martin, membre et codirecteur de recherche

CIARI, **Francesco**, membre

MAOH, Hanna, membre externe

DEDICATION

To Ponder for a Moment is Greater Than Sixty Years of Worship.

(Mohammad)

To:

Imam Hossein, Imam Reza, and Agham Abu Fazel

Beloved wife and My Family

Those who have enhanced the world through science

ACKNOWLEDGEMENTS

I would like to thank Prof. Catherine Morency, my research director, and Prof. Martin Trepanier, my co-director, who gave me this opportunity and who guided, advised, and developed my critical thinking in such a politicized field. Through their teaching and passion, they helped me acclimatize and become aware of the challenges of effective transport planning.

My heartfelt appreciation goes to my core family in Iran, including my father, mother, beloved wife, brothers, and sister, as well as to my extended family worldwide. Their support was pivotal to this project's success. Additionally, I'm deeply thankful to my friends and faith community for their constant encouragement, uplifting conversations, and the much-needed leisurely moments together.

Special acknowledgment is due to Hubert, Jean-Simon, Brigitte, and Frederique Roy for their commendable assistance. My experience at Polytechnique Montréal was enriched by the camaraderie and shared insights with my master's and doctoral peers. Those cherished moments in the office remain a bright spot in my academic journey.

RÉSUMÉ

La réduction des émissions de GES provenant des transports constitue un défi de taille en raison de la demande croissante de mobilité et de la dépendance continue aux combustibles fossiles. Même si de nombreux efforts ont été déployés pour rendre les transports plus écologiques, les émissions de GES affichent une tendance à la hausse. Selon le Conseil international pour les transports propres (ICCT), l'un des principaux problèmes consiste à assurer que les émissions des véhicules soient correctement mesurées afin que les règles de contrôle de ces gaz soient respectées. Nous utilisons ce qu'on appelle le cycle de conduite pour aider à modéliser ces émissions.

Il existe principalement deux types de cycles de conduite : SDC (Standard Driving Cycle) et LDC (Local Driving Cycle). Les SDC examinent les habitudes de conduite à l'échelle nationale, voire internationale, pour évaluer les émissions d'un véhicule. En revanche, les PMA se concentrent sur les habitudes de conduite dans des zones locales spécifiques, telles que les villes ou des routes particulières. En général, les PMA peuvent représenter des comportements de conduite plus réalistes. Cependant, de nombreuses études antérieures se sont appuyées sur un ou quelques cycles de conduite seulement pour représenter les comportements de conduite, ce qui peut conduire à des imprécisions dans l'estimation des émissions.

Il est important de comprendre comment les gens conduisent sur tous les types de routes et dans toutes les conditions météorologiques. Des études récentes montrent que des éléments tels que la pluie, la température, la pente de la route et les limites de vitesse peuvent modifier la façon dont les gens conduisent. Ainsi, les habitudes de conduite peuvent changer d'une ville à l'autre et même en fonction de la météo.

Il peut y avoir de nombreux modes de conduite différents dans une ville. Si nous n'utilisons qu'un seul modèle de conduite, nous perdons de nombreux détails sur la façon dont les gens conduisent réellement. Cela signifie que nous n'avons peut-être pas une idée précise de la quantité de gaz nocifs produits par les véhicules. Lorsque nous parlons de « perte d'informations », nous voulons dire que nous passons à côté de nombreux détails sur les habitudes de conduite.

À l'inverse, générer des cycles de conduite individuels pour tous les déplacements sur différents segments de route et conditions météorologiques nécessiterait la construction d'un nombre illimité et indéfini de cycles de conduite. En d'autres termes, le cycle de conduite est dynamique et comme les caractéristiques des déplacements changent en fonction des différentes conditions

météorologiques et des différents segments de route, il existe un nombre indéfini de cycles de conduite dans une zone urbaine. La question est donc d'identifier un nombre fini de cycles de conduite capables de représenter correctement les comportements de conduite (le profil de vitesse, d'accélération, de décélération étant utilisé comme proxy de ces comportements) dans une zone urbaine. Pour répondre à cette question, nous proposons un cadre méthodologique utilisant Montréal comme étude de cas. Le cadre est conçu pour identifier l'ensemble « optimal » de cycles de conduite. Pour ce faire, un compromis est fait entre minimiser le nombre de cycles de conduite, minimiser la perte d'informations et maximiser la variabilité des cycles de conduite générés..

L'examen de la littérature existante révèle que les conditions météorologiques et les caractéristiques de la route sont les deux principaux facteurs externes qui ont un impact sur le comportement de conduite et entraînent sa variabilité. Le principal défi dans la compréhension de la relation entre ces facteurs externes et les caractéristiques du cycle de conduite provient de leur nature volatile. Plus précisément, les conditions météorologiques peuvent changer au cours de la journée et selon les saisons, tandis que les caractéristiques de la route peuvent varier d'un segment à l'autre. Un élément crucial de notre stratégie est son adaptabilité : la méthodologie doit être suffisamment polyvalente pour être applicable dans diverses zones urbaines, indépendamment de la taille et de la complexité de la structure du réseau routier d'une ville ou des conditions météorologiques dominantes.

Pour tenir compte de la variabilité des caractéristiques de la route, un modèle de régression a été initialement développé pour déterminer les facteurs ayant un impact sur les comportements de conduite. L'utilisation ultérieure de techniques d'apprentissage automatique non supervisées, telles que les k-moyennes, l'AHC et le k-proto, ont généré des typologies de segments de route, Les résultats de trois techniques de validation de cluster ont confirmé qu'une combinaison de k-moyennes et de distance de Gower apparaissant comme la plus précise. Les validations utilisant les données de vitesse des taxis montréalais ont confirmé les capacités de différenciation de ces typologies.

Ensuite, grâce à une modélisation de régression multiple multivariée, les principaux facteurs météorologiques affectant les modèles de cycle de conduite ont été déterminés. Grâce à l'apprentissage automatique non supervisé, les conditions météorologiques ont été regroupées en fonction de leur impact sur les habitudes de conduite, ce qui a abouti à la formulation de deux

cycles de conduite distincts pour chaque type de temps : un pour les jours de semaine et l'autre pour le week-end.

S'appuyant sur les études initiales, un cadre méthodologique a été conçu, en se concentrant sur Montréal comme étude de cas, pour déterminer l'ensemble « optimal » de cycles de conduite qui représentent de manière exhaustive toutes les conditions routières et météorologiques.

En employant des typologies et des techniques d'apprentissage automatique préalablement établies, un modèle mixte multiniveau a été formulé. Ce modèle a minutieusement évalué l'influence combinée des types de route et des conditions météorologiques sur les paramètres du cycle de conduite. À partir de la série initiale de 80 cycles de conduite, une approche en deux étapes a été développée. La première étape a utilisé une matrice de similarité pour évaluer la ressemblance entre les cycles de conduite, et la deuxième étape a utilisé deux paramètres directeurs : la perte d'informations et la variabilité. La recherche conclut qu'une représentation exhaustive des conditions de conduite à Montréal nécessite 76 cycles de conduite distincts.

ABSTRACT

There is a significant challenge in reducing GHG emissions from transportation due to the rising demand for mobility and the continued reliance on fossil fuels. Even though many efforts have been made to make transportation greener, GHG emissions shows an upward trend. According to the International Council on Clean Transportation (ICCT), one big problem is making sure we measure vehicle emissions correctly so that everyone in transportation follows the rules for controlling these gases. We use something called the driving cycle to help model these emissions.

There are primarily two types of driving cycles: SDC (Standard Driving Cycle) and LDC (Local Driving Cycle). SDCs examine driving habits on a national or even international scale to evaluate a vehicle's emissions. In contrast, LDCs focus on driving habits in specific local areas, such as cities or particular roads. Generally, LDCs can depict more realistic driving behaviors. However, many past studies have relied on just one or few driving cycles to represent driving behaviors, which can lead to inaccuracies in estimating emissions.

It is important to understand how people drive on all kinds of roads and in all weather conditions. Recent studies show that things like rain, temperature, the slope of the road, and speed limits can change the way people drive. So, driving habits can change across a city and even with the weather.

There can be many different driving patterns in a city. If we just use one driving pattern, we lose a lot of details about how people really drive. This means we might not get a clear idea of how many harmful gases are produced by vehicles. When we talk about "information loss", we mean that we miss out on many details about driving patterns.

Conversely, generating individual driving cycles for all trips in different road segments and weather conditions would require constructing an unlimited and indefinite number of driving cycles. In other words, the driving cycle is dynamic and since travel characteristics change depending on different weather conditions and different road segments, there are an indefinite number of driving cycles in an urban area. The question is therefore to identify a finite number of driving cycles capable of correctly representing driving behaviors (the speed, acceleration, deceleration profile being used as a proxy for these behaviors) in an urban area. To answer this question, we propose a methodological framework using Montreal as a case study. The framework is designed to identify the optimal set of driving cycles. To do that, a trade off is made among minimizing number of

driving cycles, minimizing information loss and maximizing variability of generated driving cycles.

Reviewing the existing literature reveals that weather conditions and road features are the two primary external factors that impact driving behaviors and lead to variability in it. The main challenge in understanding the relationship between these external factors and driving cycle characteristics stems from their volatile nature. Specifically, weather conditions can change throughout the day and across different seasons, while road features can vary from one segment to another. A crucial component of our proposed strategy is its intended adaptability – the methodology should be sufficiently versatile to be applicable across diverse urban areas, irrespective of the size and complexity of a city's road network structure, or its prevailing weather conditions.

To address variability of road features, initially, a regression model was developed to determine factors impacting driving behaviors. Subsequent utilization of unsupervised machine learning techniques, such as k-means, AHC, and k-prototype, generated typologies of road segments, with a combination of k-means and Gower distance emerging as the most accurate. Validations using speed data from Montreal taxis confirmed the differentiation capabilities of these typologies.

Next, through multivariate multiple regression modeling, key weather factors affecting driving cycle patterns were determined. Using unsupervised machine learning, weather conditions were grouped according to their impact on driving patterns, resulting in the formulation of two distinct driving cycles for every weather type – one for weekdays and the other for weekends.

Building upon the initial studies, a methodological framework was conceived to determine the "optimal" set of driving cycles that comprehensively represents all road and weather conditions.

By employing previously established typologies and machine-learning techniques, a multilevel mixed model was formulated. This model meticulously assessed the combined influence of road and weather types on driving cycle parameters. Out of the initial set of 80 driving cycles, a two-step approach was developed. The first step used a similarity matrix to gauge the resemblance among the driving cycles, and the second step utilized two guiding parameters: information loss and variability. The research concludes that an exhaustive representation of Montreal's driving conditions necessitates 76 distinct driving cycles.

TABLE OF CONTENTS

DEDICATION III

ACKNOWLEDGEMENTS IV

RÉSUMÉ V

ABSTRACT VIII

TABLE OF CONTENTS X

LIST OF TABLES XVI

LISTE OF SYMBOLS AND ABBREVIATIONS XX

CHAPTER 1 INTRODUCTION 1

1.1 Context 1

1.2 Problem Statement 3

1.3 Research Objectives 5

1.4 Main Research contributions 6

1.5 Research Questions 7

1.6 Thesis Contents 7

CHAPTER 2 LITERATURE REVIEW 9

2.1 Driving cycles 10

2.1.1 Route Selection 13

2.1.2 Driving Cycle Data Collection 16

2.1.3 GPS Data Processing 18

2.1.4 Microtrips Generation 20

2.1.5 Microtrips Classification 25

2.1.6 Assessment Measures 25

2.1.7 Candidate Driving Cycle Construction 29

2.1.8 Selecting representative driving cycle(s) 33

2.1.9 Summary of Driving cycle 33

2.2 Influencing Factors 37

2.2.1 Road features 38

2.2.2 Weather Parameters 42

2.2.3 Type of day (weekend and weekday) 48

2.2.4 Summary of Influencing Factors 49

2.3 Classification of influencing factors 53

2.3.1 Classification of Road Types 53

2.3.2 Classification of weather parameters 56

2.4 Clustering 56

2.4.1 Partitioning method 56

2.4.2 Hierarchical Clustering 57

2.4.3 Cluster Validation 58

2.5 Mixed Models 59

2.6 Main Gaps 60

CHAPTER 3 RESEARCH METHODOLOGY 62

3.1 Data 65

3.2 Data Preprocessing 65

3.3 Preparation of Model Inputs 66

3.3.1 Driving cycle construction 67

3.3.2 Road network clustering 68

3.3.3 Weather parameters clustering 68

3.4 Developing a Model to Determine “Optimal” Sets of Driving Cycle 70

3.4.1 Validation of developed methodological framework 71

3.4.2 Determining "optimal" number of driving cycles 74

3.5 Publications of this thesis 76

CHAPTER 4 ARTICLE 1: NEW DATA-DRIVEN APPROACH TO GENERATE TYPOLOGIES OF ROAD SEGMENTS 78

4.1 Introduction 79

4.2 Background 81

4.2.1 Factors Affecting Driving Patterns 81

4.2.2 Road Classification Methodology 83

4.3 Methodology 85

4.3.1 Phase 1. Data preprocessing 86

4.3.2 Phase 2. Cluster analysis of mixed-type data 93

4.3.3 Phase 3. Multivariate regression model 94

4.4 Results 95

4.4.1 Generating regression model for the whole network 95

4.4.2 Generating the road segment typology 96

4.4.3 Validating Clustering Results 101

4.4.4 Statistical analysis of the speed distribution of the first approach 103

4.4.5 Investigating the association between road features and driving speed 106

4.5 Conclusion 109

CHAPTER 5 ARTICLE 2: IDENTIFYING OPTIMAL SETS OF DRIVING CYCLES IN DIFFERENT DRIVING CONDITIONS USING MACHINE LEARNING 111

5.1 Introduction 112

5.2 Related Work 114

5.2.1 Driving Cycle Generation and Weather Conditions 115

5.2.2 Combined Effects of Weather Parameters and Driving behaviors Study 117

5.3 Research Methodology 119

5.3.1 Study Area and Data Description 122

5.3.2 Driving Data 123

5.3.3 Weather Data 124

5.3.4 Data Preprocessing 125

5.3.5 Driving Cycle Construction 131

5.4 Results and Discussion 133

5.4.1 Multivariate multiple regression analysis before generating weather types 133

5.4.2 Generating Weather Types 134

5.4.3 Multivariate multiple regression model after generating weather types 141

5.4.4 Driving Cycle Characteristics Analysis 143

5.4.5 Comparative Analysis of Newly Constructed Driving Cycles Versus Established Models 148

5.4.6 Determining the Representative Sets of Driving Cycles 149

5.5 Conclusion 152

CHAPTER 6 ARTICLE 3: A METHODOLOGICAL FRAMEWORK TO DETERMINE THE OPTIMAL SET OF DRIVING CYCLES FOR A REGION 154

6.1 Introduction 155

6.2 Literature review 157

6.2.1 Route selection and road features 158

6.2.2 Collecting driving data and driving conditions 159

6.3 Methodology 162

6.3.1 Driving data description 162

6.3.2 New typology of road network 163

6.3.3 Weather typology 164

6.3.4 Driving cycle generation 165

6.3.5 Mixed models 166

6.3.6 “Optimal” number of driving cycles 168

6.4 Result 171

6.4.1 Mixed models to study the influences of fixed effects on driving cycle parameters 171

6.4.2 Determining aggregation level for the developed methodology 178

6.4.3 Determining the “optimal” number of driving cycles 180

6.5 Conclusion 183

CHAPTER 7 GENERAL DISCUSSION 185

7.1 Recall of the general objective of the thesis 185

7.2 Discussions of results 185

7.2.1 Developing a new approach to categorize road network segments 186

7.2.2 Generating a weather typology 187

7.2.3 Determining “optimal” number of driving cycles 187

7.2.4 General discussion on the contributions 188

7.3 Summary of limitations 189

7.3.1 Data 189

7.3.2 Geographical Specificity 190

7.3.3 Methodology 190

CHAPTER 8 CONCLUSION 192

8.1 Summary of the thesis 192

8.2 Future research 192

8.2.1 Climate change and driving cycle parameters 192

8.2.2 Considering the geographical specificity of the study 193

8.2.3 Adding new features and running the framework 194

8.2.4 Exploring the impact of COVID-19 on driving behavior 195

REFERENCES 196

APPENDIX 216

LIST OF TABLES

Table 2-1 Characteristic parameters in constructed driving cycles 27

Table 2-2 Traffic-related parameters for cycle assessment 28

Table 2-3. Comparison of local driving cycles 35

Table 2-4 Factors influencing driving behavior 37

Table 2-5 Comparison Influences of weather parameters on vehicle speed 46

Table 2-6 List of constructed driving cycles considering influencing features 51

Table 2-7 Comparative analysis of criteria utilized to classify road system in different countries55

Table 3-1 Required data for this project 66

Table 3-2 Assessment measures (Nouri, 2015) 67

Table 3-3 Brief description of validation index 69

Table 4-1. Description of data used in this study 86

Table 4-2 Details of the base model 95

Table 4-3 Sample of road network data 97

Table 4-4 A sample of the Gower similarity matrix 97

Table 4-5 Internal cluster validation index 101

Table 4-6 Contingency matrix between results from approaches 1 and 3 103

Table 4-7 The result of the ANOVA test 104

Table 4-8 Post hoc result analysis 104

Table 4-9 Road features and OLS regression model outputs for developed road types 107

Table 5-1 SDCs and weather parameters 115

Table 5-2 Characteristics of the selected road type 122

Table 5-3. Driving data features 124

Table 5-4 Driving cycle assessment criteria 132

Table 5-5 Number of observations in each cluster 136

Table 5-6 Number of observations in each cluster in AP method 138

Table 5-7 Number of observations in each cluster in AHC method 138

Table 5-8 Comparison R-square between clustered weather types and base 141

Table 5-9. MMRM models output for the developed weather types 142

Table 5-10 Comparison between SAFD of weekdays and weekend 145

Table 5-11 differences between the developed driving cycles local and standard driving cycles 150

Table 6-1Comparison of conducted studies and proposed approach 161

Table 6-2 Driving data features 162

Table 6-3 Driving cycle assessment criteria 166

Table 6-4 The most important exogenous factors for each driving cycle parameters 177

LIST OF FIGURES

Figure 1-1 Green House Gas emissions by source sector, EU-27, 1990 and 2018 in Quebec Source: Reproduced from (Canada 2021) 1

Figure 1-2 SDS in the transport sector by mode, 2000 – 2030 Source: (IEA 2019)..... 2

Figure 2-1 Driving cycle profile-image by author, 2023 10

Figure 2-2 Changes of fuel consumption in various road types..... 54

Figure 2-3 Hierarchical clustering structure, Image by author, 2020 58

Figure 3-1 Research methodology of this thesis- Image by author, 2020..... 64

Figure 3-2 Hierarchical data structure..... 72

Figure 4-1 Phases to perform the data-driven framework- Image by author, 2022 85

Figure 4-2 Correlation analysis between road features 91

Figure 4-3 The number of components for the PCA..... 93

Figure 4-4 The optimal number of clusters; Elbow method 98

Figure 4-5 Dendrogram to determine the optimal number of clusters 99

Figure 4-6 Elbow method; k-prototype 100

Figure 4-7 Comparison of speed distribution of the three typologies of segments..... 102

Figure 4-8 Montreal Island network typology using approach 1 106

Figure 5-1 Research methodology of the present study 121

Figure 5-2 Variation of temperature, humidity, precipitation, wind speed over the years (2019-2020)..... 125

Figure 5-3 Determination of cluster numbers using elbow (a), gap statistics (b), and Dendrogram 135

Figure 5-4 Comparing clustering results of approach 1, approach 2, approach 3..... 141

Figure 5-5 Share of modes for all weather types on weekdays (a) and weekends (b) 148

Figure 5-6 Similarity matrix of the generated driving cycles in different weather conditions ... 151

Figure 6-1 Driving cycle number spectrum 156

Figure 6-2 Conceptual model of the research methodology 163

Figure 6-3 Hierarchical structure of data of the present study 168

Figure 6-4 Changes of driving cycle parameters in different road types, weather types in weekend and weekdays 176

Figure 6-5 Comparison of information loss in different aggregation level for information loss indicators and overall information loss 180

Figure 6-6 Similarity matrix of the generated driving cycles 181

Figure 6-7 Share of similarity level in the similarity matrix 182

Figure 6-8 Similarity level, driving cycle number vs. information loss and variability 183

LISTE OF SYMBOLS AND ABBREVIATIONS

AADT	Annual Average Daily Traffic
AHC	Agglomerative Hierarchical Clustering
ANOVA	Analysis of variance
AVL	Advanced Vehicle Location
CUEDC	Composite Urban Emissions Drive Cycle
CVS	Constant Volume Sampling
DEM	Digital Elevation Model
DHC	Divisive Hierarchical Clustering
GHG	Green House Gas
HCM	Highway Capacity Manual
IEA	International Energy Agency
LCSS	Longest Common Subsequence
LDC	Local Driving Cycle
LOS	Level Of Service
MLE	Maximum Likelihood Estimation
MOVES	MOtor Vehicle *Emission* Simulator
OBD	On-Board Diagnostics
PCA	Principal Component Analysis
PV	Performance values
SAFD	Speed Acceleration Frequency Distance
SDC	Standard Driving Cycle
SVD	Singular Value Decomposition
SVM	Support Vector Machine

CHAPTER 1 INTRODUCTION

1.1 Context

Greenhouse gases (GHGs), comprising carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O), are pivotal contributors to significant environmental and health challenges encountered by contemporary societies. Extensive research evidences the linkage between these emissions and phenomena such as climate change, atmospheric contamination, and heightened risks of cardiovascular ailments (Tayarani et al. 2018). Over recent decades, the transportation sector has emerged as a primary source of GHG emissions (Lin et al. 2011). Figure 1-1 delineates the GHG contributions from various sectors, encompassing waste management, agriculture, industrial operations, and fuel combustion. Notably, from 1990 to 2018 in Quebec, Canada, the overall GHG emissions remained stable. However, within this timeframe, the transportation sector's contribution to these emissions rose from 21% to 25% (Canada 2021).

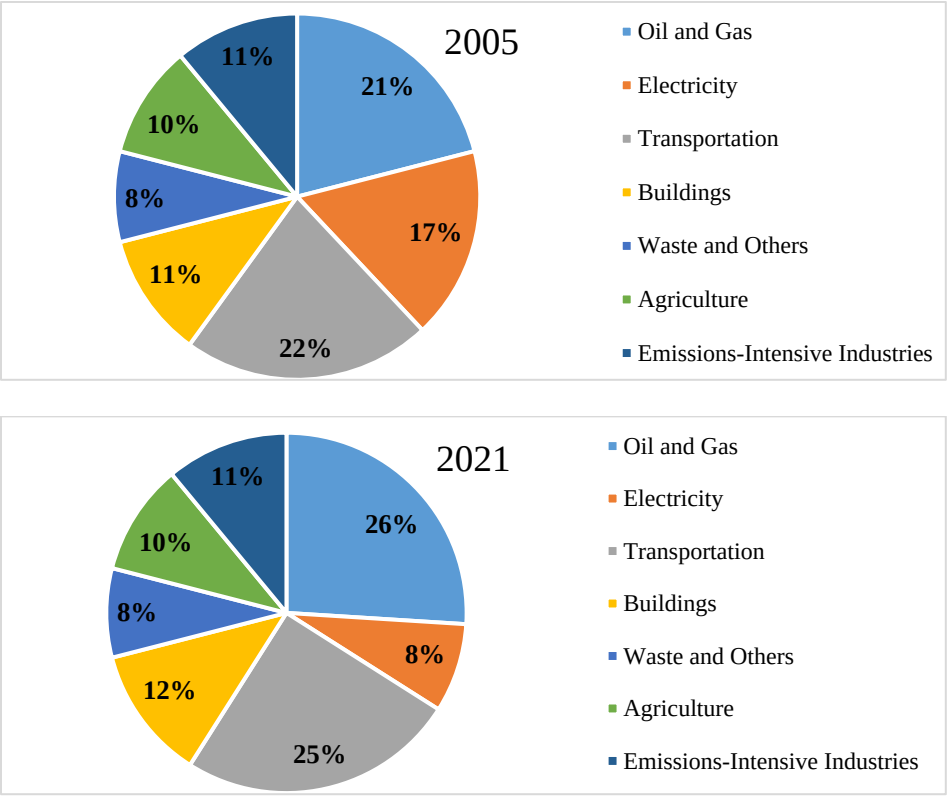


Figure 1-1 Green House Gas emissions by source sector, EU-27, 1990 and 2018 in Quebec
Source: Reproduced from (Canada 2021)

The 2019 Tracking Transport report by the International Energy Agency (IEA 2019) highlights that road vehicles, including trucks and two- and three-wheelers, are responsible for roughly three-quarters of global transport-related CO_2 emissions. The IEA posits that in order to address climate change and air pollution in the transport domain, there needs to be an average annual reduction of 3.2% in energy intensity (energy consumption per GDP unit) from 2019 to 2030 (IEA 2019). However, in 2019, the decline was recorded at 2.3%. Figure 1-2 presents the CO_2 emission trends across various transport modes as depicted in the IEA's Sustainable Development Scenario (SDS) for the period 2000-2030. The SDS articulates a collaborative framework for countries to optimize efforts in curbing air pollution and mitigating climate change impacts (IEA 2019).

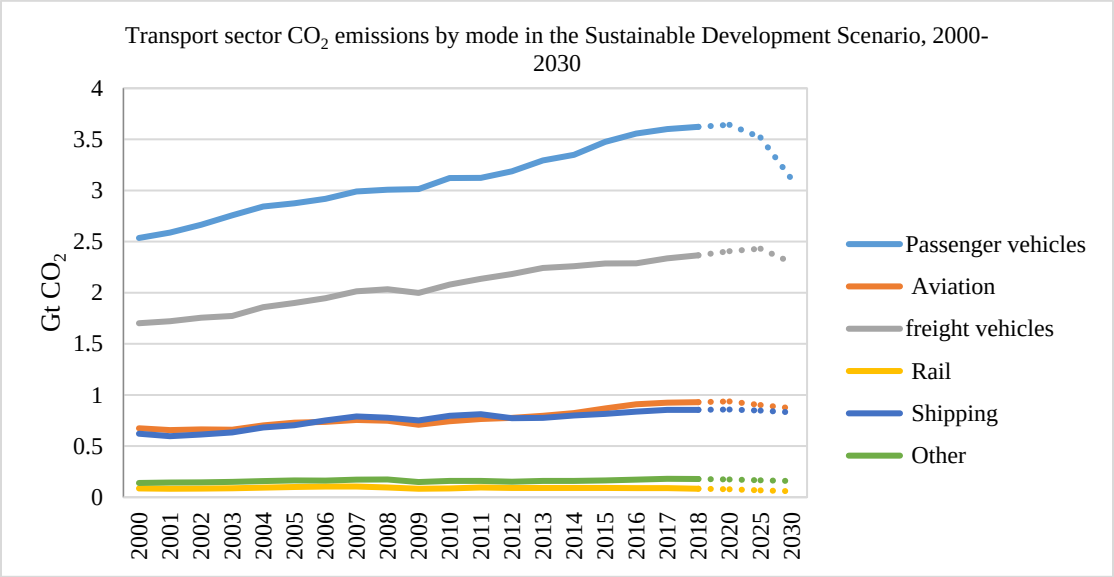


Figure 1-2 SDS in the transport sector by mode, 2000 – 2030 Source: (IEA 2019)

In response to the escalating issue of vehicle emissions, regulatory bodies worldwide have instituted stringent legislation targeting new automobile models. To gauge the effectiveness of these legal measures and project vehicle emissions, researchers have formulated two principal modeling approaches: fuel-based models and travel-based models (Galgamuwa, Perera, and Bandara 2015).

Fuel-based models derive GHG emission estimates from fuel consumption data, usually sourced from taxation records, irrespective of the specifics of traffic activities. These models express emission factors in terms of grams per specific unit of fuel consumed (Zhang, Zhao, and Shen 2011). A notable exemplary within this category is the Computer Program to Calculate Emissions

from Road Transport (COPERT), a model pioneered by the European Environment Agency (EEA) in conversely, travel-based models, as suggested by their nomenclature, employ emission factors in tandem with travel data to compile emission inventories. These models consider not just the Vehicle Mile Travel (VMT) data but also integrate insights from driving patterns or driving cycles, which are paramount in estimating fuel usage.

Driving cycles serve as foundational input for a vast majority of global vehicle emission models, including MOVES and VERSIT+ (Nouri 2015). Essentially, a driving cycle is defined as a sequence of data points that illustrate variations in speed or acceleration against time. Recognizing the significance of driving cycles in accurately estimating emissions, numerous countries periodically design and revise their own cycles to more closely mirror prevailing driving behaviors and conditions. To illustrate, Canada introduced a 5-cycle test driving pattern in 2015.

Historically, driving cycles have been tailored for specific scenarios, such as distinct meteorological conditions or specific road categories. However, numerous research findings affirm that utilizing region-specific or provincial driving cycles in emission calculations yields more precise results compared to broader national or international driving cycles (Al-Samari 2017; Anida et al. 2017; Esteves-Booth et al. 2001; Fotouhi and Montazeri-Gh 2012).

Unlike traditional methodology in which one or a few driving cycles are generated, in this study, we constructed a full set of driving cycles to represent driving behaviors in different driving conditions. It should be noted that in this thesis we used speed, acceleration, and deceleration profiles as a proxy to represent driving behavior. This methodology aims to augment the precision of the emission inventory.

In this thesis, we adopt a novel approach by incorporating a diverse range of driving conditions to formulate a full collection of driving cycles. This strategy aims to enhance the precision of the emission inventory, setting it apart from previous research endeavors.

1.2 Problem Statement

Driving cycles serve pivotal roles across multiple domains such as traffic engineering, vehicle emission calculations, and automotive manufacturing. Within traffic engineering, driving cycles are invaluable tools to evaluate the environmental ramifications of transportation initiatives,

strategic traffic planning, and the introduction of new traffic technologies (Hirschmann et al. 2010). They also form a cornerstone of the vehicle emissions estimation process.

In the realm of vehicle manufacturing, driving cycles are indispensable in designing and forecasting the efficiency of internal combustion (IC) powertrains (Lin et al. 2015; Nyberg 2015), as well as electric propulsion systems and battery performance. Further accentuating their significance, driving cycles are a primary benchmark during the type approval procedure. This process ensures that a new vehicle model adheres to all established emission standards before it is greenlit for market introduction (European Commission Joint Research Center, 2013).

Research highlights significant variances in emission estimation between standard laboratory-formulated driving cycles and real-world, locally based ones (Bodisco and Zare 2019; Samuel, Austin, and Morrey 2002), underlining the need for locally tailored driving cycles (Huertas et al. 2018). For instance, Huertas et al. (2019) found that laboratory settings can overestimate fuel consumption rates by up to 20%.

The pivotal role of exogenous factors in driving cycle variability and, consequently, emissions estimations is well-established in the literature (Giakoumis 2016). Such factors not only induce significant variations in driving cycles but also introduce uncertainty in emissions modeling. A detailed categorization of these factors can be found in Table 2-4 with an in-depth explanation provided in the "Influencing Factors" section.

Upon surveying past research, it's evident that many studies formulated driving cycles without giving due consideration to these external influences. Instead, they often opted for a limited set of conditions to procure necessary data. For instance, studies by Silvas et al (2016), Tzirakis and Zannikos (2013), Zhang et al (2019), Sankar et al (2020) disregarded road features, while others like Ali, Ahmed, and Yang (2020), Zhao et al. (2018) assumed clear weather conditions when crafting their driving cycles.

A closer look at Canada's standard driving cycle reveals a similar pattern. While a select range of temperatures (-7°C, 35°C, and 20°-30°C) are addressed, pivotal factors such as varying precipitation patterns or the intricacies of road networks are conspicuously absent. For instance, Montreal Island experiences a broad spectrum of precipitation types, temperature variations ranging from a chilly -35°C in winter to a sweltering 35°C, and a diverse array of road types in terms of slope, width, and function (like arterial, collector, and local roads). There's likely a marked

discrepancy between emission estimates derived using this standard and the actual emissions, given the significant variability in driving conditions.

Ignoring the reality of driving conditions while formulating driving cycles can undoubtedly skew emission estimates, either inflating or underreporting values. Such inaccuracies can misguide legislative actions and their perceived effectiveness.

In summary, the literature currently lacks thorough exploration of the variability in driving cycles arising from diversities in road conditions and climatic influences. This research seeks to address this gap. By viewing the driving cycle as an evolving and context-dependent entity, this thesis introduces a methodology to construct comprehensive driving cycle sets that truly reflect all road types and weather conditions.

1.3 Research Objectives

Toward improving the accuracy of vehicle emissions estimation, the main objective of this thesis is to put forward a framework to identify and develop an optimal, finite number of driving cycles required to represent a full set of conditions observed in an urban area.

In this thesis, “optimal” is defined as a compromise between a set of values. To optimize the number of driving cycles, we need to balance the trade-offs among the number of driving cycles, information loss, and variability in the driving cycles, which can be influenced by road types, weather types, and type of day. In fact, “optimal” refers to the minimum number of driving cycles while information loss is minimum, and variability of driving cycles is maximum. The number of possible driving cycles in an urban area like Montreal is unlimited and undefined. In other words, since trip characteristics depend on different weather conditions, various road segments, and the type of day, there are an indefinite number of driving cycles in an urban area. Therefore, this study employs the term 'finite' to denote a manageable number of driving cycles that can be produced given the constraints of available data, computational resources, and the need to minimize the influence of outliers.

To achieve the desired objective, we have introduced a three-step approach, with each step having its own specific objectives, to help achieve our main goal. Each method represents a distinct contribution to the current research on generating driving cycles and is discussed in one of the three papers in this thesis.

The first step sought to develop a data-driven approach to enhance the understanding of the relationship between road features and driving speed through three primary contributions. Firstly, it introduced a machine-learning-driven framework to develop a new road segment typology based on essential driving speed determinants, rather than the traditional function-based classifications. This was exemplified using Montreal's road network, highlighting the efficacy of k-means sand Gower similarity measure in identifying five distinct road types. Secondly, it tackled the challenge of validating clustering results in transportation by devising a specific validation method for vehicle speed data, confirming variations in speeds across road categories. Lastly, this thesis redefined a regression model to better capture the intricate associations between road characteristics and driving speeds, stressing that influential road features can vary by type.

The second step aims to develop a new machine learning framework to identify representative driving cycles for different driving conditions. To fulfill these objectives, this study contributes to the literature by developing a methodology to generate and identify the representative driving cycles for various weather conditions. In addition, it developed an approach to examine the relationship between changes in weather parameters and driving cycle characteristics values.

The third step addressed critical gaps in driving cycle generation literature by introducing three primary contributions. Firstly, it introduced a novel approach for data aggregation in driving cycle creation, focusing on optimizing the level of data consolidation to minimize information loss and maximize driving cycle variability. "*Information loss"* is the decrease in detail about driving characteristics that occurs when creating representative driving cycles from the extensive data collected. Secondly, the paper was the first to develop a multilevel mixed model that explored the relationships between road types, weather conditions, and driving cycle parameters, providing deeper insights into driving cycle dynamics. Lastly, this thesis presented a unique methodological framework, asserting that 76 driving cycles is "optimal" number to represent varying driving behaviors, thereby filling the gap typically present in the literature.

1.4 Main Research contributions

Beyond Single Driving Cycle Generation: While previous studies have primarily focused on generating a single driving cycle, this thesis extends this by considering the variability of driving conditions. This includes factors such as different road types, weather, and day types.

Focus on "optimal" number of driving cycles: Another contribution of this research is the development of a methodology aimed at determining the "optimal" number of driving cycles for a given region. This aspect of the study distinguishes it from previous works.

Emphasis on external factors: The core objective of this study is not the construction of a driving cycle. Instead, it seeks to identify a finite, yet "optimal" number of driving cycles that are influenced by a range of external factors. This approach is a significant departure from past methodologies, which have not thoroughly explored this aspect.

1.5 Research Questions

This research contributes to answer the following questions:

1. How many driving cycles are necessary to represent typical driving patterns in Montreal?
2. How to identify relevant typologies of road segments and weather conditions to correctly represent the diversity of driving conditions in an area?
3. What is the "optimal" level of data aggregation to minimize information loss and maximize driving cycle variability?

1.6 Thesis Contents

The remainder of this thesis is structured as follows:

1. Chapter 2 reviews various methodologies for generating driving cycles, discusses factors impacting driving behaviors, and elaborates on the analysis techniques that have been used to investigate the influences of external factors on driving behavior.
2. Chapter 3 explains the research methodology steps to perform the thesis and provides more details about the techniques that are used to achieve the desired research objectives.
3. Chapter 4 presents the first publication of this thesis and provides more details about the methodology that has been developed to generate a new road typology.
4. Chapter 5 describes the second paper of this thesis. It discusses representative number of driving cycles in different driving conditions can be identified.
5. Chapter 6 presents the third publication of this thesis. It discusses the research methodology to determine the "optimal" level of data aggregation and explains the methodology for determining "optimal" driving cycles in Montreal.

6. Chapter 7 recalls and discusses the main motivation and objectives of this thesis. Next, the author provides an overview of the main contributions and discusses the main findings of this thesis.
7. Chapter 8 reviews the research contributions, limitations, and future research.

CHAPTER 2 LITERATURE REVIEW

In this section, we delve into the various facets pertinent to driving cycles, influencing factors, clustering methods, emission modeling, and eco-driving. Furthermore, this section underscores the novel contributions this thesis aims to offer, illuminating the necessity of this research against the backdrop of existing scholarly works. The principal themes explored are as follows:

Driving Cycle includes:

- Definition and Applications: This subsection elucidates diverse interpretations and conceptualizations of a driving cycle. Additionally, the myriad applications of driving cycles, particularly in the realm of vehicular emissions, are outlined.
- Legislative vs. Non-Legislative Driving Cycles: A thorough discourse on the dichotomy of legislative and non-legislative driving cycles is presented, highlighting their distinct characteristics and implications.
- Prominent Universal Cycles: The focus here shifts to globally recognized driving cycles, with special emphasis on those formulated in the USA and Canada. The origins, developments, and salient features of these cycles are recapitulated.
- Driving Cycle Construction: A comprehensive inventory detailing the steps, methodologies, and techniques employed in the design and formulation of driving cycles is detailed in this segment.

Influencing Factors delves into:

- Impact on Emissions and Driving Cycles: This section endeavors to examine and elucidate the plethora of factors that influence over both vehicle emissions and the intricacies of driving cycles. The objective is to provide a holistic understanding of how myriad of determinants interplay to shape driving behaviors and emission profiles.

Clustering Methods encompasses:

- Overview of Clustering Approaches: Offering an overview of clustering methodologies, this subsection introduces various clustering paradigms and their applications.
- Clustering in Driving Cycle Research: A deeper exploration into prior research endeavors that have leveraged clustering techniques for categorizing influencing factors and for devising driving cycles is presented. This will provide a clearer perspective on the evolution of clustering techniques in this domain.

As we navigate through the of existing literature, the overarching objective remains: to discern where our research can carve its niche, offering both innovation and enhanced understanding in the broader context of driving cycles and vehicular emissions.

2.1 Driving cycles

Driving cycles lack a universally agreed-upon definition, with numerous interpretations proposed by different researchers and practitioners (Huertas et al. 2018). One widely accepted definition suggests that a driving cycle consists of a series of data points reflecting speed variations over time (Kalgal et al. 2017). Barlow et al. (2009) defined it as a speed-time profile used to simulate car operations for lab-based emission tests. Huertas et al. (2018) described it as a collection of parameters indicating a driver's behavior, though it remains ambiguous as to which parameters best illustrate driving habits. Nesamani and Subramanian (2011) elaborated on this by specifying parameters such as "idle, acceleration, deceleration, and steady states" to represent typical urban driving patterns. Similarly, Amirjamshidi and Roorda (2013) identified a driving cycle as a specific speed-time profile within which a vehicle might be idling, accelerating, decelerating, or cruising. Achour and Olabi (2016) view driving cycles as tools to evaluate the environmental consequences of traffic plans and to optimize vehicle powertrains to minimize fuel use. Shi et al. (2016) described it as a representation of either a specific area's vehicle operating conditions or a typical vehicle. In this thesis, a driving cycle is defined as a speed-time profile, generated from second-by-second driving data, whose length corresponds to the average length of trips from which it is derived. This concept of a driving cycle is dynamic, varying according to different road conditions, weather, and types of days, thereby resulting in a range of diverse profiles.

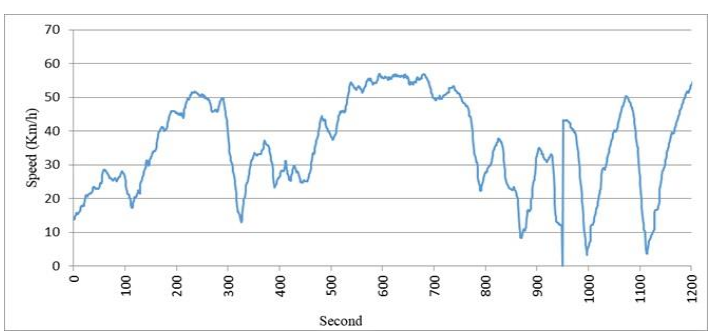


Figure 2-1 Driving cycle profile-image by author, 2023

The concept of the Driving Cycle was first introduced in the early 1960s. Over the past forty years, automotive industry leaders, including the USA, European nations, and Japan, have developed standard driving cycles, established emission goals, and formalized testing protocols. Meanwhile, certain countries have utilized these driving cycles to perform emission inventories (Galgamuwa, Perera, and Bandara 2015).

Driving cycles can be primarily divided into two categories: standard and local driving cycles. These categories differ in their methods of data collection, uses, and development processes.

Standard driving cycles (SDCs) can be largely classified into two main categories based on their structure and representation: transient and modal. Transient cycles are characterized by continuous variations in speed over time without specific stable states. These cycles provide a continuous trace of the vehicle's operation, mimicking real-world driving with its unpredictable nature of accelerations, decelerations, and cruises. The outstanding example of SDCs are New European Driving Cycle (NEDC), Federal Test Procedure (FTP-75), and Worldwide Harmonized Light Vehicles Test Procedure (WLTP). In contrast to the transient cycles, modal cycles break down vehicle operation into discrete modes or stable states, like idle, cruise, acceleration, or deceleration.

These modes are then averaged over a certain period, providing a more simplified representation of driving. IM240 (Inspection/Maintenance 240), SCC (Steady-State Cycle), and Bag Tests are examples of the modal cycles.

Building upon prior discussions, the precision of the standard driving cycle has been found to be insufficient for certain applications (Gebisa et al. 2021). This has prompted numerous scholars to devise local driving cycles tailored to specific conditions. Subsequent sections delve into the diverse methodologies employed in formulating these local driving cycles.

It is crucial to emphasize that the reliability of the devised driving cycles, representing authentic driving behaviors, primarily relies on these three pivotal factors, as underscored by (Ashtari, Bibeau, and Shahidinejad 2014; Huertas et al. 2019):

- Data Quality: The volume, accuracy, and precision of information pertaining to vehicle operation and driving conditions play a substantial role in determining the accuracy of the developed driving cycles.

- Performance Values (PV): These are metrics introduced to typify driving behaviors, including aspects like idling periods, average speeds, and cycle durations. Such PVs serve as benchmarks during the formulation of driving cycles.
- Technological Precision: The accuracy of the tools and methodologies used in the construction of driving cycles is paramount.

The precision of all three aforementioned factors is pivotal for the development of an accurate driving cycle. A survey of pertinent literature reveals that many researchers adopt similar methodologies in terms of data volume and driving conditions when crafting driving cycles. For instance, Ashtari, Bibeau, and Shahidinejad (2014) employed a dataset of 44 million data points encompassing the entirety of Winnipeg city for their driving cycle construction. Similarly Günther et al. (2017) formulated a driving cycle based on nearly 27,000 driving hours. Notably, both these studies overlooked the integration of driving conditions in their methodologies.

Furthermore, the field lacks a standardized methodology regarding the choice of Performance Values (PVs). This discrepancy is evident in the variance in the number of parameters considered by different researchers. Ashtari, Bibeau, and Shahidinejad (2014); Chauhan, Joshi, and Parida 2020; Günther et al. (2017) opted for 14, 7, and 4 parameters, respectively.

The process of formulating a driving cycle is intricate, generally unfolding in a series of systematic steps. These steps typically encompass:

- Route Selection: This involves identifying routes that aptly mirror the urban road network and its variances.
- Data Collection: Once the routes are determined, data from these chosen paths are collected, capturing a plethora of driving conditions and patterns.
- Microtrip Generation: The extensive data gathered is then disaggregated into smaller segments or 'microtrips' each showcasing specific driving behaviors.
- Criteria Establishment: To guide the development and evaluation of the driving cycles, definitive assessment criteria are delineated.
- Candidate Driving Cycle Development: Based on the data and set criteria, potential driving cycles are formulated for further scrutiny.
- Evaluation and Final Selection: These candidate driving cycles undergo rigorous assessment, post which the most representative cycles are pinpointed.

The subsequent sections delve deeper into each of these pivotal phases, elucidating their intricacies and importance.

2.1.1 Route Selection

The initial and arguably the most critical phase in the crafting of driving cycles is the selection of representative routes. A thorough review of academic literature indicates that there are predominantly two methodologies that researchers adopt to incorporate road characteristics into the driving cycle development process. These two approaches, integral to ensuring the authenticity and relevance of the driving cycles, are elaborated upon in the subsequent sections.

Selecting Representative Routes

The preliminary method adopted by researchers emphasizes selecting a limited subset of streets or particular sections of a network that encapsulate the essence of the entire urban road network. The primary rationale behind narrowing down the selection, rather than encompassing the entirety of urban routes, stems from the prohibitive costs of data collection and the multifaceted nature of road features. These constraints have, in several instances, rendered it infeasible for studies to create driving cycles for every route within a city (Anida et al. 2017; Mahayadin et al. 2017).

It is imperative that the chosen roads capture the nuances of the urban road network's characteristics. An imprecise selection, skewed towards specific road types, risks introducing biases in data collection, consequently affecting the outcomes derived from the driving cycle (Tong and Hung 2010). Route selection is imprecise if the selected route does not represent appropriately an urban network. After the identification process, researchers then proceed to accumulate driving data from the selected routes, serving as the foundation for driving cycle construction.

Parameters frequently invoked for road selection encompass traffic metrics, land use, and road typologies. A noticeable gap in existing studies is the scant consideration given to road physical features. Yet, as underscored earlier studies, numerous road attributes, such as gradients, road categories, functions, width, intersection densities, and speed limits, play a pivotal role in shaping driving conditions.

A study by Nguyen et al. (2019) employed expert insights to delineate a representative route for Hanoi, Vietnam's driving cycle. Their methodology hinged on home-to-work commute patterns,

population distribution, and road classifications, culminating in the selection of three bus routes bridging the city's core with its periphery.

In a methodical approach, Mahayadin et al. (2017) delineated a four-step process to identify representative routes in Malaysia's Penang region. Their strategy entailed gauging the proportion of peak-hour traffic, leveraging the Highway Capacity Manual (HCM 2010) to ascertain the Level Of Service (LOS) for each route, subsequently selecting routes based on their LOS and traffic volume. Their methodology culminated in evaluating the geographical spread of their chosen routes using Google Maps.

Nutramon and Supachart (2009) pivoted their focus to traffic density and road classifications in their endeavor to select representative routes for Bangkok, Thailand. The authors employed a traffic model, mapping travel speed distributions across the city's 20 primary roads, eventually narrowing down to seven key roads.

Hung et al. (2007) embarked on their study for representative routes in Hong Kong, anchoring their methodology in land use and traffic data. Their approach utilized land use patterns to discern primary travel trajectories, further leveraging the Annual Average Daily Traffic (AADT) to single out the city's most frequented highways. To ensure comprehensive coverage, they categorized urban networks into three subsets—suburban, urban, and highway, eventually finalizing their selection based on these classifications.

Lastly, Hung et al. (2005) fashioned driving cycles for China's Pearl River Delta. Their methodology, geared towards identifying a representative route, leaned heavily on traffic flow intensity and road type classifications, with the former playing a decisive role in highlighting primary commute routes for their study.

With the increasing availability of driving data in recent years, some research endeavors have sought to construct driving cycles without resorting to the selection of representative routes. One such study by Ashtari, Bibeau, and Shahidinejad (2014) tapped into a pool of 76 volunteer drivers to amass driving data in Winnipeg, Canada. Their intensive data collection spanned an entire year, amassing approximately 44 million GPS data points across the entirety of the city's urban network, facilitating the design of representative driving cycles.

Incorporating Road Features in the Construction Process

This alternative methodology diverges from the typical route-selection model. Instead, it places significant emphasis on the integration of road features, especially gradients, directly into the driving cycle's construction process. By doing so, researchers can more accurately gauge and quantify the effects these features have on driving patterns. The ensuing discussion offers a snapshot of notable contributions in this arena. It is pertinent to highlight that this method, though promising, has seen relatively limited adoption compared to its predecessor.

Tzirakis et al. (2013) spearheaded an exploration into the ramifications of road gradient on vehicular emissions. Their investigative approach hinged on the development of two distinct methods to derive driving cycles. The first method, denoted as "Method A," catered to shorter driving durations, whereas its counterpart was tailored for longer durations. Notably, while "Method A" presented a time-intensive yet highly accurate approach, the alternative method traded off accuracy for efficiency, necessitating an expansive dataset but promising expedited processing times.

Within "Method A" as devised by Tzirakis et al. (2013), the impact of road slope is intricately evaluated. Equation (2.1) is employed to compute the influence of road gradient on acceleration, and this computed value is subsequently integrated into the positive and negative acceleration values.

$$\alpha_{add} = g\,(h2 - \mathrm{h1})/d \qquad (2.1)$$

In equation (2.1), g is gravity, h is elevation, and d is distance.

For method “B”, using equation (2.2) the influence of road gradient on acceleration is calculated and added to positive and negative acceleration.

$$P_{gradient} = \left[\frac{h2 - h1}{v}\right] g\, v\, mvehicle = (h2 - h1) g\, mvehicle \qquad (2.2)$$

Where (h2−h1) calculates the change in elevation or vertical displacement, and when multiplied by g and mvehicle (mass of the vehicle), it gives the change in potential energy per unit time, which is power. The division by v (velocity) in the first part of the equation suggests a normalization over

a unit of distance, effectively determining the power needed to overcome the gradient at a particular speed.

2.1.2 Driving Cycle Data Collection

Undoubtedly, the cornerstone of driving cycle construction lies in the meticulous accumulation of precise and exhaustive data pertinent to the vehicles under study (Giakoumis 2016). Ideally, the curated database ought to encapsulate a myriad of observations, encompassing the entire gamut of driving patterns and associated conditions.

Historically, two prevalent methodologies have been championed for the acquisition of driving data essential for crafting driving cycles: the car chasing technique and the on-board instrumentation technique.

Chase Car Technique

The chase car technique, historically recognized as a conventional data collection method, involves the selection of a random vehicle, termed the "target vehicle", which is subsequently pursued by a vehicle fitted with monitoring instruments, known as the "test car". This setup facilitates the real-time recording of the target vehicle's driving behaviors, either through laser technology or the non-lock method (Nouri 2015). While laser technology directly acquires driving activities from the target vehicle, the non-lock approach relies on a GPS system integrated within the test car.

One of the salient merits of the chase car method is its economical nature. Nonetheless, executing this approach on routes with diminished Level Of Service (LOS) poses notable challenges and can lead to compromises in data accuracy (Galgamuwa, Perera, and Bandara 2015). A quintessential issue arises when either the initiation or termination of a trip remains unrecorded. This generally transpires when a target vehicle embarks or concludes its journey outside the predetermined routes, thus not being under observation throughout its entirety. Moreover, the datasets generated through this technique tend to be rather limited in scope (Ashtari, Bibeau, and Shahidinejad 2014; Zhao et al. 2018).

Globally renowned driving cycles, such as LA01 and LA92 formulated in the USA, amassed requisite data utilizing the chase car technique. Parallelly, Esteves-Booth et al. (2001) harnessed this method to curate data pivotal for creating a localized driving cycle specific to Edinburgh.

Similarly, Niemeier and Eisinger (2008) capitalized on this methodology in the development of an international driving cycle.

On-Board Instrumentation Technique

The on-board instrumentation technique primarily employs On-Board Diagnostics (OBD) to extract trip activity specifics from the test vehicle. Essentially, the OBD serves as a computerized system fitted within vehicles aimed at scrutinizing engine components. In this technique, a driver operates the vehicle unhindered on pre-established routes, with the OBD cataloging speed data at second-by-second intervals.

Unlike the chase car methodology where a delay in speed-change recording can arise owing to the driver's reaction time, on-board instrumentation derives its readings directly from the test vehicle. This method inherently reduces human interference in the data collection process. Nonetheless, the upfront cost of this system tends to be higher, and the scope of data collection remains restricted. Consequently, when extrapolated to broader studies, the associated costs escalate substantially.

There exists a dichotomy of perspectives regarding the accuracy of data curated via OBD. While Zhao et al. (2018) advocate for the technique, contending it faithfully depicts driving scenarios, Galgamuwa, Perera, and Bandara (2016) express reservations. They argue that since routes are predefined, the recorded trip activities might not genuinely mirror real-world driving behaviors. Furthermore, the knowledge that their driving is under scrutiny may subtly influence the actions of test vehicle drivers, possibly making them unrepresentative of the average driver's reactions.

Historical references like the ARTEMIS driving cycle testify to the utility of the OBD method. Throughout the course of the ARTEMIS project, routes were chosen randomly, and a commendable 2200 hours of driving data was accrued over 2000 days. Similarly, the Australian driving cycle, known colloquially as the Composite Urban Emissions Drive Cycle (CUEDC), gathered approximately 431 hours of driving data to bolster the accuracy of the devised driving cycle. Literature further highlights the appropriateness of this method for regions with a notable proportion of erratic or aggressive drivers.

In summary, harnessing OBD in daily driving scenarios is arguably the most comprehensive approach for data collection, as it possesses the capacity to encompass a diverse array of driving conditions.

2.1.3 GPS Data Processing

GPS (Global Positioning System) technology is instrumental in both previously discussed data acquisition methods. The quality of the data yielded significantly determines the precision of the resultant driving cycle. GPS data, despite its ubiquity and advancements, is not immune to errors that can skew results.

Ranacher et al. (2016) noted that GPS data is prone to discrepancies, especially concerning positioning accuracy and distance estimations between successive points. Such inconsistencies can introduce a systemic bias in recorded distances. The primary sources of these inaccuracies, as identified by (Nguyen et al. 2020), include:

1. White Noise: This type of error is characterized by a signal that retains consistent intensity across various frequencies. In the context of GPS, white noise can manifest as random errors owing to atmospheric interferences, clock inaccuracies, or electronic noise in the receiver.
2. Outliers: These are anomalous data points that deviate significantly from the neighboring values. Outliers can emerge due to various reasons such as sudden signal drop, hardware malfunctions, or environmental factors like tall buildings or dense foliage that interfere with the signal.
3. Signal Loss: Often, a GPS device may lose its connection with satellites, resulting in a temporary interruption of data collection. Factors like tunnels, dense urban environments, or electronic interference can cause such losses, leading to gaps in the recorded data.
4. Multipath Errors: Multipath issues arise when the GPS signal reaches the receiver after reflecting off local obstacles like buildings, mountains, or other structures. This reflection causes the signal to travel a longer path than it would in a direct line. As a result, the recorded position can appear offset from the actual location due to the additional travel time of the signal.

To ensure the reliability of a driving cycle developed from GPS data, it is essential to apply robust data cleaning and processing techniques. These methods can mitigate the aforementioned errors and produce a more accurate representation of driving conditions. Implementing filters, quality checks, and statistical analyses can help in identifying and rectifying such anomalies in the dataset.

The accurate representation of driving behaviors from GPS data is essential for creating a reliable driving cycle. With the numerous errors associated with GPS data acquisition, filtering techniques play a pivotal role in refining the data. Duran and Earleywine 2012; Nguyen et al. (2020) have proposed eight types of filters to enhance the data quality:

1. Microtrip Initialization: This filter segments the continuous GPS data into smaller sections or 'microtrips', usually based on periods of inactivity or stops, to facilitate a more structured analysis.
2. Converting Timeline: Instead of working with the absolute time (HH:MM:SS), this filter standardizes the time series by resetting the beginning of each microtrip to a zero timeline, making subsequent processing more manageable.
3. Omit Duplicate Data Points: If two consecutive data points have the same timestamp and a zero value, one of them is redundant and is thus removed.
4. Replacing Abnormal Velocity Value: This filter compares the velocity of each data point against preset urban speed limits. Those that exceed these limits are identified as anomalies and are replaced using interpolation methods to approximate a realistic value based on surrounding data points.
5. Removing Velocity During Idle Time: As the microtrip should ideally start with zero velocity, any initial data point indicating otherwise (due to errors like multipath signals) is corrected to zero.
6. Replacing False Zero Velocity Values: A data point with zero velocity is suspicious if its neighboring data points suggest motion. This filter removes such inconsistencies by replacing the erroneous zero values using interpolation from surrounding data.
7. Filling Signal Gaps: In urban environments with tall buildings or obstructions, the GPS signal may get temporarily lost, creating gaps in the dataset. This filter identifies these gaps and uses interpolation techniques to fill in the missing values.
8. Smoothing and Denoising: For a cleaner representation of driving behavior, smoothing out noisy data is crucial. The Kalman filter, a filtering algorithm, is particularly favored for this purpose. It uses statistical methods to estimate and correct errors, like noise, in the data. Jun, Guensler, and Ogle (2006) provided specific equations tailored for the application of the Kalman filter in the context of GPS data denoising and smoothing.

When appropriately applied, these filters significantly enhance the accuracy and reliability of GPS data, making it a robust foundation for constructing driving cycles.

2.1.4 Microtrips Generation

After cleansing the data, microtrips are generated. The generation of microtrips is the process of segmenting an entire trip into smaller pieces, referred to as microtrips. This is done to better understand average driving patterns. There are various methodologies for categorizing these processes. For example, Liu et al. 2020; Niemeier and Eisinger (2008) categorized them into four classes, including microtrips cycle construction, segment-based cycle construction, pattern classification, and modal cycle construction. Galgamuwa, Perera, and Bandara (2015) classified them into two groups: microtrips-based and stochastic modal cycle. Generally, the selection of each method depends on factors such as road type, traffic condition, and vehicle type. (Mayakuntla and Verma 2018) have shown that the microtrips-based approach is not accurate enough and tends to overestimate or underestimate real driving patterns. The authors believe this issue is related to the limitations of driving cycle parameters, which cannot adequately reflect variations in speed within a given microtrip.

Microtrips-based method is the main approach to construct a driving cycle. The most comprehensive list which covers different methodologies are as follows:

1. Adjacent stops;
2. Event bins based on acceleration;
3. Segment-based (intersections);
4. Pattern-based cycles construction;
5. Fixed time interval;
6. Fixed distance.

The initial two methodologies are predominantly utilized in the field. As delineated by (Huertas et al. 2019), both approaches exhibit stochastic characteristics. This implies that while these methods can consistently yield similar outcomes upon repetition, they do not guarantee identical results each time. Subsequent sections delve deeper into the specifics of each approach.

Adjacent stops

In the method known as "Adjacent stops", driving data is partitioned between two successive halts—these can be full stops or instances when the vehicle idles. The resulting microtrips are then sorted based on average speed and pieced together by referencing specific driving cycle parameters. This linkage can be undertaken through three distinct methods: selecting randomly, employing the best incremental strategy, or using a combination of both.

However, the technique does come with inherent limitations. Firstly, it might not accurately depict urban driving habits, especially during rush hours. This is mainly because, in such situations, the frequency of stops often surges, leading to more truncated microtrips. Additionally, during data preparation, any microtrip that is shorter than 10 seconds is typically discarded since it is believed they might not represent true driving patterns. This aspect could result in the method failing to capture behaviors in high-traffic scenarios, a critical factor for emission estimations. Secondly, it is challenging to gain insights into driving conditions, such as the type of road or its Level Of Service (LOS), from the features of microtrips. This makes recreating specific driving behaviors in certain driving conditions a challenge.

For instance, Kamble, Mathew, and Sharma (2009)**,** which aimed at devising a driving cycle for Pune, India, used the microtrips-based cycle construction. They pinpointed specific routes, derived microtrips using the adjacent stops methodology, and devised distinct criteria for evaluations. Another example is the work of Wang et al. (2008), who tailored the Hong Kong Cycle. They chose eleven cities, had experts identify prominent routes as representative of the road network, gathered data, generated microtrips via the adjacent stops strategy, and finally selected a subset of these trips to draft a driving cycle, which was then assessed against set benchmarks.

Event Bins Based on Acceleration (Markov chain model)

Vehicle emissions are closely linked to the vehicle's operating mode—namely acceleration, deceleration, idle, and cruise (An et al. 1998; Ritner et al. 2013).

In 2002, Jie Lin and Niemeier proposed a unique methodology for microtrip generation, designed to categorize driving pattern data based on these specific vehicle operation modes. Two foundational assumptions underpin this approach: first, the selection of an operating mode is solely influenced by its immediate preceding mode, and second, the Markov chain can be employed to model driving behaviors.

Within this framework, the Maximum Likelihood Estimation (MLE) algorithm, in conjunction with the k-means clustering method, segments driving data by aligning it with modal activities. These activities can be broadly categorized into acceleration, deceleration, cruising, and idling. During the microtrip generation phase, the MLE partitions the driving data into segments based on acceleration patterns. Subsequently, using the same algorithm, these segments are grouped or clustered based on criteria such as their average, peak, and lowest speed, and acceleration metrics.

To sequence the generated microtrips in a logical and representative manner, a transition matrix is formulated. This matrix offers insights into the likelihood or probability of transitioning between different operational modes. In the end, this structured approach culminates in the formulation of a driving cycle, conceptualized as a Markov chain. The assumptions of this technique are discussed in (Lee and Filipi 2011; Zähringer, Kalt, and Lienkamp 2020).

Segment-based driving cycle construction

The segment-based approach to construct driving cycle generate microtrips for segments between intersections, exhibits certain similarities to the adjacent stops technique. However, it diverges in its segmentation criteria. Instead of basing segments on adjacent stops, this approach segments the vehicle speed/time profile by roadways and Levels of Service (LOS). This is more suitable for generating driving cycle for a highway since there are no adjacent stops and a microtrip approach cannot be implemented. Consequently, this method enables the representation of driving behaviors and prevailing traffic conditions on specific roadways. Notably, the segment-based approach finds more resonance among traffic engineers than those engaged in emission studies.

However, this approach is not without its complexities. Unlike the adjacent stops method where microtrips primarily begin and conclude at zero speeds, segment-based microtrips frequently commence and terminate at variable speeds. As a result, the challenge arises in ensuring the seamless concatenation of consecutive microtrips. Factors such as speed, acceleration, and time become paramount in ensuring this integration.

Highlighting its practical application, Zito and Primerano (2005) employed the segment-based approach to devise the Australian Composite Urban Emissions Drive Cycle (CUEDC). Their process initiated with the classification of urban road networks into four distinct categories: congested, residential, highway class, and arterial classes. Subsequently, microtrips are assigned to

these categories based on their average speed values. After this categorization, microtrips are concatenated, and driving cycles are constructed.

However, it is pivotal to acknowledge the primary limitation of the segment-based methodology. Predominantly crafted for transportation engineering purposes, its application appears less aligned with emission studies. Actually, this approach often consists of predefined, repetitive segments (like urban, rural, highway) and might not capture the real-world variability in driving patterns. Real-world driving patterns includes a wide range of speeds, accelerations, idling, and other factors that can significantly impact emission levels.

Pattern-based cycles construction

The pattern-based cycle construction approach emphasizes the identification of kinematic sequences within a driving trip. In this methodology, a single trip is perceived as an amalgamation of distinct kinematic sequences. These sequences are then segregated into heterogeneous classes. Following this stratification, succession probabilities are leveraged to anticipate the likelihood of specific kinematic sequence transitions. A transition matrix, encapsulating these probabilities, is subsequently formulated. Using this matrix and through random selection, activity classes are concatenated, culminating in the creation of a driving cycle.

This approach finds precedence in the construction of non-legislative driving cycles in Europe. One such notable instance is the crafting of the European Urban Cycles. The development process entailed the characterization of the kinematic sequence through 20 distinct variables. These included parameters such as duration, idle time, distance throughout the sequence, mean values, maxima, standard deviations, percentiles for driving modes (20% and 80%), instantaneous speed and acceleration, and the spatial distance between successive accelerations. To cluster these kinematic sequences, the Principal Component Analysis (PCA) method was adopted. The analysis identified four primary classes: congested urban traffic, free-flow urban traffic, extra-urban, and motorway driving conditions.

Highlighting its broader applicability, Zhao et al. (2018) employed this methodology in the design of a driving cycle tailored for Electric Vehicles (EVs) in Xi’an, China. The impetus for this endeavor was to bridge the disparity between real-world driving cycles and international ones, especially in the realm of fuel consumption estimation. After the route selection and data procurement using ODB, microtrips were fashioned through the pattern-based paradigm. An

equation was deployed to fragment the entire trip, spanning 428 hours, into 36,388 kinematic segments. To manage the complexity of computations and analytical challenges, PCA was applied for dimensionality reduction of microtrips. Following this, a hybrid technique, melding k-means and the Support Vector Machine (SVM), was engaged both to categorize the microtrips and to finalize the driving cycle.

The pattern-based cycle construction technique presents a few salient limitations. First and foremost, its direct applicability in estimating vehicle emissions is questionable since it does not include exhaust emissions data, and vehicle emissions do not directly match with driving patterns. Secondly, the method's reliance on the chi-square distance for the classification of sequences, particularly within the joint distribution of speed (v) and acceleration (a), adds a layer of complexity and potential inaccuracy to the process as indicated by (Giakoumis 2016). Lastly, the approach is notably time-intensive, necessitating vast datasets to ensure accurate route categorization.

Fixed distance driving cycle construction

The fixed-distance method, as conceptualized by (Nouri 2015), partitions driving data into uniform distance intervals, with subsequent cycle construction steps aligning closely to the methodologies employed in the adjacent stops cycle construction.

In a study by Nouri and Morency (2017), a localized driving cycle was engineered specifically for Montreal Island, Canada. Data acquisition was realized using the Chase car and Onboard Instrumentation Techniques. With the data at hand, they engaged six distinct microtrip strategies, as delineated in 2.1.4, to ascertain the most precise method for curating a driving cycle. The SAFD (Speed-Acceleration Frequency Distribution) matrix was utilized to chart the recurring combinations of speed and acceleration within the generated microtrips. Subsequently, dissimilarity matrices were fashioned, offering a comparative insight into the “speed acceleration time distribution matrix” for each individual microtrip. It is worth noting that speed acceleration time distribution is referred to as SAFD.

To further segment the microtrips, the k-means clustering technique was deployed, yielding seven distinct classes. Transition matrices, indicating the probabilistic transition between clusters, were formulated. Following this, using the principles of the Markov chain model, a total of 12 driving cycles were constructed. In ensuing phases, these driving cycles underwent evaluation against 14 cycle assessment parameters, with the aim of distinguishing the most representative driving cycle.

Conclusively, the study underscored the superior accuracy of the fixed distance method in microtrip generation, relative to other approaches.

Fixed time interval

In this approach, a constant time interval is defined to segment trip data and the rest of the method is same as fixed-distance driving cycle construction approach (Nouri 2015).

2.1.5 Microtrips Classification

Once the microtrips have been generated, the subsequent phase entails classifying these microtrips. This step is indispensable in clustering microtrips that exhibit analogous driving behaviors or patterns, ensuring a more streamlined and organized dataset.

To initiate this, certain driving characteristics are chosen to evaluate the likeness or differences among the microtrips. Prominent metrics include average speed, average acceleration, and the Speed-Acceleration Frequency Distribution (SAFD). These metrics serve as the foundation upon which distance or dissimilarity matrices are crafted, essentially providing a structured representation of how each microtrip compares to others in terms of the selected features.

Though a myriad of techniques exists to calculate the distance between data pairs, some stand out due to their efficacy and widespread use. Algorithms such as Euclidean have become popular tools in this process due to their precision and adaptability (Mafi et al. 2022; Mi et al. 2019).

Having established the distance or dissimilarity matrices, the next endeavor is to systematically group the microtrips. Clustering algorithms are particularly adept at this, and among the gamut of algorithms available, k-means sand k-medoids are the most prevalently employed. Their computational efficiency in segmenting data into distinct clusters, based on the pre-determined metrics, makes them a preferred choice in the realm of microtrip classification.

2.1.6 Assessment Measures

Driving cycles can be characterized by many parameters such as average speed, maximum acceleration, idling time etc. These parameters are used in a cycle assessment, which makes sure that the constructed cycle reflects the real driving patterns correctly. Thus, these must be selected in a way that reflect the features of real driving patterns correctly.

Assessment measures are mostly selected from driving activity parameters and Vehicle Speed Power (VSP). Activity measures are statistics like speed, acceleration, and deceleration. VSP is the estimation of required power during running a vehicle. It is a function of instantaneous speed, acceleration and road grades. Yu et al. (2010) confirmed that VSP parameters is 2% more accurate in emission estimation than driving activities.

The number of considered parameters in driving cycle construction process is different in various papers. For example, Table 2-1 shows the variation in the number of characteristic parameters in famous local driving cycles around the world. In addition, characteristic parameters are formulated differently from one study to another. For instance, Wang et al. (2008), Chen et al. (2003) considered starting time when $a > 0.1$ m/s^2 to calculate "time spent accelerating"; Arun et al. (2017) calculate it when $V > 5$ km/h and $a > 0.1$ m/s^2.

Generally, it can be recognized that most countries employ 10 to 13 parameters to assess the constructed driving cycles although some studies, like (Ashtari, Bibeau, and Shahidinejad 2014), did not provide any reasoning on their choices. Nevertheless, other researchers expounded on how the parameters have been chosen for their approaches.

The most important local driving cycles around the world and their assessment criteria are listed in Table 2-1. It shows what criteria they considered to construct their driving cycle and how many driving cycles they constructed to represent driving behaviors. The dash sign means the authors did not consider the parameters. V (km/h) is average speed including zero speed, D (m) is distance traveled, a(m/s^2) is average acceleration, d(m/s^2) is average deceleration, t(m)is total test time, and IT(%) is idling time where V and a are zero, AT is time proportion where $a>0$ m/s^2 , DT is time proportion where $a<0$ m/s^2, CT shows time proportion where the vehicle is in cruising mode, and NC indicates the number of constructed driving cycles.

Table 2-1 Characteristic parameters in constructed driving cycles

Author	Region	V	D	*a*	*d*	t	CT	IT	AT	DT	NC
(Z. Wang et al. 2019)	Beijing	38.5	-	-	-	1200	28%	16%	29%	27%	1
(Y. L. T. Nguyen et al. 2019)	Hanoi	19.4	10.0	0.41	−0.38	1862	20%	10%	33%	37%	7
(Mayakuntla and Verma 2018)	Bangalore	16.2	9.4	1.51	−1.76	2088	14%	22%	35%	30%	7
(Pouresmaeili, Aghayan, and Taghizadeh 2018)	Mashhad	20.3	-	0.53	−0.54	-	3%	22%	37%	38%	1
(Arun et al. 2017)	Chennai	22.1	-	0.61	−0.71	-	20%	16%	34%	29%	16
(Roso et al. 2015)	Santa Maria	42.9	11.7	-	-	1294	3%	0%	45%	53%	2
(Nouri 2015)	Montreal	25.42	898	.43	-0.47	2760	12.51%	19.97%	34.57%	31.28%	1
(Ho, Wong, and Chang 2014)	Singapore	32.8	21.5	-	-	2344	26%	21%	29%	25%	1
(Knez et al. 2014)	Celje	25.5	13.0	0.79	−0.84	2453	25%	25%	26%	25%	4
(Kamble, Mathew, and Sharma 2009)	Pune	19.6	-	-	-	1533	56%	18%	14%	11%	1
(W. T. Hung et al. 2007)	Hong Kong	15.5	6.3	.55	-.59	1471	9%	31%	31%	29%	5

The most important traffic-related parameters for cycle assessment along with statistical categories are presented in Table 2-2Table 2-1.

Table 2-2 Traffic-related parameters for cycle assessment

Statistical category	Parameters
Maximum and minimums	Speed, acceleration, deceleration, acceleration rate
Percentages	Idle time, acceleration time, deceleration time, cruise time, creeping time
Means and averages	Average speed, average running speed, average acceleration and average deceleration, mean length of a driving period, average number of acceleration deceleration changes within one driving period.
Standard deviations and percentiles	Speed standard deviation and acceleration standard deviation, speed at the 95th percentile, acceleration/ /deceleration rate at the 95th percentile.
Other approaches	Roots mean square acceleration, Positive acceleration Kinetic Energy, Speed Acceleration Probability Distribution (SAPD) or Speed Acceleration Frequency Distribution (SAFD)[1].

Abas, Rajoo, and Zainal Abidin (2018) undertook the development of the Malaysia's inaugural urban driving cycle, with a primary focus on Kuala Lumpur, attributed to its high vehicular density and population. Utilizing instrumented vehicles, data was collected from five principal roads, restricting the collection to weekdays. Diverging from conventional methodologies, this study incorporated engine parameters in addition to the standard traffic and operation mode metrics. To manage the high-dimensionality inherent to their data, a Principal Component Analysis (PCA) was implemented, resulting in the selection of six engine-specific and eleven traffic-related parameters. Comparative analysis revealed that the devised urban driving cycle exhibited an 8.5% deviation[2] from real-world conditions, presenting a notable improvement over the 43.5% deviation observed in the New European Driving Cycle (NEDC).

[1] . SAFD focuses on the frequency or count of specific speed-acceleration events, while SAPD deals with the probability or likelihood of these events occurring.
[2] Deviation refers to difference between estimated vehicle emission and real-world vehicle emissions.

Fotouhi and Montazeri-Gh (2012) embarked on the development of urban driving cycles specific to Tehran, leveraging the Advanced Vehicle Location (AVL) system for a six-month data collection phase without geographical restriction. Recognizing the paramount influence of 'average speed' and 'idle time percentage' on vehicular emissions, these parameters were prioritized. The study employed an adjacent stops methodology for microtrip generation and subsequently, the k-means algorithm for microtrip classification. The strength of this approach lies in its computational efficiency and adaptability across diverse datasets.

Ericsson (2001) set forth to devise a driving cycle tailored to Lund, Sweden. Through extensive data collection, a dataset comprising 19,230 driving patterns was acquired via GPS, encompassing a broad spectrum of 62 parameters, including traffic dynamics, Vehicle Specific Power (VSP) metrics, engine states, and gear-shifting behavior. The primary objective was to discern parameters with the most significant impact on emissions. An initial factorial analysis was instrumental in refining these parameters to 16 pivotal determinants. A subsequent regression analysis was conducted, revealing critical assessment criteria: four factors associated with acceleration and VSP, three linked to gear-shift dynamics, and two aligned with speed metrics.

2.1.7 Candidate Driving Cycle Construction

The most intricate phase in driving cycle formulation involves linking microtrips, known as the candidate driving cycle construction. This cycle is preliminary and requires evaluation for its accuracy in mirroring real driving patterns. When creating this, microtrips are combined, with common methodologies being Random-selection, Clustering, and the Markov chain model. However, classification of these approaches varies among researchers. Nouri (2015), for instance, grouped them into "Stochastic methods" and "Markov chain model", while Czégé, Vámosi, and Kocsis (2020) delineated them into four categories: Random selection, Clustering, Statistical methods, and the Markov chain model.

Random selection method

The Random-selection method is frequently employed in the literature for driving cycle development. In this method, a microtrip starting with zero velocity is chosen at random, indicating a trip's inception. Subsequent microtrip selections rely on quasi-random selection, continuing until the concatenated microtrips reach the desired length of a candidate driving cycle. A quasi-random technique is a method of selection or sampling that combines elements of randomness with a

structured approach. It's not completely random, as it follows specific patterns or rules to ensure a more uniform distribution, but it also incorporates some degree of unpredictability, unlike purely deterministic methods. The quasi-random selection in driving cycle construction is an approach where microtrips are selected from a larger dataset to form representative driving cycles. A notable drawback, as highlighted by (Sithananthan and Kumar 2020; Achour and Olabi 2016), is the potential omission of some microtrips, possibly leading to certain driving conditions being unrepresented.

This technique is best suited for specific regions, particularly for vehicles like buses or garbage trucks with predictable routes, ensuring more accurate cycles.

Wang et al. (2016) addressed the Random selection method's limitations while developing driving cycles in Beijing. The inherent uncertainties of Random selection can lead to unreflective driving patterns. To rectify this, they:

1. Categorized microtrips using the adjacent stops technique and normalized varying dimensional assessment parameters before concatenating them.
2. Calculated the sum of parameters per microtrip.
3. Normalized the average parameter values across all segments, then computed the sum of these normalized values.
4. Measured the total duration for each category.
5. Determined the absolute deviations between steps one and three, arranging the segments by increasing deviation.
6. Sequentially concatenated segments, progressing from least to most deviation, until reaching the required duration.

Comparison of this enhanced approach with the traditional method, using real driving data, demonstrated that the refined method offered increased precision and acceptability.

Cluster-based method

The cluster method emerged as a solution to address issues found in the Random-selection technique. In this method, microtrips are categorized into different driving conditions (e.g., congested, free flow) and designated assessment criteria for each cluster. Mostly, microtrips are clustered based on their average speed and positive acceleration. However, when it comes to assembling the microtrips and creating candidate driving cycles, the process mirrors the Random-

selection method. Hence, researchers like Amirjamshidi (2015) argue that the cluster-based method should not be viewed as a distinct approach for driving cycle creation.

He (2020) harnessed the Mean Shift clustering method to enhance the precision of the cluster-based driving cycle formulation technique. Initially, real-world data was collected, producing 1701 kinematic segments or microtrips. Out of 12 defined assessment parameters, three primary determinants were pinpointed using PCA. For the first time, the Mean Shift clustering technique was applied to cluster microtrips for driving cycle creation. The achieved result was then juxtaposed against the traditional k-means clustering method's accuracy. k-means possesses inherent challenges that introduce uncertainty in driving cycle creation. Firstly, the number of clusters must be pre-defined, adding a subjective element. It is also susceptible to the initial center value; if set to a high value, the output can be misleading. Additionally, k-means fails to recognize and segregate noisy data, clustering them indiscriminately, leading to increased uncertainty.

Conversely, the Mean Shift method obviates the need to pre-specify cluster numbers and adeptly isolates noisy data, grouping them separately. This enhances the accuracy of the clustering results. When contrasting the traditional clustering technique's output with the new approach using real-world data, it was evident that the Mean Shift method's results had minimum deviation from real-world data, marking it as more precise.

Markov chain model

The Markov chain model is frequently referenced in scholarly literature for driving cycle construction. The approach encompasses two primary stages. Initially, the Markov chain process is employed to formulate a transition matrix that delineates the sequential probability of different operational states. The first microtrip is chosen randomly then using an iterative procedure where a microtrip is chosen at random. Based on the transition matrix, subsequent microtrip selections prioritize the ones with the highest likelihood. This selection process continues until the aggregated length of the microtrips matches the predefined duration of the candidate driving cycle.

Shi et al. (2016) conducted a theoretical investigation into the precision of the Markov chain model for driving cycle construction and prediction. They employed Wavelet theory and Dynamic analysis to elucidate the relationship among velocity, acceleration probability, and the transition matrix. The Wavelet theory breaks down a driving cycle (specifically its velocity profile) into multiple signals of varying frequencies, allowing for an analysis of velocity fluctuations.

Concurrently, the dynamic theory examines the interdependence between successive velocities. The study's conclusions highlight that the transition matrix proves to be a more suitable method for concatenating microtrips when compared to the velocity-acceleration distribution probability.

In a 2019 study, Huertas and colleagues crafted driving cycles for four distinct regions in Mexico: two urban locales, a mountainous region, and a composite area merging urban and mountainous terrains. Throughout the research, driving cycles were developed leveraging the Markov chains–Monte Carlo (MCMC) model (Silvas et al. 2016), a Microtrip-centric method, and a Fuel-focused technique. Subsequently, the precision of these three methodologies concerning fuel consumption and driving patterns was juxtaposed for each region. To facilitate this, the team observed fuel consumption and vehicle emissions using specialized sensors while cataloging trip activities via high-precision GPS. Adjacent stops were utilized to generate microtrips, with 23 parameters under consideration. Given the inherently stochastic nature of microtrips and the Markov chain model, the concatenation procedure was reiterated 1,000 times to achieve reliable results, with both average values and standard deviations documented. Prior to initiating the candidate driving cycle creation process, its accuracy was ascertained. This validation involved employing a synthesized trip as the input, subsequently comparing the resultant output with the original trip. Upon conducting tests across all designated regions, findings indicated superior accuracy of the Fuel-based driving cycle, both in terms of mirroring actual driving patterns and approximating vehicle emissions, outperforming the other two methodologies.

In a 2019 study, Ma and his colleagues employed advanced data mining methodologies to explore the disparities in driving cycles within Beijing, China. Recognizing the significance of collecting data across diverse road conditions and at varying times (such as peak and off-peak periods) for a credible driving cycle, the team gathered a comprehensive dataset spanning 3.3 million kilometers of travel over 17,000 sampling days. To methodically establish driving cycles, they factored in road categories (comprising freeways, arterial roads, and residential streets) along with specific travel time periods (distinguishing between peak and off-peak intervals). The data was then segmented based on operational patterns, utilizing the k-means clustering approach. This ensured that similar vehicular activities, like idling, accelerating, and decelerating, were grouped under a unified category. After this, the Markov chain model was applied to generate candidate driving cycles tailored for specific road types and driving periods. As a culmination of this process, two

emblematic driving cycles were crafted: the Off-peak cycle, which is indicative of high-speed driving, and the Peak cycle, reflecting driving scenarios characterized by frequent stops.

2.1.8 Selecting representative driving cycle(s)

The final phase of driving cycle development involves choosing the most representative driving cycle(s).

After generating candidate driving cycles, for each of them, a comparative analysis is conducted where the metric values derived from the cycle are compared against the predetermined target values of these assessment metrics. This comparison involves calculating the difference of each cycle's metrics from the target profile.

The core of the process is the ranking mechanism, where each driving cycle is evaluated based on the magnitude of its difference from the predefined assessment criteria. The cycle that exhibits the minimum difference, thus aligning most closely with the target profile, is determine as the most representative. This ranking determines the driving cycle that most accurately represent the real-world driving scenarios.

2.1.9 Summary of Driving cycle

Driving cycles are pivotal in estimating vehicular emissions. Broadly, they are categorized into legislative and non-legislative driving cycles. Legislative cycles are sanctioned by governments to regulate vehicle emissions at the national level, and their data predominantly stems from laboratory environments. In contrast, non-legislative cycles are primarily researcher-driven, operating at local scales, and are built from real-world data.

For the creation of a non-legislative driving cycle, certain stages are essential. The process initiates with data collection, typically acquired through on-board instrumentation or the chase car method, with both techniques leveraging GPS for data accumulation. Before data gathering starts, two crucial determinations are made: the selection of representative routes and the specification of assessment metrics or driving cycle parameters. Some studies target specific routes as representative of the entire network, while others encompass the whole network for data extraction. After data acquisition, preprocessing techniques refine the raw data for further analysis. Microtrips are then derived from the data using one of six approaches: adjacent stops, fixed distance, fixed time interval, event bins, patterns-based, and segment-based. These microtrips provide insights into

prevailing driving patterns. Following this, microtrips with similar driving characteristics, as gauged by the driving cycle parameters, are clustered together. Candidate driving cycles are then produced through various methodologies, such as the random selection method, the Markov chain method, and the cluster-based method. In the concluding stage, the most representative driving cycle(s) is chosen based on predefined assessment measures. Table 2-3 offers a comparative analysis of various driving cycles concerning construction techniques, assessment metrics, data collection methods, and other relevant aspects.

Table 2-3. Comparison of local driving cycles

Driving cycle	Country	DC Type	Route selection	Data collection	Assessment criteria	Cycle construction
Tehran	Iran	LDC	Entire urban network	AVL	1, 8	Cluster-based
Mashhad	Iran	LDC	Traffic Information, road function and type	GPS	1, 19, 16, 2, 8, 3, 20	Cluster-based
ECE WMTC,	Europe	SDC	In-use driving behaviour	OBD	1,18,19,14,7,8	Random selection
Edinburgh	UK	LDC	Home to work trips	GPS	12	TRAFIX method
Chennai (CADC)	Southern India	LDC	City arterial	OBD	1,18, 19, 16, 8, 20, 2	Random selection
Delhi	North India	LDC	Home to work trips	GPS	16, 19, 1, 18, 21, 8, 20,2,3	Random selection
Khon Kaen	Thailand	LDC	Local traffic	OBD	1, 19, 16, 15	Random selection
KHM	Southern Taiwan	LDC	four routes in downtown	Chase car	1, 19, 16, 15, 21	Random selection
Hanoi,	Vietnam	LDC	Entire network	OBD	20, 1, 19, 16, 1, 3, 15, 8, 20, 21	Random selection

Table 2-3. Comparison of local driving cycles

Driving cycle	**Country**	**DC Type**	**Route selection**	**Data collection**	**Assessment criteria**	**Cycle construction**
TMDC	Taipei	LDC	Entire network	Chase car	18, 21, 13, 22, 8	Random selection
Sydney	Australia	LDC	traffic density	Chase car	9, 1, 2, 4	Random selection
Melbourne	Australia	LDC	Land use, arterial Highway	Chase car	2, 3, 4, 5	Random selection
FTP 75	USA	SDC	Home to work trips	OBD	6, 1, 15, 17, 10	Selecting the best fit trip
Assessment parameters index	**1.** Average speed **2.** Root mean square acceleration **3.** Positive kinetic energy **4.** Percentage idle **5.** Speed/acceleration probability density function **6.** Stops per distance **7.** Average number of stops within a microtrip **8.** Proportions of idling, acceleration, and deceleration **9.** Speed/acceleration matrix **10.** SAPD **11**. SAFD **12.** Speed codes **13.** Average acceleration/deceleration **14.** Mean length of microtrips **15.** Maximum speed/acceleration**16**. Average deceleration **17**. Number of stops **18**. Average running speed **19.** Average acceleration **20**. Percentage of time spent in cruise mode **21**. Total length 22. Mean length					

2.2 Influencing Factors

Drivers and vehicles are intrinsically influenced by their external environment. Environmental conditions play a significant role in determining drivers' abilities, influencing vehicle dynamics, impacting road infrastructures, and modulating traffic flow. As an illustration, adverse weather conditions such as rain or snow can diminish visibility, leading drivers to adopt a more cautious driving approach (Akin, Sisiopiku, and Skabardonis 2011; Ericsson 2001b; Faria et al. 2020). Furthermore, transitioning from one road type to a steeper gradient can necessitate changes in drivers' gear-shifting strategies, consequently affecting their speed and overall driving pattern.

Table 2-4 enumerates the various factors influencing driving behavior. Notably, among these factors, weather conditions and street characteristics exert a more pronounced impact (Hjelkrem and Ryeng 2017). Additionally, data pertaining to these elements are more readily accessible compared to information on driver and vehicle attributes.

Based on road characteristics, driving data analysis have indicated that elements like road functions, road types, land use, speed limits, road width, junction density, and surface conditions can influence variations in driving behavior (Ericsson 2000).

Vehicle speed is pivotal in examining driving behaviors as well as fuel consumption. While it provides a robust representation of driving patterns, the distribution of speed also has a correlation with fuel consumption and the rate of vehicle emissions.

Furthermore, speed serves as a critical variable in various speed-centric emission models designed for estimating vehicle emissions. Typically, these models can be categorized as average speed-based models, instantaneous models, and power demand models.

In conclusion, vehicle speed stands out as a versatile variable instrumental in both emission models and analyses of driving behavior. Subsequent sections provide a comprehensive review of research focused on evaluating the influence of determinants on driving behaviors and vehicle emissions.

Table 2-4 Factors influencing driving behavior

Influencing factors	Grouping variables	Variables
Street features	Street function classifications	1. Freeway

		2. Arterials 3. Collector 4. Local Street
	Type of environment	1. Residential 2. Central Business District (CBD) 3. Industrial
	Road design	1. Slope 2. Speed limit 3. Curvatures 4. Roundabout 5. Shoulder 6. Signage 7. Width
	Density of intersection	1. No signalized intersection 2. Segments longer than 200 m between signalized intersection 3. Segments shorter than 100 m between signalized intersection
Weather factors	1. Temperature 2. Snow 3. Humidity 4. Fog 5. Wind speed 6. Water level 7. Rain	

2.2.1 Road features

Street features encompass various attributes that can significantly influence driving behavior. Broadly, these attributes can be categorized into four primary classes: road design, land use (or type of location), road function, and intersection density. Each of these classes carries unique

characteristics that can impact the way drivers navigate and interact with the road environment. The subsequent sections delve deeper into the specifics of these categories.

Road design

Road design encompasses several features that have the potential to influence driving behavior. These features range from structural elements like barriers (for instance, guardrails) to more aesthetic considerations like the presence of trees alongside the road. Other important aspects of road design include the availability of an emergency lane, the width of the road, its vertical alignment, particularly in terms of gradient, pavement type characteristics, and the presence of curves or bends. Out of these, the slope, curvature, and width of the road have been most frequently discussed and considered in previous studies.

However, scrutiny of existing literature reveals a common trend: a substantial proportion of research in this area has relied on simulation tools to carry out their investigations. While simulators offer controlled environments to study specific variables, they also come with limitations. A glaring oversight in many of these simulation-based studies is the disregard for crucial factors such as traffic flow, the gradient of the road, and its curvature. These are pivotal elements that can significantly sway driving behavior. Conducting studies in a simulated, isolated environment might inadvertently introduce discrepancies, creating a divide between actual driving conditions and those represented in the simulations.

Among the enumerated road design attributes, existing literature unanimously acknowledges that slope plays a paramount role in influencing vehicle speed. Curvatures too, albeit to a lesser degree compared to slope, have been shown to have a tangible impact on driving behavior. Drivers often find themselves compelled to adjust their speeds when navigating uphill or downhill terrains, as well as along straight paths or tangents. As highlighted by (Yue and Wang 2019), such terrain-based and curvature-induced speed modulations are not mere preferences but often necessitated by the demands of the road's design and the safety requirements they entail.

To address the aforementioned gap in understanding the impact of road gradients and curvatures on driving behavior, (Nama et al. 2016) conducted an in-depth study in India. They used the 85th percentile speed as a benchmark and employed statistical analyses to decipher the intricacies of driving behaviors under varying road conditions. Their findings revealed that in hilly regions, 45%

of drivers tend to exceed the designated speed limit. Furthermore, the study determined that the effect of curvature on vehicle speed is more pronounced in areas with a relatively flatter terrain.

Dinh and Kubota (2013) delved into the impact of carriageway width, among other variables, on driving speed. Using the 85th percentile speed as their benchmark for reference, they statistically analyzed the correlation between carriageway width variations and driving speed. In their analysis, they categorized carriageway width into three distinct variables: roadway width, left sidewalk, and right sidewalk. Their findings revealed that the width of the roadway can indeed shape driving behavior. Notably, the presence of sidewalks, both on the left and right, exhibited a strong positive correlation with increased driving speeds.

The impact of road surface on driving behavior is a nuanced and multifaceted issue, encompassing aspects like human response time and sensitivity, vehicle speed and comfort on rough roads, and vehicle handling and tire forces. In terms of human response time and sensitivity, a study by Ranjitkar and Nakatsuji compared driving behavior under icy and dry road conditions. This research focused on perception response time, sensitivity factor, and stability factor, revealing that response time is significantly higher in icy conditions compared to dry roads (Ranjitkar and Nakatsuji 2013). Regarding vehicle speed and comfort on rough roads, Nguyen, Nguyen, and Hoa Tran (2020) explored the effect of road surface roughness, as measured by the International Roughness Index (IRI), on recommended vehicle speed. This study suggested that road sections of specific lengths with certain IRI ranges can enhance driving comfort. Abeygunawardhana, Sandamal, and Pasindu (2020) explored impact of road roughness on speed patterns for different roadway segments. They used smartphone-based roughness measurements. These measurements were found to be reliable and comparable to precise International Roughness Index (IRI) assessments. The research revealed that road roughness significantly impacts vehicle speeds, especially affecting passenger cars more than heavy vehicles. It also highlighted the correlation between road roughness and varying speed-flow patterns, offering critical insights for pavement management systems and decisions related to road maintenance and safety.

Type of land use (built environment)

The surrounding environment of a roadway, specifically the type of land use or neighborhood, plays a role in influencing driving behavior. In essence, the specific land use dictates the nature of the landscape alongside a road. Studies from the fields of psychology and safety have substantiated

that such landscapes can alter parameters like average driving speed and the lateral position of a vehicle on the road. Aditjandra (2013) reported that driving speeds decreased by 5% in areas with shopping facilities and by 10% in residential zones. However, the influence of certain landscape features on driving speed has yielded inconsistent findings across studies. For instance, Abele and Møller (2011) found that the presence of trees alongside a road tends to reduce driving speeds. Contrarily, Van Der Horst and De Ridder (2007) found no significant impact of roadside trees on driving speeds.

Antonson et al. (2009) delved into the relationship between changes in driving behaviors and three distinct landscape types: forested, open, and varied terrains. Additionally, the study aimed to discern the influence of these landscapes on drivers' perceptions of their driving contexts. To conduct this investigation, the landscapes were replicated in the Driving Simulator, with eighteen participants driving a 35 km distance within this simulated environment. The findings highlighted a notable disparity in average speeds across the different landscape types, with drivers exhibiting higher speeds in the open areas.

Road function

Various road types, such as arterial roads, freeways, collectors, and local streets, play distinct roles in influencing driving behavior. Previous research indicates that driving behaviors are inherently linked to the function of the road. These road categories differ in aspects like speed limits, access, the presence of intersections, and other infrastructural features. Consequently, these differences culminate in diverse driving behaviors, which in turn result in variations in driving cycles.

Pilgerstorfer (2011) employed the Naturalistic Driving method to study 4,551 trips, aiming to discern the variations in driving behaviors across diverse roadway categories such as collector, arterial, regional, inter-urban, and highway roads. The Naturalistic Driving approach involves the simultaneous utilization of a video recorder and a sensor (typically a GPS) to continuously capture the driver's actions, driving conditions, and the vehicle's position. The outcomes revealed noticeable differences in driving behaviors depending on the road category: highways witnessed high speeds (exceeding 90 km/h), whereas arterial and collector roads predominantly saw speeds below 40 km/h. However, this study's methodology had certain drawbacks. Firstly, the Naturalistic Driving Data produced a huge volume of data, primarily due to the video recordings. More critically, the results were presented at an aggregate level; the study used average speeds to depict

the speed levels associated with each road category, which does not truly capture the variability of speed over time.

Damsere-Derry et al. (2008) took a different approach compared to the previously mentioned study. They introduced two metrics: mean and dispersion, to delve into not just the average driving speed, but also the variability in speed values. This investigation was conducted on regional, inter-regional, and national roads in two distinct regions – settlement and rural areas in Ghana. The dispersion metric denoted the standard deviation derived from the recorded speeds. Upon statistically analyzing the data, it was evident that the national roads in both areas witnessed the highest average speed and variability, with values around 85 ± 18 km/h. On the contrary, the inter-regional roads in both regions recorded the lowest figures, averaging at 64.7±12.3 km/h.

Density of intersection

Intersection density refers to the number of intersections within a given area. In a study conducted by Ericsson (2001), an analysis of various factors influencing a driving cycle was carried out. The results of this study highlighted that the density of intersection had a more pronounced impact compared to other factors. Generally, as the density of intersections increases, there is a corresponding decrease in both the mean speed and speed deviation.

2.2.2 Weather Parameters

Weather conditions can have a profound effect on driving behaviors and traffic flow. Multiple studies have demonstrated that adverse weather can lead to significant changes in traffic dynamics and reduce the capacity of roads and highways.

For instance, severe rainfall can reduce road capacity by 14% to 19%, as pointed out by (Christian and Jensen 2014; Chung et al. 2006). Dense fog poses another challenge as it significantly diminishes visibility and the level of service (LOS) provided by a roadway. Consequently, drivers often reduce the distance they maintain from the vehicle in front, leading them to adopt a more defensive driving style, as highlighted by (Jiang et al. 2020).

However, interestingly, even in conditions with extremely low friction, such as very slippery surfaces, drivers tend to exceed speed limits. (Kilpeläinen and Summala 2007) observed that many drivers still travel faster than recommended even when road friction is below 20%.

Extreme temperature conditions, either too hot or cold, can also influence driving behavior. Such temperatures can impact a driver's psychological state, leading to increased chances of aggressive driving behavior, as noted by (Morris 2015).

Moreover, the Highway Capacity Manual (HCM 2010) states that light and heavy snowfalls can decrease Free Flow Speed (FFS) by substantial margins. Specifically, light snow reduces FFS by 8-10%, while heavy snow leads to reductions of 30% and 40%, respectively.

In essence, understanding the implications of various weather conditions on driving behaviors are essential for traffic management, road safety, and infrastructure planning.

Literature review indicates that while many studies have explored the effects of adverse weather conditions on safety, level of service (LOS) of highways, and overall traffic flow, fewer have delved into the relationship between weather factors and individual driving behaviors. Most research to date has largely focused on the macroscopic impacts of weather on transportation. For example, studies might quantify the increase in accident rates during heavy rainfall or the decline in freeway capacity during snowfall. However, the microscopic impacts of weather on individual driver behaviors remain less explored. These could encompass studies on how fog affects individual braking distances, how wet roads alter turning behaviors, or how glare from sun or snow influences lane-changing decisions.

Various studies pointed out this gap in the literature, noting the limited research specifically focusing on the influence of weather factors on individual driving behaviors. While macroscopic studies are certainly valuable, gaining a microscopic understanding is crucial for several reasons:

1. Driver Training and Awareness: Insights into how individual drivers react to specific weather conditions can inform driver training programs, making them more tailored and effective.
2. Infrastructure Design: Understanding microscopic behaviors can help in designing roads, intersections, and other infrastructures that account for these behaviors during adverse weather conditions.
3. Advanced Driver Assistance Systems (ADAS): With the advent of smarter cars, understanding individual driving behaviors under different weather conditions can lead to better ADAS functionalities that can warn or assist drivers accordingly.

A potential challenge facing researchers in this area is the complexity and variability of individual driving behaviors. Two drivers might react very differently to the same weather condition based on their experience, confidence, and the vehicle they are driving.

Furthermore, it is worth noting that many of the studies in this area resort to qualitative expressions due to the multifaceted nature of the topic. Quantifying the impact of weather on driving behaviors requires considering a myriad of factors, and often, quantitative observations can offer richer insights than strictly qualitative ones.

Hoogendoorn et al. (2010) provided a notable exploration into the effects of fog on driving behaviors using a driving simulator. Then, these impacts were simulated on traffic flow using micro-simulation traffic model. The study found substantial reductions in speed and acceleration during foggy conditions as compared to sunny weather, reflecting the typical caution drivers exercise under reduced visibility. However, the change in mean deceleration was not statistically significant, suggesting that braking behaviors might not be as greatly influenced by fog as other driving parameters.

The study by Min (2015), set in Ontario, Canada, offers a comprehensive look into how road maintenance and winter weather conditions influence vehicle speed distribution. This, in turn, impacts fuel consumption and emission rates. Using a multiple regression model, Min was able to quantify the relationships between these factors. By integrating speed models with the MOVES (Motor Vehicle Emission Simulator) model, the research provided a deeper insight into how variations in speed under different conditions directly influence emission outputs.

One of the significant findings from this study is the revelation that a mere 10% improvement in road surface conditions can lead to a potential 2% reduction in emissions. This emphasizes the crucial environmental and public health benefits that can stem from consistent road maintenance, especially in areas prone to severe winter conditions.

From an economic perspective, this potential decrease in emissions suggests that improved road maintenance might lead to tangible cost savings for both drivers, in terms of reduced fuel consumption, and the broader society, which would benefit from reduced healthcare costs linked to pollution-induced health issues.

Looking forward, there is room for more detailed exploration in this field. It would be insightful to study how varying levels of maintenance or the severity of winter conditions impact driving behavior. A comparison between different types of vehicles, such as cars versus trucks or electric versus petrol vehicles, would also be intriguing.

However, while the study is impactful, it is essential to acknowledge its limitations. External factors like traffic density, adaptations in driver behavior to changing road conditions, or the distribution of vehicle types were not extensively considered.

Chowdhury (2015) research delved into the intricate connection between temperature and driving performance. The central premise of the study was to isolate temperature as the primary variable while maintaining consistency in other potential influencing factors, such as road conditions, driver age, and other meteorological variables. To achieve this, the research focused on a temperature range of 10.5 to 26.6 °C.

To provide a more structured understanding of the temperature's effects, Chowdhury grouped the studied range into three distinct categories based on the "comfort temperature," a concept that signifies the temperature at which a driver is most at ease:

- Low Temperature Range: 10.5 - 15.5 °C
- Medium Temperature Range: 15.5 - 21.1 °C
- High Temperature Range: 21.1 - 26.6 °C

By using this classification, the study aimed to discern if, and how, deviations from the "comfort temperature" influence driving behavior. The hypothesis likely revolved around the idea that temperatures either below or above the comfort range could lead to variations in driver alertness, reaction time, or overall driving capability. Given that extreme temperatures, both hot and cold, have been observed to affect human cognitive and physical performance in other fields of study, it is logical to assume that driving, a task that demands both mental and physical attention, would also be influenced.

In a comprehensive study of various weather factors and their impact on driving behaviors on an urban freeway in Minnesota, Agarwal, Maze, and Souleyrette (2005) endeavored to quantify the effects of elements like wind intensity, temperature, rain, visibility, and snow on driving speed. To make their results more interpretable, they chose to categorize the intensity of these weather factors

into three to five distinct classes and subsequently calculated the average speed associated with each class. Table 2-5 compares the classification of weather parameters in different studies.

Table 2-5 Comparison Influences of weather parameters on vehicle speed

Variable	**Intensity range**	**Corresponding range in HCM 2000**	**Average Speeds % reduction**	
			HCM 2000	**Agarwal**
Rain	0-.2m54 cm/h	Light	2-14	1-2.5
	.254 to 0.635 cm/h	Light	2-14	2-5
	>0.635 cmt/h	Heavy	5-17	4-7
Snow	<=1.27 cm/h	Light	8-10	3-5
	0.152 to .254 cm/h	Light	8-10	7-9
	.279 to 1.27 cm/h	Light	8-10	8-10
	>1.27 cm/h	Heavy	30-40	11-15
Temperature	10-1 °C	-	NA	1-1.5
	0-(-20) °C	-	NA	1-2
	<-20 °C	-	NA	0-3.6
Wind speed	16-32 km/h	-	NA	1
	>32 km/h	-	NA	1-1.5
visibility	1.6 to .82 km	-	NA	6
	.8 to .4 km	-	NA	7
	<.4 km	-	NA	11

From their findings, wind speed and temperature seemingly had a negligible impact on driving speed. This was an interesting observation as one might intuitively feel that factors like strong gusts

of wind or extreme temperatures might significantly alter a driver's comfort or vehicle handling. However, their research painted a different picture.

On the other hand, visibility emerged as a significant factor, causing a decrease in driving speed ranging between 6 and 10 percent. This finding resonates with the intuitive understanding that reduced visibility conditions, whether due to fog, rain, or other factors, would necessitate more cautious driving.

When it came to the impact of snow, their findings exhibited some differences with the Highway Capacity Manual (HCM 2000). While both studies concurred on the effects of light snow on reducing driving speeds, they diverged on the topic of heavy snow. Specifically, Agarwal, Maze, and Souleyrette's study indicated that the speed reduction due to heavy snow was more profound than what the (HCM 2000) proposed.

To provide a comprehensive understanding of their findings, a comparative table (Table 2-5) was used to juxtapose the results from Agarwal, Maze, and Souleyrette (2005) with those from HCM 2000. This side-by-side comparison allowed for a clearer picture of where the two studies aligned and where they differed, offering invaluable insights for future research.

In the study by Zhao, Sadek, and Fuglewicz (2012), they established a regression model to ascertain the correlation between meteorological parameters and variations in average traffic speed. Their findings were as follows:

- A direct relationship exists between visibility and speed: a 1 km increase in visibility corresponds to a 2 km/hr uptick in traffic speed.
- There is an inverse correlation between weather type and speed, though specifics of this correlation were not elucidated upon.
- When temperatures rise above the freezing mark, there is a noticeable increase in speed, approximately 1.58 km/hr higher compared to sub-zero conditions.
- Wind speed presents a negative correlation with traffic speed. A surge in wind speed by 10 km/h leads to a decrease in traffic speed by approximately 1.3 km/h.

The study's outcomes suggest that among the meteorological factors analyzed, high temperatures exert the most significant influence on driving patterns. This underscores the heterogeneity of driving patterns across regions, contingent on prevailing weather conditions. Among weather

factors cumulative precipitation and visibility have the biggest negative and positive impacts respectively.

2.2.3 Type of day (weekend and weekday)

Driving behavior tends to vary between weekdays and weekends. This variation is often attributed to differences in traffic patterns, purposes of travel, and the overall pace of life on these days. On weekdays, driving behavior is frequently influenced by commuter traffic, leading to busier roads during peak hours. Conversely, weekend driving might be more recreational or involve running errands, leading to different traffic patterns and potentially less congested roads during traditional rush hours. These differences can impact factors such as average speed, the likelihood of congestion, and overall driving stress levels.

(Heydari et al., 2019) investigated the likelihood of speeding on weekend nights compared to other times, specifically analyzing speeding trends during weekends and weekdays. The study used speed data collected automatically from traffic analyzer sensors on local streets in Montreal. Contrary to expectations, the findings revealed that speeding on weekends is less common during night hours and more prevalent during evening and midday hours. Additionally, the study observed that one-way streets and roads with a speed limit of 50 km/h witnessed slightly less speeding during weekends. These insights are valuable for designing targeted road safety interventions, such as publicity campaigns and police enforcement strategies, aimed at reducing speeding behaviors. The study's findings contribute to a deeper understanding of speeding patterns and can inform more effective approaches to enhancing road safety.

Ellison & Greaves, (2010) analyzed speeding behavior among 133 motorists in Sydney, using GPS technology. It reveals that 20% of the total distance traveled by these motorists was at speeds exceeding the posted limits. The study observed that speeding tends to occur more frequently in both high-speed zones (100-110 km/h) and low-speed zones (40-50 km/h). An interesting pattern emerged showing that speeding is more prevalent during weekday mornings and on weekend nights.

Huang et al., (2022) examined the speed and clearance of commuter buses in urban areas, with a focus on comparing data between weekdays and weekends, under varying weather conditions including sunny and heavy snow. Data was gathered from thirty participants driving across ten routes in natural conditions, encompassing a range of traffic flows from free to unsteady. The

findings highlight that on weekdays, commuter buses tend to travel at significantly lower speeds compared to weekends. Specifically, there was an observed reduction in speed ranging from 11.2% to 16.5% on weekdays. This difference becomes even more pronounced during heavy snow conditions, where the speed decrease on weekdays varied from 33.8% to 40.7% compared to weekends. The study also notes that the lowest average speeds for commuter buses were recorded in the most congested traffic situations, identified as Grade IV. This reduction in speed was more acute on weekdays. Furthermore, it was observed that on weekdays, especially under heavy snow or in heavy traffic conditions, commuter buses maintained a minimum distance between vehicles. This research underscores the substantial impact of weekdays and weekends on the dynamics of urban commuter bus travel, emphasizing the need for tailored urban planning and transportation policies to accommodate these differences effectively.

Ellison & Greaves, (2010) studied whether motorcyclists behave differently during workdays compared to weekends, facing challenges in making this comparison. The observation locations varied between workdays and weekends, and there was a limited number of observations on 30 and 90 km/h double lane roads during weekends. Additionally, no school zone with a permanent 30 km/h limit was surveyed on weekends, and only one site on a 90 km/h double lane road was visited. Due to these limitations, certain data combinations were omitted. The study focused on observations at free speed to minimize the impact of congestion on results. It was found that, on average, motorcyclists drove 3 to 8 km/h faster during weekends than on workdays. This difference was even more pronounced on highways, where the average speed was almost 9 km/h higher on weekends. These findings were statistically significant for 50 km/h roads and highways.

2.2.4 Summary of Influencing Factors

Driving behaviors is shaped by an interplay of various factors including meteorological conditions, roadway attributes, vehicular characteristics, traffic, and individual driver traits. While every component is crucial in its own right, literature reveals a heightened emphasis on weather and road characteristics. This is primarily attributed to the wide accessibility of relevant data and their pronounced impact on the transportation ecosystem.

1. Weather Conditions: Key meteorological parameters affecting driving behaviors include snowfall, rainfall, fog, ambient temperature, and visibility. A scan of relevant literature reveals that much of the research pivoting around the impact of weather conditions on

transportation revolves around safety and Level of Service (LOS) concerns. A smaller subset, however, delves specifically into the direct influence of these weather parameters on driving behaviors.

2. Road Characteristics: Fundamental road attributes that modulate driving behaviors encompass road design, location type, functional classification, and intersection density. Road design encapsulates physical facets like horizontal and vertical alignments, roadway width, prescribed speed limits, and the presence of road shoulders. Intersection density, on the other hand, offers a metric representing the frequency of intersections within a given area. Functional classification systematically categorizes roads by their intended role in overall traffic flow and connectivity, ranging from high-capacity urban arterials to local roads primarily serving neighborhood traffic.
3. Type of day: driving behavior differs between weekdays and weekends. Weekday driving is influenced by work-related commuting, leading to busy roads and rush hour congestion. Weekend driving, on the other hand, tends to be more leisurely and errand-based, resulting in different traffic patterns and potentially less congestion during usual peak times.

Table 2-6 compares the key studies that have delved into these influencing factors against a research methodology delineated in the 3.2section 2.3. This comparative framework aims to provide a holistic overview, mapping historical insights to contemporary research approaches.

Table 2-6 List of constructed driving cycles considering influencing features

Authors	Techniques	Influencing Factors	Objective(s)
(Ragione and Giovanni 2016)	1. Portable Emissions Measurement System is utilized to gather data from two streets 2. PCA is used to cluster variables into 4 groups 3. Multivariate regression to build KEM model	Road gradient (divide grades to uphill, downhill, and flat)	To Develop second version of Kinematic Emission Model (KEM)
(B. Yue et al. 2015)	1. DEWE-50 and CAN-Bus is utilized to collect data. 2. Markov chain model is employed to construct driving cycle	Road slope	To Analyze the capability of Markov Chain Model in considering road grade in designing driving cycle
(Tzirakis and Zannikos 2013)	1. Collecting driving data using instrumented vehicle 2. Adding impact of grade to mean value of acceleration using equation below αadd=g (h2-h1)/d 3. where g is acceleration of gravity, h is elevation and d is distance.	Road gradient	To investigate the impacts of road grade on driving cycle

Table 2-6 List of constructed driving cycles considering influencing features

Current thesis	1. Classifying road type using machine learning methods 2. Classifying weather type using machine learning methods 3. Developing and identifying driving cycles for different weather condition 4. Determining optimal number using information loss and variability indicators	1. Road characteristics 2. Weather parameters	To Develop a methodology to determine optimal number of driving cycles.

2.3 Classification of influencing factors

In this section, we examine literature that delves into the categorization of factors influencing driving behavior. Classification essentially involves grouping objects or concepts based on their shared traits, facilitating a clearer understanding of data for researchers and professionals alike. While various factors were detailed in2.2 section 2.2, it was previously highlighted that weather conditions and road characteristics stand out due to their pronounced impact on driving patterns.

2.3.1 Classification of Road Types

The endeavor to capture driving patterns across an entire urban network can certainly enhance the precision of driving cycle outputs. However, a handful of driving cycles might not encompass the vast spectrum of driving behaviors due to the diverse range of road types, each bearing unique characteristics. In past studies, despite the multiplicity of road types, only a finite number of driving cycles were formulated. A remedy to this limitation involves categorizing the myriad road types and then ascertaining the ideal number of driving cycles for each category. This ensures that the chosen driving cycles collectively provide an accurate representation of the region's driving behavior.

By adopting this methodology, roads with analogous characteristics are clustered into singular categories. This not only offers researchers a deeper insight into the complexities of a road network but also streamlines the process of determining the necessary number of driving cycles for urban settings.

Classification of road types

The intention behind the categorization process can lead to various classifications of an urban road network. For instance, Gorges, Öztürk, and Liebich (2018) aimed to derive comprehensive insights into urban networks to further refine the design of two-wheeled vehicles. They achieved this by creating a tri-layered system: a curve estimator, a slope estimator, and a road profile estimator. The curve estimator gauged factors like road curvature, elevation changes, and unevenness. In contrast, the slope estimator computed the gradient of road segments, and the road profile estimator focused on classifying distinctive road profiles.

Vitkienė, Puodžiukas, and Žilionienė (2017) embarked on a comparative study of road classification systems across seven nations, namely Germany, Austria, Poland, Denmark, Greece,

Norway, Portugal, USA, and Australia. Their methodology involved utilizing eight distinct criteria for this comparison, including aspects like access control, usage parameters, road surface characteristics, administrative control, the significance of surrounding land use, geographical setting (place status), modes of transportation, and primary functionalities. As depicted in Table 2-7, the study highlighted the absence of a universally adopted approach to road system classification across nations.

Lee and Öberg (2015) conducted a study to understand the correlation between road categories and driving behaviors in Sweden, with an aim to enhance fuel efficiency in fleet management. The categorization of roads was based on the velocity of the vehicles. Using this method, vehicles traveling at speeds of 0-50 km/h, 50-70 km/h, and above 70 km/h were indicative of urban roads, rural roads, and highways, respectively. Furthermore, driving behaviors were categorized into calm, normal, and aggressive based on acceleration and jerk patterns. A driving pattern is deemed calm if the acceleration frequency is below the 1st quartile, aggressive if it is beyond the 3rd quartile, and normal if it falls between the 1st and 3rd quartiles. Acceleration denotes the change in speed over time, while jerk represents the rate of change in acceleration over time. Data was collected from 93 journeys using On-board Diagnostic (OBD) systems. Based on the recorded speeds, the roads were categorized, and then acceleration variations for each road type were analyzed. Figure 2-2 illustrates the differences in fuel usage across various road types.

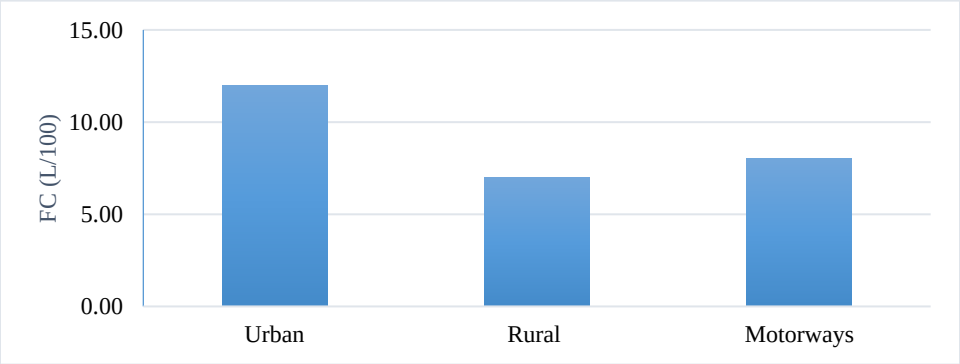


Figure 2-2 Changes of fuel consumption in various road types

Table 2-7 Comparative analysis of criteria utilized to classify road system in different countries

	Road classification criteria							
	Access control	**Road surface**	**usage**	**Administration**	**Link role**	**Place status**	**Transport mode**	**Function**
Austria				*	*			*
Germany		*			*	*		*
Poland		*		*	*			*
Denmark			*		*	*		*
Greece	*		*		*			*
Norway			*		*			*
Portugal	*		*		*		*	*
USA	*		*		*	*		*

2.3.2 Classification of weather parameters

In prior discussions, it was highlighted that weather conditions can significantly influence driving behaviors. Yet, upon a thorough review of the existing literature, it becomes evident that no studies have been undertaken to categorize specific weather parameters for a more in-depth understanding of their impact on driving cycles.

2.4 Clustering

Clustering, a form of unsupervised learning, is a descriptive analytical approach that groups data into distinct categories, known as clusters, based on their similarities. The objective is to ensure that data within each cluster is as similar as possible, while also being distinct from data in other clusters. Unlike supervised learning techniques or regression models, clustering does not rely on labeled data. Instead, in supervised methods such as classification or regression, algorithms are trained on labeled data, whereas in clustering, the absence of labels necessitates the algorithm to identify patterns and categorize data based on inherent similarities.

Unsupervised learning can be divided into several types, such as partitioning, hierarchical, and Affinity Propagation techniques. Given that the first two methods, partitioning and hierarchical, are elaborated upon in the following two sections this discussion will focus on the latter one.

Affinity Propagation (AP) is a clustering algorithm that identifies clusters by considering all data points as potential exemplars through "message passing." It automatically determines the number of clusters based on the data, offering a distinct advantage over many clustering methods. Using a similarity matrix, AP iteratively exchanges messages between data points to refine exemplar decisions. While AP's ability to handle various similarity data and produce robust cluster assignments is commendable, it has challenges. It can be computationally intensive for large datasets and is sensitive to its preference parameter, which influences cluster numbers, necessitating fine-tuning for optimal results.

2.4.1 Partitioning method

Partitioning techniques, in the domain of clustering, work by dividing data into distinct clusters based on similarity measures. Their primary characteristic is the necessity to specify the number of clusters in advance. Among these techniques, the k-means algorithm stands out as the most prevalent. The following sections provide insight into their important features and pros and cons.

K-means algorithm

The k-means algorithm is a key machine learning technique for clustering, where the user defines the number of clusters (k). It involves selecting initial centroids, assigning data points to the nearest centroid based on distance (like Euclidean, Manhattan, or Minkowski), and iteratively updating centroids to minimize intra-cluster variance until a stopping criterion is met. The choice of distance metric is crucial, as it influences clustering accuracy and efficiency.

Understanding the data, including interrelationships, similarities, and distribution characteristics, is vital for selecting an appropriate distance function. K-means typically uses Euclidean, Manhattan, Chebyshev, or Minkowski distance metrics, each influencing clustering outcomes differently.

Studies like those of Singh, Yadav, and Rana (2013), and Kumar, Chhabra, and Kumar (2014) have shown that the effectiveness of a distance measure varies with the nature of data and the clustering approach. K-means is preferred for its simplicity, scalability, and computational efficiency, making it suitable for large data sets and distributed processing.

However, k-means has limitations, including the subjective choice of 'k', potential inconsistencies due to random initial centroid selection, sensitivity to data noise, and difficulties with non-linear and categorical datasets. These factors can affect the final accuracy of the clustering results.

2.4.2 Hierarchical Clustering

Hierarchical clustering techniques, unlike partitioning methods, do not require a predetermined number of clusters. They evolve by merging or splitting data points, eliminating the need for predefined cluster numbers. These methods are mainly Agglomerative Hierarchical Clustering (AHC) and Divisive Hierarchical Clustering (DHC).

DHC starts with all data in one cluster and splits it into smaller clusters recursively, while AHC begins with each data point as a separate cluster, merging them based on similarity criteria. The dendrogram is a common visual tool representing the hierarchical structure of clusters.

In AHC, clusters are merged based on a proximity matrix using a specified distance function. As the process iterates, clusters gradually combine into a single comprehensive cluster. Various distance metrics, like Single-linkage (minimal distance), Complete-linkage (maximal distance), and Average-linkage (average distance), influence the resulting hierarchy.

Single linkage focuses on the shortest inter-cluster distance and can result in elongated clusters. Complete linkage uses the greatest distance, leading to less cohesive clusters sensitive to outliers. Average-linkage, combining aspects of the other two, results in more compact, spherical clusters less affected by outliers.

Overall, hierarchical clustering offers diverse approaches to group data, each with unique implications for cluster formation and representation.

To visually illustrate, consider Figure 2-3 depicting a hypothetical dataset with six distinct objects. Figure 2-3 then showcases a dendrogram, a tree diagram frequently used to represent the hierarchical structure produced by the algorithm, indicating the gradual fusion or division of clusters as one moves up or down the hierarchy.

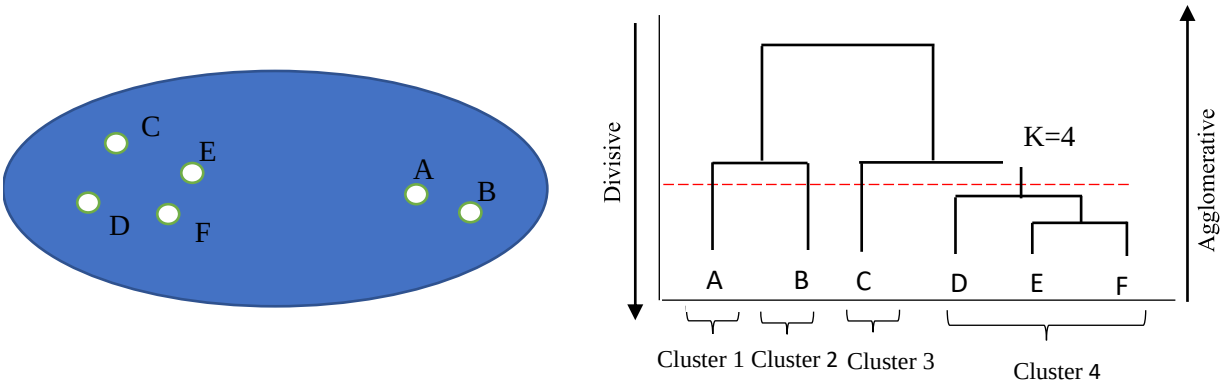


Figure 2-3 Hierarchical clustering structure, Image by author, 2020

2.4.3 Cluster Validation

Cluster validation is an indispensable step to gauge the effectiveness of clustering results. Essentially, cluster validation techniques can be broadly classified into internal and external cluster validation methods, as delineated by (Liu et al. 2010). The section 2.3 leans heavily towards the internal cluster validation, while this segment offers a cursory glance at external validation, with a more in-depth exploration of the internal methodology in subsequent sections.

External cluster validation gauges the quality of clustering outcomes, much like its internal counterpart. The key distinction lies in the benchmark it uses while external validation leverages external information such as predefined class labels to compare with results obtained post-clustering, internal validation solely relies on inherent cluster data to evaluate outcomes.

Consequently, external validation offers a comparative perspective, juxtaposing clustering outcomes against prior knowledge or ground truth.

Another nuance that differentiates the two is their applicability. Given that the cluster count in external validation is pre-established, it primarily serves as a tool to cherry-pick the most suitable clustering algorithm. On the other hand, internal clustering validation, with its intrinsic focus, aids in both identifying the most apt clustering methodology and discerning the best compromise cluster count.

2.5 Mixed Models

Mixed models are advanced statistical techniques designed to analyze data that possess a nested or hierarchical structure. These models are particularly useful when dealing with data where observations might not be independent, a scenario commonly encountered in longitudinal studies, repeated measurements, or clustered designs. Traditional regression models are often unsuitable for such data, as they fail to account for this inherent structure, potentially leading to biased or misleading results.

In a mixed model, there are two main components: fixed effects and random effects. Fixed effects, analogous to the coefficients in standard regression models, capture the average relationship between predictors and the outcome for the entire population. They describe the overall patterns and are assumed to be consistent across all levels or groups.

Random effects, on the other hand, allow the model to consider variations across different levels or groups in the hierarchy. For instance, if studying students' scores across various schools, random effects can capture the variability in average scores across these schools, even if the general trend (fixed effect) indicates a particular relationship between age and score. By modeling this extra layer of variability, mixed models provide a more robust framework that can account for both intra-group correlation and unexplained variability.

Furthermore, the term "mixed effects" typically indicates that the model incorporates both types of effects: fixed, which are constant, and random, which can vary. The versatility and robustness of mixed models make them indispensable in fields ranging from social sciences to medicine, wherever complex datasets require nuanced interpretation.

2.6 Main Gaps

The main gaps that this thesis covers can be summarized as follows:

Lack of a methodological approach for determining "optimal" number of driving cycles:: As discussed, driving cycles should be representative of driving behaviors to have an accurate emissions estimation. In the existing literature, the volatility of driving cycle parameters in different driving conditions and in various road segments is neglected, and mostly one or a few driving cycles are considered to represent driving patterns in an urban area in all weather conditions. This approach leads to a discrepancy between real driving behaviors and the generated driving cycles, which can be one of the main sources of uncertainty in emissions estimations.

Gap in determining "optimal" data aggregation level in driving cycle production: Previous studies aggregated all driving data collected in different road and weather conditions to construct the driving cycles. This approach maximizes information loss because all data are combined and only one driving cycle is produced. Thus, maximum data loss can maximize the discrepancy between real driving behaviors and selected representative driving cycles. On the other hand, generating driving cycles for all trips in different road and weather conditions surges computation complexity, and the potential for redundancy and inconsistencies. To solve this problem, it is necessary to invent an approach to determine the best compromise level of driving data aggregation.

Insufficient insights on the impact of road features and weather conditions on driving cycle parameters: Although the external factors impact driving cycle parameters, there are still many questions about the associations between road and weather features and driving cycle parameters, and the combination effects of these factors have remained unanswered. Answering these questions provides actionable insights for decision-makers and drivers.

Absence of a methodology for identifying road typology influencing driving behaviors: In the literature, road segments are mostly sorted into different groups according to the type of their functions, for example, highway, residential, freeway, and so on. This categorization does not represent how each road segment can impact driving behavior. To cover this gap, this thesis develops a new machine learning approach to identify and group road segments according to features influencing driving behaviors.

Need for a methodology to accurately identify weather types: Hourly weather factors can vary over a day, month, and year. In addition, weather conditions can be characterized by several factors such as rain, wind speed, humidity, and snow which can impact driving behaviors. To cope with this variability, previous studies broke down continuous values of weather factors into groups like heavy rain and light rain, nevertheless, it was not explained how a threshold is identified to group weather conditions. In this thesis, a methodological framework is developed to quantify the variability of hourly weather data.

CHAPTER 3 RESEARCH METHODOLOGY

Figure 3-1 depicts the conceptual framework of this thesis. A brief overview of its stages is provided below:

In the initial phase, a four-step procedure is employed:

1. Initially, a comprehensive review of the literature pertaining to driving cycle construction methodologies, the influence of external factors on driving behavior, and clustering analysis is conducted.
2. Subsequently, data concerning meteorological conditions, roadway characteristics, driving patterns, fleet compositions, and fuel types are gathered.
3. The final step of this phase is marked by data preprocessing, preparing the data for the next stage.
4. It is imperative to preprocess the data to ensure accurate results in the succeeding stages. Therefore, various techniques are employed in Python to prepare the road, weather, and driving data.

The second phase focuses on creating the weather and road typologies necessary for the third phase. The accuracy of this phase is pivotal for the final model results. The progression of this phase is as follows:

1. A variety of unsupervised machine learning techniques are employed to produce different road typologies. The most accurate approach is then determined using clustering validation techniques and driving speed data, as highlighted in Figure 3-1. This step marks the first significant contribution of this thesis, with a paper being published on this topic.
2. Clustering techniques are applied to hourly weather data to develop weather typologies. The weather typology is then chosen using clustering validation techniques and by visualizing the clustering output.

In the subsequent phase, driving cycles are created in accordance with the road and weather types developed in the previous phase, using code by Roy and Morency (2020).

The third phase revolves around the formulation of the final model and encompasses:

1. All driving data corresponding to a specific road type and weather type are chosen and amalgamated to generate a driving cycle. An application developed by Roy and Morency

(2020) is used in this thesis to construct the driving cycle. It should be noted that the generation of the driving cycle is iterative, continuing until driving cycles for all road and weather types are created.

2. Subsequently, two indicators measuring information loss at various data integration levels are introduced. It is vital to construct a driving cycle with the least possible information loss.
3. A sensitivity analysis is then conducted to comprehend the relationship between driving cycle parameters and external determinants.
4. A similarity matrix is then formulated to gauge the similarity between the constructed driving cycles.
5. In the concluding step, “optimal” finite number of driving cycles is determined. To do that, a tradeoff is made between the number of driving cycles, information loss and variability of driving cycles. To achieve it, the driving cycles with the highest similarity are combined. The information loss and variability of the remaining driving cycles are then assessed. This process is iterative, continuing until the minimum number of driving cycles, highest variability and lowest information loss are achieved.

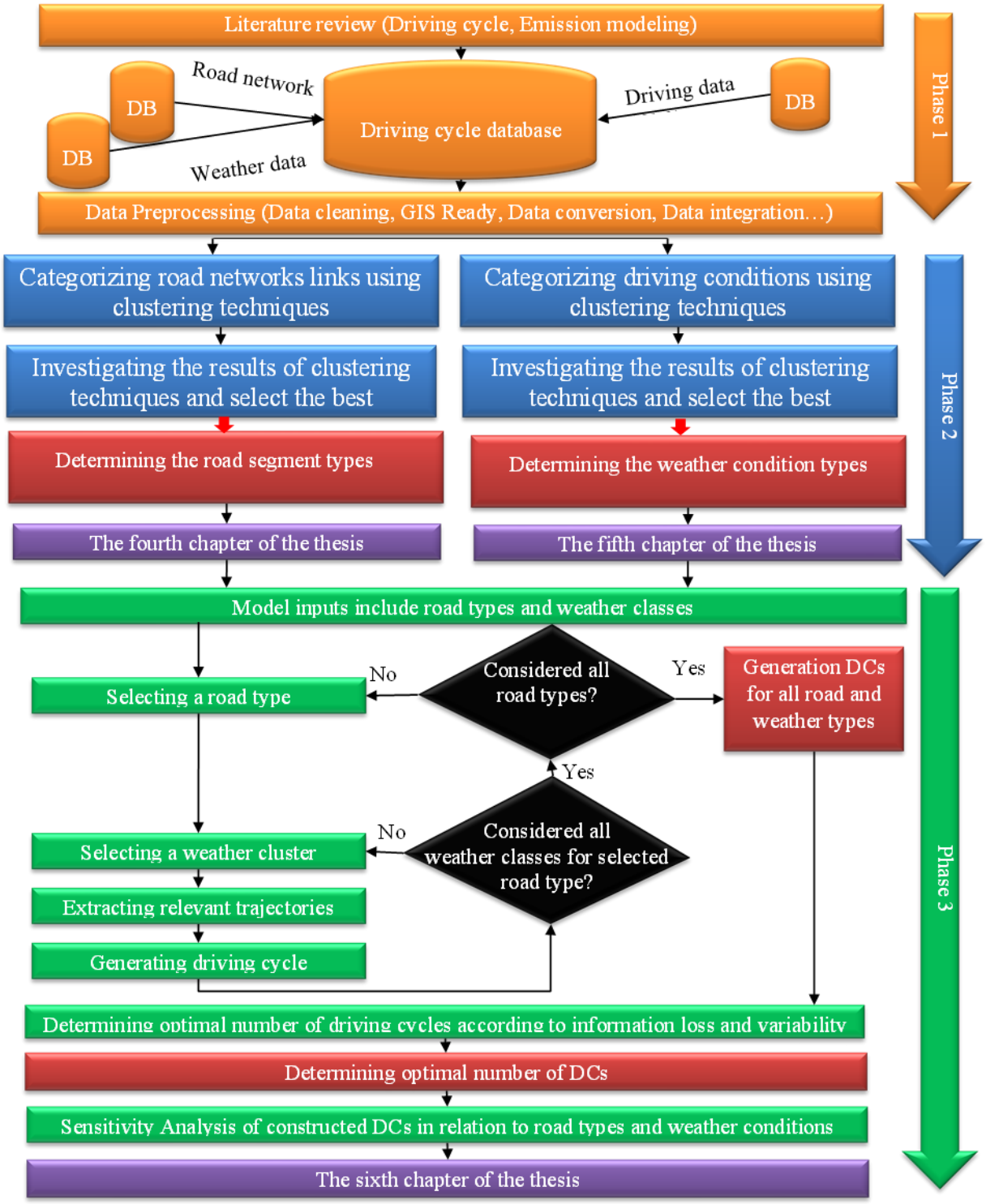


Figure 3-1 Research methodology of this thesis- Image by author, 2020

3.1 Data

Figure 3-1 highlights the foundational importance of data collection and preprocessing in the initial phase of this thesis. As delineated in the literature (section 2.2), multiple factors shape driving behaviors and patterns. To accurately capture the dynamics of driving cycles, it is imperative to first gather the relevant data. Table 3-1 enumerates all the required data components for the envisioned model.

In the past decade, there has been an exponential surge in data availability. However, contemporary datasets, encompassing GPS coordinates, meteorological details, and vehicular statistics, are now widely available.

Within the thesis context, data play an important role for model formulation. Table 3-1 is meticulously crafted to elucidate the specific data prerequisites for each phase. Data marked under the 'Developed model' column will be pivotal in refining and finalizing the model, ensuring the precise determination of “optimal” driving cycle subsets.

3.2 Data Preprocessing

Data preprocessing is one important steps in this thesis to refine raw data and make it primed for analysis. This involves a plethora of operations, ranging from attribute editing, noise, and redundancy removal, to handling missing values and rectifying skewed attributes to mitigate the influence of outliers. Tools such as QGIS, Excel, and Python—with specific emphasis on Pandas, NumPy, and Scikit-learn libraries—are used to facilitate these tasks.

A significant feature in the road data is the 'slope', derivable from the Digital Elevation Model (DEM) through spatial analyst tools in QGIS. Once calculated, each slope value is mapped to its associated road segment. Furthermore, the lengths of these segments, integral to network parameters, can be deduced through geometric computations.

Existing studies highlight the inherent inaccuracies in GPS data, a pivotal resource for constructing driving cycles. This necessitates an extensive array of filtrations and other preprocessing measures. Delving into this, essential filters delineated in works (Duran and Earleywine 2012; Nguyen et al. 2019) are elucidated in section 2.1.2.

Table 3-1 Required data for this project

Influencing factors	Grouping variables	Groups
Street features	Street function	• Freeway • Arterials • Collector street • Local street
	Street type	• Street • Boulevard
	Type of environment	• Residential • Industrial • Central Business District • Green area
	Road design	• Vertical alignment (Slope) • Horizontal alignment (curvatures) • Shoulder • Width • Speed limits
	Density of intersections	
Weather factors	• Temperature • Humidity • Wind speed • Rain • Snow • Fog	

3.3 Preparation of Model Inputs

The proposed model is made to find the best set of driving cycles for Montreal Island. To do that, three main inputs are used: typology of road network segments, typology of weather conditions, and driving cycles. As depicted in Figure 3-1, unsupervised machine learning techniques are deployed to categorize both the Montreal Road network and its diverse weather conditions, given

their notable influence on driving cycles. Furthermore, driving cycles are constructed based on taxi trajectory data. It is worth mentioning that only driving data when a passenger was onboard were used in this thesis.

3.3.1 Driving cycle construction

The methodology for constructing the driving cycle is adopted from (Nouri 2015) using code developed by Roy and Morency (2020).

A Microtrip-based approach is used to construct driving cycles. These are derived from GPS data using the fixed-distance technique, a method that emerged as superior after Nouri (2015) conducted a comparative analysis of six techniques. These microtrips can be characterized and segmented based on either average metrics, like speed or acceleration, or the more encompassing Speed Acceleration Frequency Distance (SAFD). Owing to its capacity to encapsulate variations in speed and acceleration, SAFD has been the favored choice in several studies, including those by (Nouri 2015; Huertas et al. 2018). A comparative analysis between the speed acceleration matrix and the microtrips results in a matrix detailing their dissimilarities. Subsequent application of k-means clustering groups similar entities together.

Post clustering, 14 distinct features pinpointed by Nouri (2015) lay the groundwork for the assessment criteria, which evaluates the designed driving cycle. Following this, a Markov chain model stitches the disparate microtrips together, presenting a comprehensive driving cycle profile. All such synthesized driving cycles subsequently feed into the model slated for development in phase three. The evaluative parameters underpinning this procedure are delineated in Table 3-2.

Table 3-2 Assessment measures (Nouri, 2015)

	Parameters	**unit**
1	Difference in SAFD	%
2	Average speed of the entire driving cycle	Km/h
3	Average running speed	Km/h
4	Maximum speed	Km/h
5	Average acceleration of all acceleration phases	m/s2
6	Average deceleration of all deceleration phases	m/s2

Table 3-2 Assessment measures (Nouri, 2015)

	Parameters	unit
7	Average number of acceleration-deceleration changes (and vice versa) within one driving period	NA
8	Root mean square acceleration	m/s2
9	Road power time proportions of driving modes:	kW
10	Idle (speed = 0)	%
11	Acceleration (acceleration ≥ 0.1 m/s2)	%
12	Cruising (- 0.1 m/s2 < acceleration < 0.1 m/s2, average speed > 5 m/s)	%
13	Deceleration (acceleration ≤ -0.1 m/s2)	%
14	Creeping (- 0.1 m/s2 < acceleration < 0.1 m/s2, average speed < 5 m/s)	%

3.3.2 Road network clustering

To develop a typology for the Montreal Road network, the methodology revolves around the utilization of unsupervised machine learning algorithms. The objective is to group road segments that exhibit similarities with respect to features that influence driving behavior. The specifics of these features are previously outlined under the section titled “Influencing Factors” and are itemized in Table 3-1.

Selecting a clustering algorithm is important to generate road types. There are various clustering techniques having their own pros and cons. In this thesis three techniques including k-means, AHC, and k-prototype are selected to develop road typology and determine the paper with better accuracy. Accuracy in clustering is defined according to homogeneity and heterogeneity of objects in a cluster.

Once the clusters have been defined, the Silhouette measure is invoked to fine-tune and ascertain the best number of clusters.

3.3.3 Weather parameters clustering

In the second phase, a pivotal task involves the categorization of weather parameters. While literature reviews indicate an absence of studies delving into the interplay between ambient conditions and driving cycle variability, classifying these weather conditions could shed light on this. However, the method and number of clusters chosen are of paramount importance. One of the uncertainties surrounding this task is the number of clusters, which has implications both for

processing loads and the accuracy of the resulting model. Table 3-1 details the essential weather parameters for this thesis.

Once clusters are formed, validating the number of clusters becomes vital. This validation can be approached through internal or external indices, measuring the similarity of an object within its cluster versus its dissimilarity with other clusters. Widely recognized validation indices have been developed. The Silhouette index, dendrogram, and Gap statistic are employed to determine the number of road and weather types. For clarity, Table 3-3 offers a concise overview of the validation indices, elucidating their roles in determining the best compromise sets of driving cycles.

Table 3-3 Brief description of validation index

Method	**Description**
Silhouette index	1.Calculate two distances for each data point: • The average distance from the data point to other points in its own cluster. • The smallest average distance from the data point to points in a different cluster, indicating the next nearest cluster. 2.Compute the silhouette score for each data point: • The score is calculated as (b-a)/max(a, b), where a and b are two data points • This score ranges from -1 to +1. A high score means the data point is well matched to its own cluster and poorly matched to neighboring clusters. 3.Average score for all points: • The average of all individual data points' scores is the overall silhouette score. • A higher average score suggests better defined and separated clusters.
Gap statistic	1.Cluster the data and compute the Within-Cluster Sum of Squares (WSS):

Table 3-3 Brief description of validation index

Method	Description
	• For different cluster counts (k), partition the dataset into k clusters, and calculate the total WSS, which measures the compactness of the clusters. 2.Generate reference data and compute WSS for reference data: • Create multiple reference datasets by randomly sampling from a uniform distribution that matches the original data. Cluster each reference dataset for the same range of k values and compute their WSS. 3. Calculate the Gap statistic: • For each k, the Gap statistic is calculated as the logarithmic difference between the WSS of the reference data and the WSS of the original data. • Gap (k) = mean (log(WSSref))-log(WSSoriginal) Where WSSref: Within-Cluster Sum of Squares of the reference data WSSoriginal: Within-Cluster Sum of Squares of the original data 4. Determine the optimal k: • The optimal number of clusters is the smallest k such that Gap(k) ≥ Gap (k+1) – SE (K+1), where SE is the standard error of the mean log (WSS) of the reference data.

3.4 Developing a Model to Determine "Optimal" Sets of Driving Cycle

The output of the second phase is the number of road network clusters and number of weather data clusters. In the third phase, as illustrated in Figure 3-1, the initial step involves selecting a road type, and a weather type and then developing driving cycle for the driving data which are related to them. This is an iterative process and driving cycle should be generated for all road types and in all weather types.

To investigate the developed model's reliability, it is necessary to validate its output. To do that, this thesis performs two following validation approaches:

1. Developing a mixed model: Developing a mixed model is essential to confirm that the values of the constructed driving cycles are significantly associated with road types,

weather conditions, and types of days. Essentially, if the values of the constructed driving cycles respond sensitively to changes in these external factors, it indicates that the developed methodology is effective in identifying relationships between driving cycles and these variables. For this purpose, a mixed model has been developed.

2. Data aggregation level: It is necessary to demonstrate how information diminishes at the chosen level of data aggregation. The significance of information loss lies in its ability to highlight discrepancies between parameters values for all driving data and those for the selected driving cycle parameters.

After validation of the developed framework, the following steps should be taken to determine the "optimal" number of driving cycles:

1. Similarity level: After generating driving cycles for all road types and all weather types One of the main concerns is to measure the similarity between them. In fact, it is required to know the constructed driving cycles in different conditions should be varied.
2. Measuring information loss and variability of driving cycles: The number of driving cycles is considered "optimal" when they capture the most variability and have the least information loss. Greater variability in driving cycles indicates they represent a wider range of driving behaviors and lower information loss means the constructed driving cycles are more similar with real driving cycle parameters.

3.4.1 Validation of developed methodological framework

Mixed models

As road types, weather conditions, and days are clustered, a mixed model is an appropriate statistical tool to investigate their impacts on driving cycle parameters. This approach efficiently handles the grouped nature of the data, allowing for an analysis of how these specific factors influence driving behavior.

The developed model uses a hierarchical data structure, so mixed models might be the best choice for analysis. The data is hierarchical because the relationship between road types, weather conditions, and driving cycles forms a tree-like structure. Figure 3-2 displays the data structure for the proposed model. Clearly, the highest level represents an urban network, followed by road type (denoted as RT). Below that is the weather type (represented as WT), and at the bottom are the

driving cycles. This setup means that each road type encompasses at least one weather type, and each weather type contains at least one driving cycle.

To perform sensitivity analysis, the developed driving cycles are integrated into road types and weather types. By integrating them a hierarchical structure is created then using the mixed model it is possible to determine association of driving cycle parameters and weather and road types.

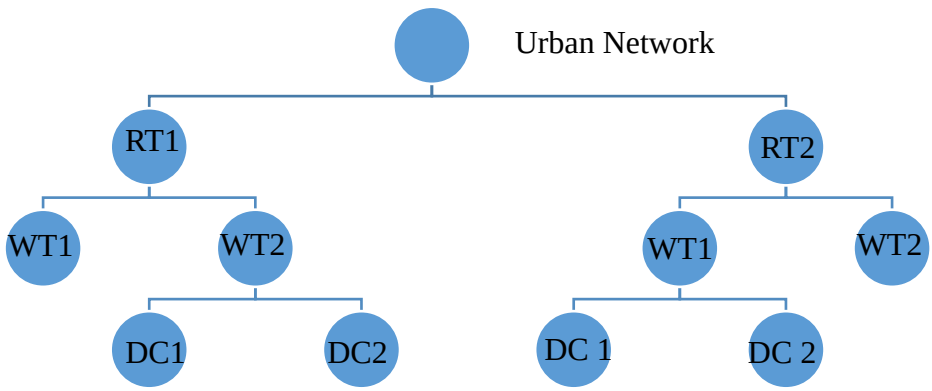


Figure 3-2 Hierarchical data structure

The design of the model is based on the Python programming language and employs the Statsmodels library. As illustrated in Figure 3-2, the model has a hierarchical organization, considering factors like road and weather conditions, and differentiating between weekdays and weekends as fixed effects. In contrast, drivers are regarded as random effects; they are chosen randomly, and specific driving patterns are created for each. The final stage involves formulating equation 3.1, which is used to analyze how these fixed effects interact and their impact on the random effects.

$$DC_p = \boldsymbol{GI_p} + \sum_{i=1}^{n} \beta_{1i,p}\, RT_i + \sum_{j=1}^{m} \beta_{2j,p}\, WT_j + \sum_{k=1}^{l} \beta_{3k,p} DT_k + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} \beta_{ijk,p}\left(RT_i * WT_j * DT_k\right) + \boldsymbol{\varepsilon}_p \tag{3.1}$$

Where:

DC_p: Dependent variable for the p-th driving cycle parameter, where p ranges from 1 to 13. Each number from 1 to 13 represents one of the 13 driving parameters that are being analyzed.

n: Total number of road types,

m: Total number of weather conditions, and

l: Total number of day types.

$GI_{\mathbf{p}}$: Global intercept of the chosen parameter.

$\boldsymbol{\varepsilon}_{p}$: Random error term for the p-th parameter.

$\beta_{1i,p}$: Coefficient for the effects of the i-th road type on the p-th parameter.

$\beta_{2j,p}$: Coefficient for the effects of the j-th weather condition on the p-th parameter.

$\beta_{3k,p}$: Coefficient for the effects of the k-th type of day on the p-th parameter.

$\beta_{ijk,p}$: Coefficients for the three-way interaction effects among the fixed factors for the p-th parameter.

RT_i, WT_j, DT_k: Indicator variables correspond to each level of the categorical fixed effects.

Data aggregation level

In literature, researchers often aggregate all driving data to produce just one or a few driving cycles. However, aggregating all driving data poses two main issues. Firstly, it leads to a significant loss of information. Secondly, the resulting driving cycle might not accurately reflect true driving behavior. This is because different road segments and varying weather conditions can influence driving behaviors differently. Aggregating their data can make the derived driving cycle a poor representation of actual driving patterns.

The level of aggregation can vary. It might encompass the entire road network over a year, or it could be based on road type, weather type, and so forth. Each level produces a different number of driving cycles and can result in varying amounts of information loss.

To calculate the optimal level of data aggregation two distinct metrics to evaluate information loss are developed. Information loss is defined as discrepancy between the developed driving cycle parameters with values of driving cycle parameters of microtrips which are developed from all driving data.

The first metric, defined in equation (3.2) proposed by Liu et al. in 2020, measures the ratio between the range of chosen driving cycle parameters and the range of parameters for all drivers. This is termed as the ratio of sampled to actual range. To calculate this, since estimating parameters for

every driver is impractical, the complete dataset of trips is input into a driving cycle tool to generate microtrips. As a driving cycle consists of these microtrips, their range is considered equivalent to the range of driving cycle parameters. Therefore, the range of microtrip parameters serves as a proxy for the actual range of driving cycle parameters. The formula is:

$$\text{Ratio of selected DC to all DCs (IL1)} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{R_{dcp}^{s}}{R_{mtp}^{a}}\right) \quad \textbf{(3.2)}$$

where

i : set of driving cycle parameter i ϵ {1,2, …, N}

R_{dcp}^{s}= $|pmax_i^s - pmin_i^s|$, range of ith selected driving cycle parameter in an aggregation level

R^a_{mtp} = $|pmax_i^a - pmin_i^a|$, actual range of corresponding microtrip parameter

The second indicator, equation 3.3 calculates the deviation of quartiles of the selected driving cycle parameters from quartiles of microtrip parameters.

$$\text{Distribution-based information loss (IL2)} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{3}\sum_{j=1}^{3}\frac{\left|PQ_{ij}^{s} - PQ_{ij}^{a}\right|}{\left|PQ_{ij}^{a}\right|}\right) \quad \textbf{(3.3)}$$

where

i: set of driving cycle parameter i ϵ {1,2, …, N}

j: set of quarter j ϵ {1,2, 3}

PQ_{ij}^{s} = The jth quartile of ith selected driving cycle parameter in an aggregation level

PQ_{ij}^{a} = The jth quartile of actual range of corresponding driving cycle parameter

3.4.2 Determining “optimal” number of driving cycles

Measuring similarity between the constructed driving cycles

To determine the “optimal” number of driving cycles in an urban area, it is crucial to highlight their differences. When constructed driving cycles diverge, this indicates that external factors, such as road and weather conditions, influence driver behavior. To assess the similarity of these cycles, a similarity matrix is used. This matrix, composed of arrays, calculates similarity values through pairwise comparisons of driving cycles. These values range from 0 to 1: a value of 0 denotes complete dissimilarity, 1 indicates identical cycles, and values in between reflect varying degrees of similarity.

To create this matrix, the SciPy library in Python is utilized, employing the Euclidean distance method to determine the similarity between two driving cycles based on their respective values.

Variability of driving cycles and information loss

This thesis determines the "optimal" finite number of driving cycles according to variability and information loss. The number is optimized when the number of driving cycles represent driving behaviors in urban road networks under various conditions are minimized, variability is maximum and information loss is minimum. To achieve this, the process involves measuring the similarity of driving cycles and merging those with the highest similarity. Subsequently, the variability and information loss of all constructed driving cycles are assessed. The goal is to identify a level of similarity where information loss is minimized, and variability is maximized. While information loss and its indicators have already been addressed, the subsequent sections will focus on explaining variability and discussing its relevant indicators.

Variability provides a measure of the spread or dispersion of a set of data. In this thesis variability degree of difference or dispersion among the constructed driving cycles; therefore, a decrease in variability signifies a decrease in the representativeness of the selected driving cycles.

Although there are various variability indicators to use, selecting them strongly depends on distribution of the generated driving cycles. For example, the coefficient of variation can be misleading for skewed distributions.

The Proportional Variability (PV) indicator, as introduced in equation (3.) and developed by Heath and Borowski in 2013, is utilized in this study to assess the variability of driving cycles. PV is effective in a variety of conditions and dynamics, and its values are proportionally scaled between 0 and 1, making it universally applicable.

$$\mathbf{PV} = \frac{\sum_{\mathbf{p=1}}^{\mathbf{M}}\left(\frac{\mathbf{2} * \sum_{\mathbf{comb.}} \mathbf{D}(\mathbf{z_{pi}}, \mathbf{z_{pj}})}{\mathbf{n}(\mathbf{n}-\mathbf{1})}\right)}{\mathbf{M}} \qquad \textbf{(3.4)}$$

Where i,j be indices such that i,j∈{1, 2, ..., n} and i≠j. These indices are used to calculate the relative differences between the parameters of driving cycles. For a given parameter p, z_{pi} represents the value of p in the i-th driving cycle, and z_{pj} represents the value of p in the j-th driving

cycle. The function $D(z_{pi}, z_{pj}) = \frac{|z_{pi} - z_{pj}|}{Max(z_{pi}, z_{pj})}$ is defined to quantify the relative difference between these two values. n is the total number of driving cycles and M is total number of driving cycle parameters.

This process of evaluating relative differences across all unique pairwise combinations allow for a comprehensive assessment of the variability within the dataset, making PV a robust and flexible measure for this purpose.

3.5 Publications of this thesis

In the course of this research, three papers have derived from the findings of this thesis as follows:

12. The first paper, "*New Data-Driven Approach to Generate Typologies of Road Segments"*, has been published in *Transportmetrica A: Transport Science*. The main contributions of this paper are introducing a novel typology for road segments, outlining a method for validating clustering results in the field of transportation, and presenting models that enhance our understanding of the relationships between road features and driving speeds.
13. The second paper, "Identifying Representative Number of Driving Cycles To Represent Diverse Driving Conditions" has been submitted for peer review and are currently under consideration for publication in International Journal of Sustainable Transportation. This research primarily addresses three key issues related to the variability of driving cycles under different weather conditions, using thorough scientific research and analysis. These concerns include: 1) The inadequacy of existing driving cycle development methods in accurately reflecting real-world driving conditions; 2) A lack of comprehensive analysis on how weather characteristics influence driving cycle parameters; 3) A notable gap in systematic methodologies for classifying various weather types, particularly in the context of vehicle emissions.
14. The third paper, "*A Methodological Framework to Determine the "Optimal" Set of Driving Cycles for A Region*" has been submitted for peer review and are currently under consideration for publication in *Transportmetrica A: Transport Science*. This research paper focuses on bridging three crucial gaps in the field of driving cycle generation. Firstly, it seeks to identify the "optimal" number of driving cycles required for accurate representation. Secondly, the paper introduces a new approach for aggregating driving data.

Lastly, it develops a mixed model to analyze the impact of external factors on the parameters of driving cycles.

CHAPTER 4 ARTICLE 1: NEW DATA-DRIVEN APPROACH TO GENERATE TYPOLOGIES OF ROAD SEGMENTS

Written by:

Asad Yarahmadi, Catherine Morency, Martin Trepanier

Published in January 2, 2023

Journal: Transportmetrica A: Transport Science, https://doi.org/10.1080/23249935.2022.2163206

Abstract

The main objective of this study is to put forward a typology to describe better associations between road segments and driving patterns as reflected by driving speed. As the first step, a regression model is developed to examine the association between road segments and driving speed. Then, various unsupervised machine learning techniques, including k-means, AHC, and k-proto[3], are used to develop typologies of road segments. Speed data from a fleet of taxis operating in Montreal, Quebec, are used to validate the discrimination power of the various typologies. Results demonstrate that combining k-means and Gower distance produces the most accurate road typology. Various statistical tests, including ANOVA, Leven, and post hoc analyses, confirmed that the speed values of the various road types are significantly different. Finally, R2 of regression models developed for various road types demonstrated that the generated road types better elucidate the variability of driving speed.

Keywords: Road type, driving speed, road features, data-driven model.

[3]. K-prototype

4.1 Introduction

A road segment can be characterized by various features, such as the gradient, curvature, the number of lanes, etc. Numerous studies show that these characteristics have a wide variety of effects on driving speed and vehicle emissions. Brundell-Freij & Ericsson (2005) show that by increasing the density of signalized intersections operating speed decreases by 41%, and vehicle emissions escalates. However, this study does not expound precisely on the quantity of this incrementation. Svenson & Fjeld (2016) explore the relationships between road geometry and fuel consumption and conclude that undulation, curvature, and surface roughness are responsible for a 107% increase in fuel consumption. Goralzik & Vollrath (2017) reveal that the drivers' speed choice on the road with a speed limit of 50 km/h can be strongly affected by road width. Very poor road conditions decrease vehicle speed by 55% compared to the best driving conditions (Varella et al. 2019). As a result, vehicle emissions increase by 2.49%. Nouri (2015) demonstrates that a change in the road gradient of almost 10%, on average, increases fuel consumption by 5%, although they are non-linearly related. Therefore, road features influence driving speed and vehicle emissions (Liu et al. 2020). Carlson et al. (2013), through project MIRAVEC (Modeling Infrastructure Influence on RoAd Vehicle Energy Consumption), show that vehicle fuel consumption is a function of road width, speed limit, curvature, and gradient.

Previous studies have proposed typologies of road segments based on their functions, which are related to the type of service that a road segment provides. However, this approach is limited to answering questions such as whether all road segments with the same function have similar effects on driving speed and vehicle emissions. In addition, the research works that analyze the relationship between road factors and driving speed mainly focused on a few road links with a certain road function. For example, Abele et al. (2021) explored the influence of road design on driving speed and solely focused on a 5 km rural road. Nama et al. (2016) studied the driving speed in road segments with both road gradients and curvature. Such an approach cannot provide insights into whether equivalent results would be obtained for segments presenting other road functions or if the effects of the road features remain constant when other combinations are tested. For example, does the impact of slope's effect change with different curvature radius?[4]

[4] . In other words, does the effect of a slope's inclination change with varying radii of curvature?

This study is conducted to bridge the identified gaps. The main objective of this study is to put forward a typology to describe better associations between road segments and driving patterns as reflected by driving speed. To this end, machine learning techniques are utilized to assort road segments into various road types according to the factors affecting driving patterns. It is expected that segments of the same road type have a similar influence on driving speed. This will be further explored in the future steps of this research aiming to propose optimal sets of driving cycles to represent the driving behaviors of a region. The road network links of the Montreal Area, used as the case study in this research, are categorized. Hence, this paper also aims to analyze the influence of changes in road features' combinations on driving speed.

This new typology can be important for scholars, transportation managers, and practitioners to better understand the role of road networks in vehicle emissions and the interaction between driving speed and road features. In addition, it helps decision-makers improve transportation planning, road infrastructure design, and traffic management. For example, according to this typology, managers may have better insights into which road segments have the most major influence on driving patterns and how these segments are distributed over the area. Moreover, this approach could open a new gate for future researchers to develop different typologies for other transportation issues, such as road safety.

The research questions in this study are as follows:

- How many road types are necessary to correctly describe an urban road network?
- What is the best approach to identify the road types? How could the result of the classification be validated?
- Do the effects of street features on driving speed vary for various road types?
- What factors in each road type have the most significant influence on driving speed?

The contributions of this study are as follows:

- A new typology is generated to identify and categorize all road network segments that have similar influences on driving patterns.
- Various models are developed to explain more efficiently the relationship between road types and driving speed. Table 4-9 proved that the R2 in developed models is higher than that of the base model.

- To increase the reliability of road network typology, a two-step validation method is developed.

Developing the typology of the road network enables us to determine the number of road types having different influences on driving speed in an urban area. This typology is the most crucial step in designating optimal sets of driving cycles which is the next step of this study. The driving cycle, representing average driving behavior, is one of the most important input parameters in vehicle emission models like the MOVES model.

The remainder of this paper is structured as follows. Previous efforts that generated road types are investigated, and arguments for and against these typologies are exposed. The “Methodology” section describes the technical steps to categorize road links using various techniques. The following section presents the typologies that were elaborated on in detail. Finally, the results are discussed, and some perspectives are proposed.

4.2 Background

The background section is composed of two parts. In the first part, approaches developed to investigate associations between road type and driving speed are reviewed, and their advantages and disadvantages are explored. The second part critically examines research conducted to generate a typology of road networks. These two components allow us to identify the relevant research gaps.

4.2.1 Factors Affecting Driving Patterns

An urban road network is complex since it comprises many segments that can present various features. Among these features, road function classification (i.e., highway, collector, and local) (Ericsson 2001a), road gradient (Tzirakis and Zannikos 2013), speed limit (Silvano and Bang 2016), number of lanes (Dinh and Kubota 2013), horizontal alignment (Ferreira, Rodrigues, and Pinho 2019), the density of signalized intersections (Ericsson, Larsson, and Brundell-Freij 2006), accessibility (Aronsson 2006), and surrounding land use (Han et al. 2020) have of noteworthy influence on driving speed and vehicle emissions. Exploring the effects of these features on driving speed is challenging. Firstly, each feature can have a wide range of values. The result of features can hence change depending on their values. For example, Chen et al. (2017) used a t-test model to verify the grade's impact on driving speed on 90 highway segments. They demonstrate that speed values fall from aro110 km/h to 80km/h when the slope changes from -4% to 4%. However, no

details about segment attributes and spatial resolution of the Digital Elevation Model (DEM) were presented in the paper. Furthermore, the combined effect of multiple features could change the influence of one factor. For example, Nama et al. (2016) analyzed the combined effect of the gradient and curve on driving speed on mountain roads in India. They showed that the curve's influence on driving speed increases when the gradient decreases. Still, their study has limited value due to the small sample used, especially regarding the length of curves, which are distributed from 10 m to at most 250 m and are too short to investigating the speed changes.

The primary approach to overcome this complication was to focus on a particular road function and a limited number of road factors. Abele et al. (2011) simulated 5 km of rural roads to probe the association between the presence of roadside trees, road shoulder, horizontal alignment, and driving speed. The authors concluded that the road shoulder and trees along the driveway did not affect driving speed. However, owing to the small sample size, they recommended considering their findings provisional results. Liu et al. (2016) explored the influences of lane width, lane position, and shoulder width on driving speed for a three-lane underground urban expressway. According to statistical analysis, they concluded lane width has a more significant influence than two other factors, with drivers going on average 20 km/h faster in wider lanes.

Although considering a few road factors, this approach simplifies the current complication and can provide some knowledge about the interaction between driving speed and road features; it could also be a source of contradicting results[5]. For instance, Silvano & Bang (2016) showed that the road environment could heavily impact driving speed, but Andersen & Reinau's (2016) results show that the land use cannot affect driving speed. Babić et al. (2020) and Abele et al. (2021) indicated that the shoulder does not affect driving speed, while Ben-Bassat & Shinar (2011) demonstrated that shoulders could have considerable effects on driving speed. Indeed, many examples of contradicting results exist in the literature. It should be noted that another source of contradiction could be related to the simulation method used to investigate the relationship between road features and driving speed. Driving simulation is a computer-based method for examining driving speed in a controlled environment. The drawbacks of the simulation-based approach are that simulation tools primarily cannot consider many vital details of the real world, such as traffic conditions, in their computation. This can cause uncertainty in the achieved results. Additionally,

[5]. Although considering a few road factors, this approach simplifies working on interaction between driving speed and road features; it could also be a source of contradicting results.

the simulation's accuracy, validity, and credibility depend highly on the simulation parameters, such as Field of View (FOV) (Durkee, 2010). The field of view (FOV) is the maximum area a simulator can present to a user.

4.2.2 Road Classification Methodology

Categorizing road types according to their effects on driving speed has not yet been performed. Previous studies that categorized road segments can be grouped into two groups: research that categorizes the road segments according to their functions, named *Functional Classification*. It is worth mentioning that the functional classification determines the role that a road link plays in serving traffic flow in the street network (FHWA, 2013). Road functions differ from country to country, but the standard functions are highways, collectors, locals[6], and external segments (FHWA, 2013). The second group of studies develops an algorithm to identify the segments according to their road surfaces, such as asphalt and concrete.

Almost all former studies generated typology for road networks according to the concept of functional classification. To accomplish this typology, the main criteria are the characteristics of the traffic service, such as the flow rate and efficiency for motorized vehicles (Kliment Ohridski-Bitola, 2009), and the environmental aspects of transportation (i.e., vehicle emissions) are neglected. City Council developed the City of Toronto Road Classification System (Stevanovic-Briatico 2013). In this system, thirteen criteria are included: “Traffic movement versus property access,” “Typical daily motor vehicle traffic volume,” “Minimum number of peak period lanes,” “Desirable connections,” “Flow characteristics,” “Legal speed limit,” “Accommodation of pedestrians,” “Accommodation of Cyclists,” “Surface transit,” “Surface transit daily passengers,” “Heavy truck passengers,” “Typical spacing between traffic control device,” and “Typical right-of-way width.” Consequently, the Toronto Road segments are categorized into four types: expressways, major arterial roads, minor arterial roads, collector roads, and local roads. Vitkienė et al. (2017) reviewed ten countries' road classification systems and road hierarchies, including the USA, Germany, Austria, Poland, Denmark, Greece, Norway, Portugal, and Lithuania. The authors concluded that there is no ordinary road classification system. Instead, links, roles, and road functions are the most typically used among the criteria considered for classifying road type.

---

[6]. Road functions differ from country to country, but the standard functions are highways, collectors, locals[6] (FHWA, 2013)

The second group of studies is more focused on developing data science techniques to identify road networks and classify them according to the types of road surfaces. Most focus on developing approaches to extracting information from images, vector data, and mobile phone data (Gonzalez et al. 2017). Saleh et al. (2010) classified the road type according to the images gathered from Google Street View and Google Image. However, they did not include essential features such as geometry and road design, and their method identified roads only according to the surface type (sand, asphalt, concrete). Han et al. (2020) took advantage of topological indicators to classify road network links. To measure the similarity between street networks, they first used topological principles to define a tree-like network structure with connectivity, hierarchy, and accessibility variables. Then, these variables were used to measure the similarity between the links. Finally, the developed model was estimated on various urban road links. The main limitation of this project is related to factors considered in the classification. The authors solely considered the geometric forms to classify road types without including other essential elements such as the speed limit, slope, etc. Gonzalez et al. (2017) proposed a novel technique developed from the Bag of Words (BoW) representation and used accelerometer data to classify Roadway Surface Disruption (RSD) patterns. RSD is a permanent obstacle (i.e., potholes and speed bumps). The authors proved that combining BoW and Artificial neural Networks has excellent potential to capture RSD patterns from the accelerometer data and generate high-accuracy RSD classes. However, the developed algorithm has high computational complexity and needs to be performed in a cloud environment.

In conclusion, it is worth noting that the road network variables influencing the speed and vehicle emissions were not considered in the examined classifications. Moreover, most of the methodologies developed to categorize road types included only limited features, such as road functions and types of road surfaces. They did not investigate the influence of various features. Hence, the results have limited relevance for travel pattern analysis. The critical factors affecting driving speed must be considered to achieve results relevant to supporting the research of driving patterns.

4.3 Methodology

This section describes the data used in this research and provides background information about the techniques deployed. Figure 4-1 shows the various steps of this study.

Figure 4-1 Phases to perform the data-driven framework- Image by author, 2022

As Figure 4-1 displays, this study comprises three distinct phases, expounded in detail in sections 4.3.1, 4.3.2, and 4.3.3. In the first phase, all required data are gathered, then feature engineering techniques are used to prepare them for analysis. In the last step of the first phase, a regression model, named *base*, is generated to investigate the relationship between road features and driving speed. Next, in the second phase, all road segments are categorized using three clustering analysis approaches. Afterward, various validation techniques validate clustering results[7]. Finally, a

[7]. Afterward, various validation techniques are employed to validate clustering results.

regression model is utilized to analyze sensitivity and to investigate how the generated road types describe driving speed changes.

4.3.1 Phase 1. Data preprocessing

Before describing the data preprocessing, the data used in this study are described below.

Data description

Unlike former studies that mostly focused on a particular road function and a limited number of features, this research uses all street links of the Montreal Island network. Road functions in Montreal are Highway, Primary arterial, Secondary arterial, Ramp, Tertiary, and Local. All road functions and their varied features are included in this research, except for segments with the local function. The local function composes roughly 64% of Montreal Island network segments.

According to the literature, the street characteristics listed in Table 4-1 significantly affect driving speed and vehicle emissions. The Montreal Road data was downloaded from Open Street Map (OSM), with roughly 9,409 segments[8] for a total length of about 131 km.

Speed data are collected in the actual driving conditions, including peak and off-peak hours, using GPS embedded in taxis when they were occupied from April 2020 to December 2020. Descriptive details about the data are presented in Table 4-1.

Table 4-1. Description of data used in this study

Variables		**Statistical values of variables**
Road functions	Highway	Length: 25.675 (km)
		Share in % segments: 23%
	Primary arterials	Length: 14.921 (km)
		Share in % segments: 9%
		Length: 61.436 (km)

[8] . A road segment is a link between two intersections.

Table 4-1. Description of data used in this study

	Secondary arterials	Share in % segments: 41.52%
	Ramp	Length: 5.238 (km)
		Share in % segments: 7%
	Tertiary	Length: 24.519
		Share in % segments:18%
Speed limit		30 (km/h)
		Share in % segments: 3%
		40 (km/h)
		Share in % segments: 11%
		50 (km/h)
		Share in % segments: 60%
		60 (km/h)
		Share in % segments: 4%
		70 (km/h)
		Share in % segments: 10%
		100 (km/h)
		Share in % segments: 12%
Number of lanes		Min: 1

Table 4-1. Description of data used in this study

		Max: 7
Vertical alignment		Min: -8%
		Max: +8%
		STD: 1.2
		Avg: 0
The density of signalized intersections (km2)		Min: 0
		Max: 15
		STD: 2.8
		Avg: 2.5
Horizontal alignment		Min: 94 (m)
		Max: 927 (m)
		STD: 517
		Avg: 546
Land use:	Residential	Share in data(area): 44%
	City center	Share in data(area): 9%
	Industry	Share in data(area): 23%
	Conservation area	Share in data(area): 1%
	Large public infrastructure	Share in data(area): 0.6%

Table 4-1. Description of data used in this study

	Mixed activities	Share in data(area): 18%
	Park	Share in data(area): 2%
Accessibility		Min: 0.007
		Max: 0.64
		STD: 0.14
		Avg: 0.15
Vehicle speed		Min: 1 (km/h)
		Max: 97 (km/h)
		STD: 16.64 (km/h)
		Avg: 31.28 (km/h)

As Figure 4-1 displays, the first step of this study is data preprocessing. Specifically, this project is crucial because the data used is of mixed types, unscaled, highly skewed, and few features are correlated. These issues can strongly affect the accuracy of results from machine learning techniques.

The preprocessing step comprises two main phases: generating required road features (data enrichment) and feature engineering. Each phase is explained in the following sections.

Step 1. Preparing required road features in QGIS

A road network segment can have many characteristics, but according to the literature, a limited number have considerable effects on driving speed. To determine the most influential factors, the literature was reviewed. Table 4-1 lists the most significant factors. Among these variables, “radius of the curve,” “gradient,” “density of signalized intersections,” and “accessibility” were calculated. The radius of the curve was calculated from road geometry; road gradient was computed from Digital Elevation Model; the density of signalized intersection was determined by counting the number of signalized intersections in a cell-sized 500 m × 500 m; accessibility was estimated using

the proximity measures data produced by Statistics Canada and Canada Mortgage and Housing Corporation. Accessibility is a measure showing the ease of reaching various activities distributed in space. A proximity measure relies on a gravity model that shows the distance of different activities from the center of a dissemination block[9]. It is worth mentioning that the value of the proximity measure is normalized between 0 and 1; the zero value means there is no access to activity, and 1 represents the highest proximity. This study calculates the value of accessibility, which is equal to the average of various measures, including education, shopping store, employee, and population. Then using a spatial query, it is assigned to the closest segment.

Step 2. Feature engineering

Feature engineering is a set of techniques used to make input data compatible with machine learning algorithms and reduce computational complexity (Chen et al. 2020).

The first feature engineering step is correlation analysis. As its name implies, the correlation between segment variables is explored in this step, and highly correlated variables, referred to as multicollinearity, are deleted. A feature is strongly correlated if its correlation coefficient is more than 0.7 with another part (Ratner 2009). To perform the correlation analysis, the Dython library in Python is utilized[10]. Figure 4-2 displays the heatmap in which the values indicate the correlation coefficients between road features. The green box in Figure 4-2 highlights the correlation coefficients of road function with the speed limit and the number of lanes which are 0.84 and 0.63, respectively. Accordingly, it is concluded that the road function strongly correlates with these two factors (Schober and Schwarte 2018), and this factor is excluded from this study's following steps. It is worth mentioning that eliminating collinear features is one of the most critical steps in feature engineering because it affects the statistical power of the multivariate regression model, which is developed in section 4.4.1, by decreasing the precision of the estimated coefficients in the regression model.

[9] . Reference: (Statistics Canada, 2020)

[10] . This function automatically selects the appropriate correlation measure: Pearson's correlation for continuous variables, Cramér's V for categorical variables, and the correlation ratio for mixed pairs (continuous and categorical). It effectively handles both categorical and continuous variables, ensuring statistical appropriateness for each variable pair and presenting the results in a comprehensive matrix form.

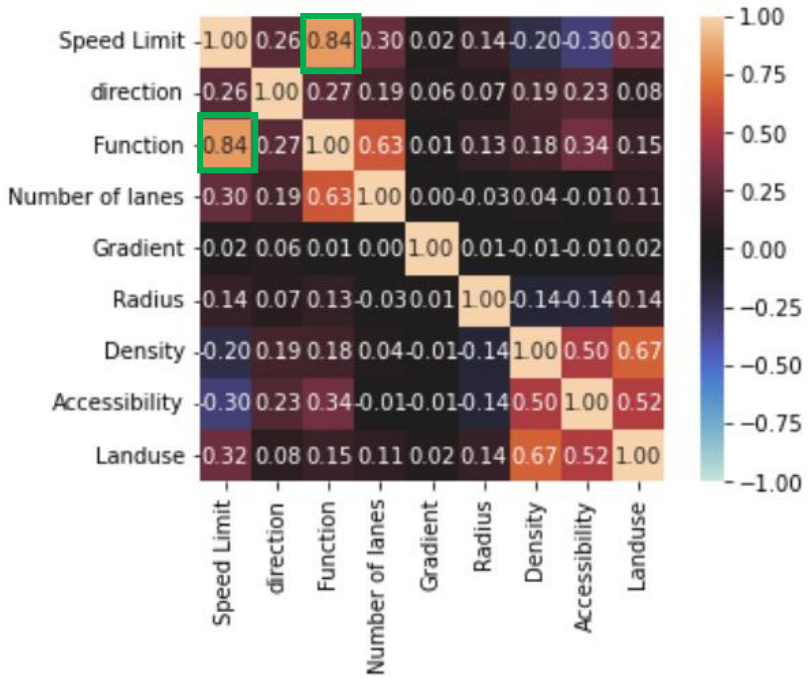


Figure 4-2 Correlation analysis between road features

Most machine learning algorithms (e.g., regression models) explicitly assume that the data follow a normal distribution[11]. So, the second step of feature engineering was to evaluate the distribution of every segment feature. To do that, skewness was employed. Skewness shows how symmetric data distribution is. If the skewness values of a column exceed 0.5, it is skewed. To calculate the skewness, the SciPy library in the python language programming is used. The result shows that the speed limit, radius of curves, and density of intersections were highly skewed with values of 1.95, 5.4, and 1.71, respectively. To resolve skewed distributions, logarithmic transformation and square root are efficient. In the log transformation, the logarithm of values is used rather than the raw data. As its name implies, the square root of values is used instead of the original values. According to the literature, the log transformation transforms skewed data better. Still, it could not be performed on the network data because the radius and density of signalized intersections include many zero values. Most importantly, the slope contains many zero and negative values. First, the minimum value was determined to eradicate negative values, which was -8 then nine was added to all slope values. As a result, slope values start from 1 to 17. Next, one was added to all values to shed zero

[11] . "In almost all statistical and machine learning analyses, it is necessary to perform some data transformations (i.e., data transformation, scaling, centering, standardisation and normalisation) on the raw (but tidy and clean!) data before it can be used for modeling."(rstudio-pubs-static.s3.amazonaws.com).

values in curvature. Eventually, the scikit-learn library is used to perform the log-normal transformation, and skewness values of the speed limit, radius of curves, and density of intersections dropped to 0.14, 0.48, and 0.02.

Another issue that could affect the accuracy of machine learning technique output is the inconsistency of units of continuous data which were unscaled. For instance, the unit of the radius of the curve was the meter; the gradient was percent, and so on. To resolve this matter, all continuous values were standardized. To do so, each value is subtracted from the mean value of that feature and divided by the standard deviation[12]. The generated values were distributed from 0 to 1.

The last feature engineering technique that was performed was Dimensionality Reduction. This technique is mainly used to find and safely drop items with the lowest variance values compared to other variables. Generally, there are three techniques: Principal Component Analysis (PCA) is for continuous values, Multiple Correspondence Analysis (MCA) is suitable for categorical values, and Factor Analysis Mixed Data (FAMD) is used for the mixed data type. This research uses FAMD to prepare data for clustering approach one and two, respectively. The PCA generates a covariance matrix of all features and then calculates their eigenvectors and eigenvalues, showing their magnitude of variance. Next, the number of components needed to describe should be determined according to the eigenvalues. To do that, the cumulative explained variance ratio is utilized, and Figure 4-3 illustrates that 95% of components are needed to explain variance. Another technique is FAMD. Implementing the FAMD on mixed data types is necessary because it balances the influences of categorical and numerical values on the analysis. To do that, first, all numerical values were scaled to variance, and the MCA was utilized to scale categorical values. The MCA is composed of two following steps. First, a one-hot encoded version is computed, and then a corresponding analysis is executed to generate a contingency table by analyzing dependencies between categorical variables.

[12] . Scikit-learn, StandardScaler

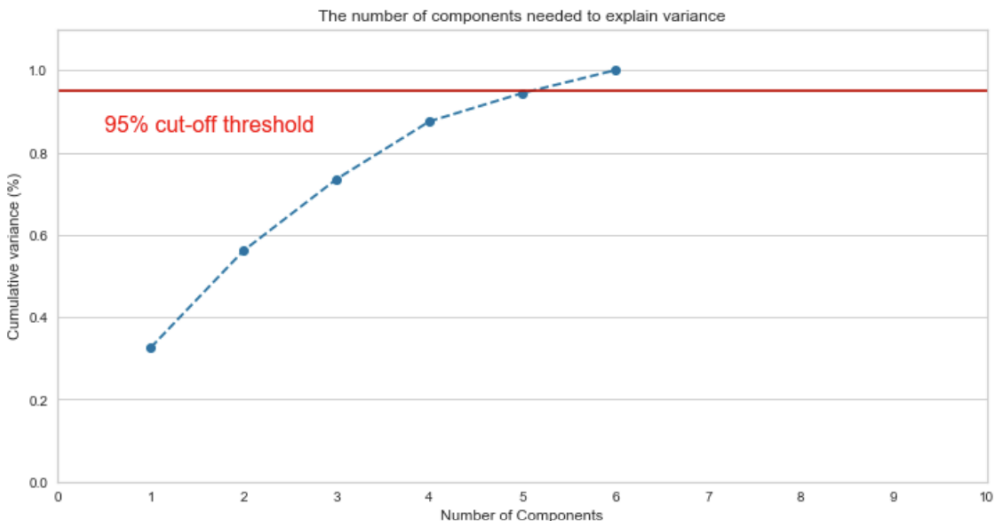


Figure 4-3 The number of components for the PCA

4.3.2 Phase 2. Cluster analysis of mixed-type data

Clustering or unsupervised learning is a descriptive analysis method to categorize data into groups called clusters according to their similarities. Data in each group has maximum similarity while having maximum dissimilarity with the data clustered within other groups.

Two common clustering methods are Partitioning and Hierarchical (He, Agard, and Trépanier 2020; Viallard, Trépanier, and Morency 2019). In this study, both methods investigate which technique could produce more accurate results.

Partitioning techniques separate data according to a similarity index based on a prespecified number of clusters. The most conventional methods are K-means, k-modes, and k-prototype[13]. These are performed on numerical, categorical, and mixed-type data, respectively. The main challenge of the partitioning method is in prespecifying cluster numbers before executing the technique. The primary approach to determining an optimal number of clusters is the Elbow method.

As the name implies, the hierarchical clustering technique generates a hierarchy of clusters. Contrary to partitioning methods, with this method, the number of clusters can be designated by the dendrogram. A dendrogram is a graphical approach by which users can determine the best cluster number. The most common technique in hierarchical clustering is Agglomerative

[13] . Reference: (Vos, 2021)

Hierarchical Clustering (AHC). AHC is a bottom-up approach that considers each piece of data as a cluster, then, according to similarity, iteratively joins the data and recursively produces larger clusters (Ghaemi et al. 2017). This process is terminated when all clusters are combined, and a supercluster is generated.

The most widely known clustering techniques work with continuous values or categorical values. Two approaches exist for handling mixed-type data: encoding the categorical data to numerical data and performing the Gower Similarity Index (GSI)[14]. The GSI uses two similarity measures, namely Manhattan distance and Dice coefficient for numerical and categorical, respectively, to measure similarity between road segments, and produces a similarity matrix with values between 0 and 1. Zero corresponds to maximum similarity, and 1 represents the minimum. The encoding technique maps a categorical value to unique numerical values so that the clustering technique can be performed. In this study, both approaches investigate which method generates the more accurate results.

The last part of clustering is validation. Cluster validation helps evaluate the clustering quality and determine the optimal number of clusters. In this study, a two-step validation technique is developed. First, the clustering results are evaluated using regular clustering validation, like Gap statistics and silhouette. Then, using speed distribution data, the generated road types are evaluated. It is assumed that similar road segments should represent similar driving speeds. Therefore, if road segments were correctly grouped, they should induce similar driving speeds, which, in this paper, are approximated using speeds. Their speed distribution should differ from other road types.

4.3.3 Phase 3. Multivariate regression model

It is necessary to benefit from regression techniques to provide insight into the influences of road features on driving speed. The regression model is used in two steps, the last step second phase, and the fourth phase. There are various regression models, but as there is a numerical dependent variable and the relationship between road characteristics and operating speed is linear[15], the multivariate regression model can provide an accurate result. In addition, this technique is straightforward. The weakness of the multivariate regression model is that it is sensitive to outliers

[14]. The Gower Similarity Index is a measure used in statistics to quantify the similarity between two entities (often, two data points or cases) when the data is mixed, meaning it includes both numerical and categorical variables.

[15] . Yang et al. (2019) and Andersen and Reinau (2016) shows there is a linear relationship between road characteristics and driving speed.

and collinearity. As described in section 4.3.1, the features engineering techniques were employed to cope with these issues.

4.4 Results

This section analyzes the results of the developed regression models and proposed data-driven methods to generate the typology of a road network.

4.4.1 Generating regression model for the whole network

Before generating the road typology, the Ordinary Least Square (OLS) is utilized to explore the association between changes in road characteristics and driving speed. To do this, first, according to spatial features, latitude, and longitude, the speed data are integrated with road segments. Then, the road segments and speed data are introduced as independent and dependent variables in the OLS to produce a model.

To develop models all attributes mentioned in Table 4-2 are used to develop the model for entire network segments. First, the categorical variable is converted to the dummy variables. Then, the most deterministic factors are selected using the backward elimination technique. The p-value is considered to determine the significance of features on driving speed. All features with a p-value less than 0.05 significantly influence driving speed. In addition, as discussed earlier, road features have various units, then, to investigate the influence of road features on driving speed, the standardized coefficient is calculated (Kim 2011).

Table 4-2 Details of the base model

Significant Features	Coefficient	Standardized Coefficient	P-values	R2
CN	2.52	0.33	0.00	0.53
ACS	-3.71	-0.01	0.00	
SL	0.63	0.44	0.00	
NL	1.25	0.21	0.00	
DSI	-1.1	-0.17	0.00	
RC	0.001	0.35	0.00	

Table 4-2 Details of the base model

LUA	17	0.65	0.00	
LUCC	-1.49	-0.48	0.013	
LUP	4.39	0.02	0.00	
LUI	-5.08	-0.52	0.00	
LUMA	-2.96	-0.23	0.00	
RG	-2.90	-0.11	0.00	
CN: Constant ACS: Accessibility SL: Speed Limit NL: Number of Lanes DI: Density of Signalized Intersection RC: Radius of Curvature RG: Road Gradient LUA: Land Use Agriculture LUCC: Land Use City Center LUP = Land Use Park LUI: Land Use Infrastructure LUMA: Land Use Mixed Activity				

Table 4-2 indicates the base model's output for the OLS model. As it shows, the R2[16] of the produced model is 0.53, and land use of agriculture and infrastructure have the most significant positive and negative influences on driving speed, respectively.

4.4.2 Generating the road segment typology

According to Figure 4-1, this study uses various approaches to generate multiple typologies rather than only one clustering technique. The main reason for deploying three approaches is to identify the best approach to cluster road segments and generate the most accurate typology. As discussed in section 4.2, most former studies have not categorized road segments according to road characteristics by clustering techniques. On the other hand, due to the road data structure composed of different data types, it is necessary to examine varied approaches to determine which method produces a more accurate result. To do so, three following approaches are developed.

- Integrating the GSI and the k-means technique
- Combining the FAMD and the AHC technique
- Deploying the k-proto technique.

[16]. The R^2 reported is the adjusted R^2

Table 4-3 demonstrates that the road network data sample used in the clustering and road function was excluded from clustering because of a strong correlation with other features.

Table 4-3 Sample of road network data

Segment	SL (km/h)	NL	RG (%)	DSI	RC (m)	LU	ACS
S1	40	1	-3	0	120	park	0.564
S2	50	2	0	1	440	mixed activities	0.453
S3	60	3	2	12	865	industrial	0.787
S4	100	4	5	7	548	residential	0.936
SL: Speed Limit Number of Lane: NL Road Gradient: RG Density of Signalized Intersection: DSI Radius of Curvature: RC Land Use: LU Accessibility: ACS							

Approach 1: combination of GSI and k-means

For Approach 1, the Gower technique generates a similarity matrix in which a single array displays similarity between two segments. The number of segments is 9,409, so a 9,409 by 9,409 matrix is generated. The average overall similarity between segments is 0.23, and the maximum similarity is 0.66. Table 4-4 shows the sample of the Gower matrix.

Table 4-4 A sample of the Gower similarity matrix

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		**9409**
0	0.00	0.17	0.01	0.18	0.02	0.20	0.19	0.19	0.19	0.18	...	0.204
1	0.17	0.00	0.16	0.01	0.17	0.02	0.03	0.03	0.03	0.02	...	0.194
2	0.01	0.16	0.000	0.17	0.010	0.19	0.20	0.20	0.20	0.19	...	0.194
3	0.18	0.01	0.17	0.00	0.16	0.01	0.02	0.02	0.02	0.01	...	0.195
4	0.02	0.17	0.01	0.16	0.00	0.18	0.19	0.19	0.19	0.18	...	0.195

After generating the matrix, the next step is to deploy the k-means clustering technique. As previously discussed, the number of clusters must be pre-specified. To do that, the Elbow method is used [61]. Figure 4-4 illustrates that the optimal number of clusters is determined to be 5.

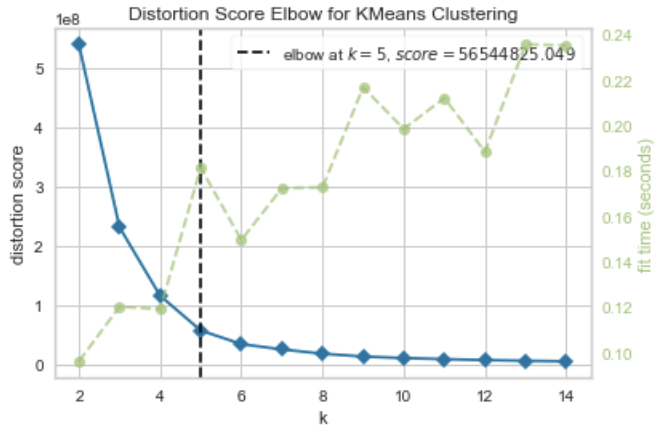


Figure 4-4 The optimal number of clusters; Elbow method[17]

The k-means are performed, and five road types are generated. The essential attributes of the clusters are as follows:

Cluster 1, representing road type 1, contains the fewest number of segments compared to other clusters and involves 965 segments. Regarding the road function, around 67% of the segments are motorways with speed limits of 60 km/h, 70 km/h, and 100 km/h. In addition, these segments are in a flat area with a slope of 0% to 1%. Most of the segments have a signalized intersection density of zero.

Road type 2 mainly contains the secondary road type. All segments are surrounded by residential land use and have speed limits between 30 km/h and 60 km/h.

Road type 3 represents road links in the city center; many have speed limits from 30 km/h to 40 km/h and the highest density of signalized intersections.

Road type 4 is the only cluster that contains all road functions, and all its links are distributed in industrial areas with slopes from -1% to 1%.

The number of segments in road type 5 is the same as in road type 4, which are 1,519 and 1,599, respectively. However, unlike the former clusters, which have one or two land uses, all kinds of land uses can be found in road type 5. Regarding other attributes, gradient values are from -2% to 2%, and the minimum and maximum speed limits are between 40 km/h and 70 km/h.

[17]. The distortion score is a metric that quantifies how far data points in a cluster are from the cluster's center, calculated by summing the squared distances of these points to the center.

The clustering results should be validated to measure how accurately segments are grouped. The conventional indices that the literature confirms are more reliable is Silhouette[18]. This research uses this index . The silhouette value ranges from -1 to 1, with -1 being the worst and 1 being the best. The silhouette value for this typology (approach 1) is 0.571.

Approach 2 Combination of FAMD and AHC

AHC, like k-means, cannot directly analyze categorical values. To cope with this issue, dummy variables must be used to convert categorical features to numerical values. The FAMD technique not only converts categorical values to continuous ones but also reduces data dimension. As discussed in section 4.3.2, the FAMD was employed to overcome this issue. To designate the number of clusters, a dendrogram is produced[19]. Figure 4-5 illustrates its result. Colors in the dendrogram represent the seven clusters. In the final step, AHC is fitted to the data to generate the seven road types.

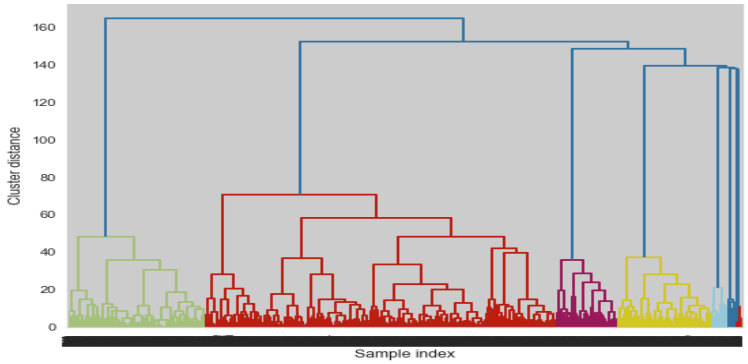


Figure 4-5 Dendrogram to determine the optimal number of clusters

The main drawbacks of this approach are the unbalanced distribution of the segments over the clusters and the heterogeneity of segments in the clusters. For instance, cluster 1 has 76 members; cluster 6 comprises 233 segments; cluster 7 contains 105 segments; cluster 3 has 4,891 segments. In addition, the segments in the clusters are not homogenous. For example, in cluster 1, 30% of segments have a speed limit of 50 km/h while 30% have a speed limit of 100 km/h, and in cluster 2, segments have broad ranges of speed limits from 30 km/h to 100 km/h. In cluster 3, segments are entirely heterogeneous. The density of intersections goes from 0 to 15, segments have gradients from -8% to 8%, and radius values are distributed from 100 m to 1,117 m. The mentioned issues

[18]. (Arbelaitz et al. 2013)
[19] . Using Ward's method

are witnessed in clusters 4, 5, 6, and 7. The statistical analysis of cluster elements confirms that road segments are not grouped most appropriately.

Approach 3: Performing k-prototype

Another technique that, unlike the previous approach, can directly analyze a mixed-type dataset is the k-prototype. This is a partitioning technique, so the user should initialize the number of clusters. As in Approach 2, the Elbow method is used to do that. In Figure 4-6, the elbow of the curve as the number of clusters chosen should therefore be three, and the dashed green line illustrates the required time to train the clustering model per cluster number.

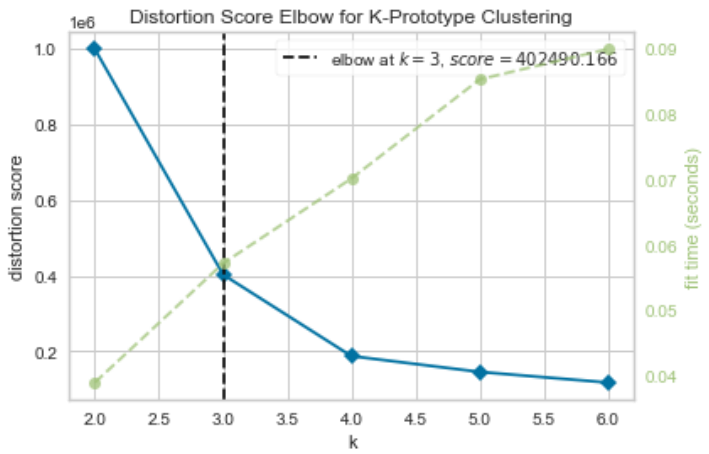


Figure 4-6 Elbow method; k-prototype

The k-prototype was executed, and three road types were generated as follows:

Road type 1 includes approximately 50% of all segments. The road function of more than 70% of the segments is secondary, and 20% are tertiary, with speed limits between 40 km/h and 60 km/h. Moreover, the density of signalized intersections is distributed from 0 to 5, with more than 90% of land in order of importance being residential, mixed activities, and industry. It should be noted that road segments in the city center are not categorized in this cluster.

Road type 2 contains almost 35% of all segments, most of which are in the city center with land use designated as residential. All segments with a density of signalized intersections of more than 10 are in this cluster, and motorways and ramp segments are not categorized in this cluster.

Road type 3 has the lowest number of members at less than 15% of segments. All segments' functions are motorways with speed limits of more than 70 km/h. The density of signalized intersection is less than 3, and the gradient of 97% of the segments is between -1% to 1%.

4.4.3 Validating Clustering Results

There are two main approaches to comparing the performance of clustering techniques: internal and external measures. This study cannot use external measures because they need ground truth labels. Table 4-5 shows the result of internal measures for all approaches.

Table 4-5 Internal cluster validation index

Clustering technique	**Approach 1**	**Approach 2**	**Approach 3**
Silhouette value	0.57	0.45	0.51

In this study, an additional validation step is considered because, in some studies (Hämäläinen, Jauhiainen, and Kärkkäinen 2017), it is discussed that some factors, such as the type of data, can affect the performance of internal measures. Then, to deal with this issue and to increase the reliability of the clustering results, another validation step is employed. To this end, it is assumed that drivers should have similar behaviors on road segments belonging to a road type. Speed is used as a proxy for driving behavior, and speed distributions are compared between clusters. shows the speed distributions for the clusters obtained using the three approaches.

As shown in Figure 4-7, in Approach 2, the speed distribution in clusters 1 and 2 and clusters 6 and 7 are similar, which means the clustering does not allow categorizing groups related to differentiated speeds. However, the speed distribution of different clusters in approaches 1 and 3 are distinct, suggesting a better ability to separate segments. These results are compatible with those of the first step of validation.

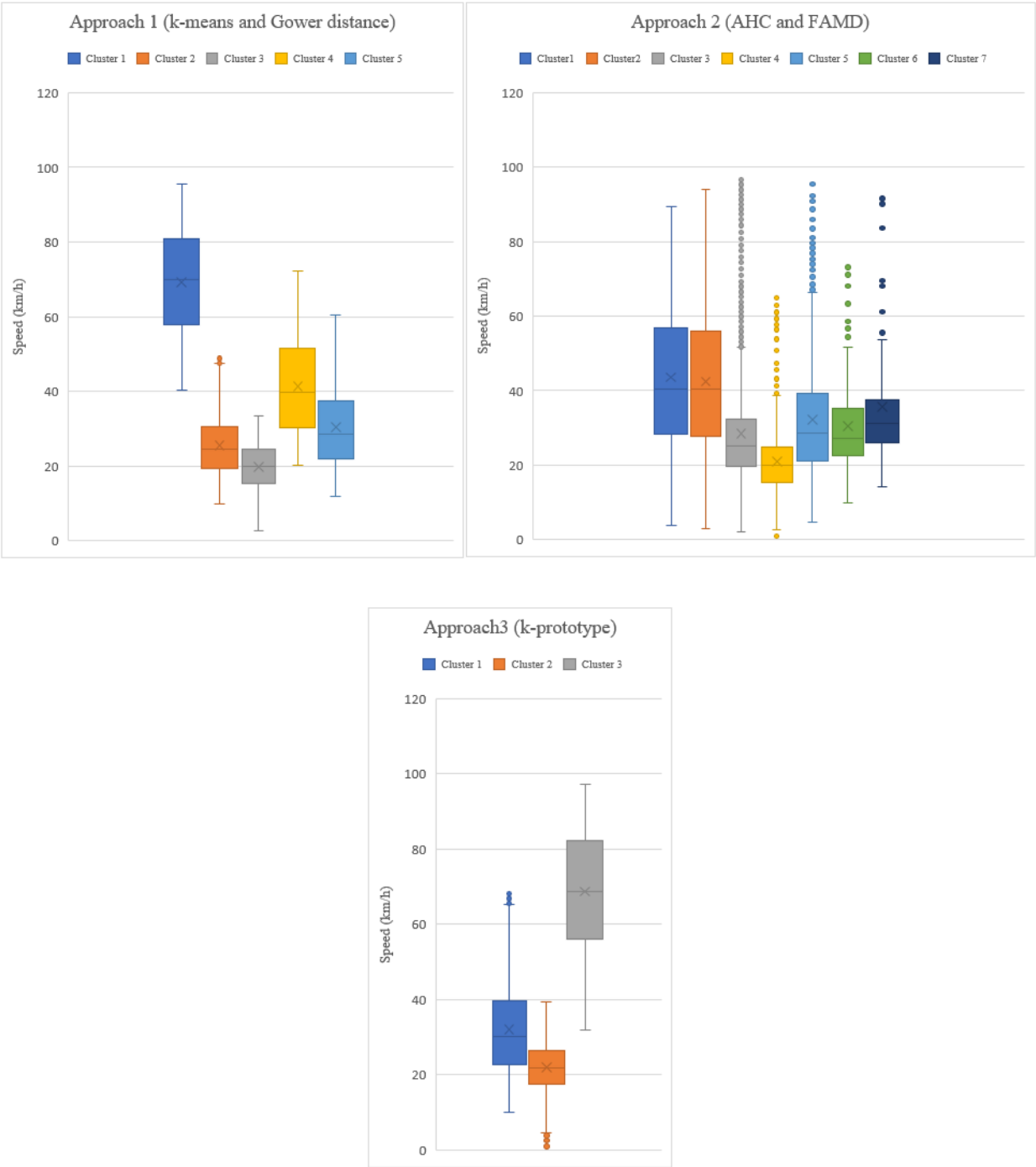


Figure 4-7 Comparison of speed distribution of the three typologies of segments

Table 4-5 and Figure 4-7 confirm that approach 1 produces more accurate results (accuracy here means better splitting segments into groups with more similar behaviors, as approximated by speed distributions). A contingency matrix is used to investigate the similarity between these two

approaches. This matrix summarizes the overlap between these two approaches, and its array denotes the number of objects in common between the two clusters. Table 4-6 shows the contingency matrix between results from approaches 1 and 3.

Table 4-6 Contingency matrix between results from approaches 1 and 3

		Approach 3		
Approach 1		Cluster 1	Cluster 2	Cluster3
	Cluster 1	386	0	102
	Cluster 2	75	848	41
	Cluster 3	198	1239	1403
	Cluster 4	1457	61	81
	Cluster 5	969	483	66

4.4.4 Statistical analysis of the speed distribution of the first approach

To further assess the quality of the typology obtained using the selected techniques (k-means and Gower), the following questions are raised: how homogenous are speeds in a road type, and how distinct are the speeds in the cluster from the speeds from other clusters? Are these distinctions significant? Answering these questions can help better assess the road typologies' quality.

Different statistical tests are employed for this purpose. One of the most famous statistical tests is the One-way Analysis of Variance (ANOVA), and it helps evaluate whether there are any statistically significant differences between the means of two or more independent (unrelated) groups (road types). Two hypotheses are formulated:

H0: All the means of speeds of the road types are equal

H1: At least one mean of speeds of road type is different

To decide whether to reject H0 or not, the p-value is used. If the p-value is less than 0.05, H0 is rejected, and there is a significant difference between the mean speeds of at least two groups.

According to Table 4-7, the p-value is 0, meaning there is a significant difference between the mean speeds of various groups. However, it does not reveal in which groups the speeds are significantly different.

Table 4-7 The result of the ANOVA test

	DF	sum_sq	mean_sq	F	PR(>F)
Speed values	4.0	8.170947e+05	204273.667215	1740.95917	0.0
Residual	8653.0	1.015291e+06	117.333979		
DF: Degree of Freedom, sum_sq: Sum square, mean_sq: mean square, F: F critical value					

To solve this issue, a post hoc technique is used. There are various techniques to perform the post hoc, and selecting the method is up to homogeneity of variances and the number of samples. To check the homogeneity, Levene's test is used. If the result of Levene's test is less than 0.05, it means the variances of groups are not equal and vice versa. The result of Levene's test for speed values was zero, which means the variances of groups are not equal. Therefore, the Games-Howell technique is deployed to perform the post hoc test.

Table 4-8 shows the result of the post hoc analysis. The column Mean Difference (I-J) demonstrates the differences between the mean speeds of road type I and road type J.

Table 4-8 Post hoc result analysis

Clusters (I)	Clusters (J)	Mean Difference (I-J)	Std. Error	Significant
C1	2	49.18	0.87	0.00
	3	44.41	0.89	0.00
	4	26.30	0.95	0.00
	5	33.74	0.91	0.00
C2	1	-49.18	0.87	0.00
	3	44.76	0.22	0.00
	4	-22.88	0.39	0.00
	5	-15.44	0.30	0.00

Table 4-8 Post hoc result analysis

C3	1	-44.41	0.89	0.00
	2	-44.76	0.22	0.00
	4	-18.11	0.44	0.00
	5	-10.67	0.36	0.00
C4	1	-26.30	0.95	0.00
	2	22.88	0.39	0.00
	3	18.11	0.44	0.00
	5	17.44	0.48	0.00
C5	1	-33.74	0.91	0.00
	2	15.440	0.30	0.00
	3	10.67	0.36	0.00
	4	-17.44	0.48	0.00

The results of Table 4-9 show that the speeds in road type 1 have the most significant difference from other road types. In addition, road types 2 and 3 have negligible differences, and the driving speed in road type 2 is greater than that of road type 3. Furthermore, the mean difference of road type 4 demonstrates that, apart from road type 1, the highest speeds occurred in this group. Figure 4-8 illustrates the spatial distribution of road types in the Montreal Island Road network.

Figure 4-8 Montreal Island network typology using approach 1

4.4.5 Investigating the association between road features and driving speed

OLS is again used to develop the multivariate regression model to investigate the relationship between road features of the produced types and vehicle speed. As five road types have distinctive characteristics and driving speed in a road type differs from others, developing a multivariate regression model for each road type provides better insight into the relationship between driving speed and road characteristics. Most importantly, as a combination of road characteristics are different in road types, this approach enables us to explore the effects of changing the combinations of road characteristics. Table 4-9 shows the details of the result of the developed OLS models for road types. Obviously, the models' R^2 for the created road types, other than road type 2, are more than the base model, meaning the engendered road types explain better variability of driving speed. Pertaining to road type 2, further research needs to understand why this is the case. But now, it can

be discussed that this is the only road type where most of its segments are located in residential areas. So, demographic factors such as population density likely significantly affect driving speed (Jing Liu et al. 2020) that, due to data shortage, are not employed in this study.

Table 4-9 Road features and OLS regression model outputs for developed road types

Road type	Significant features	Coefficient	Standardized coefficient	P-value	R2
RT 1	SL	0.46	0.29	0.000	0.82
	NL	6.89	0.19	0.000	
	DSI	-6.23	-0.23	0.000	
	RC	0.005	0.16	0.000	
	LUMA	4.65	0.21	0.000	
RT 2	ACS	-1.80	-0.10	0.000	0.22
	SL	0.52	0.35	0.000	
	NL	2.45	0.17	0.000	
	DSI	-0.62	-0.12	0.000	
	RC	0.003	0.08	0.000	
RT 3	SL	0.06	0.15	0.000	0.79
	NL	0.25	0.08	0.007	
	DSI	-0.12	-0.19	0.000	
	LUP	2.58	0.13	0.000	
RT 4	SL	1.1893	0.63	0.000	0.96
	RG	-0.9093	-0.05	0.007	
	RC	0.003	0.04	0.043	
	NL	1.1559	0.05	0.015	
	DSI	-1.77	-0.11	0.000	
RT 5	ACS	-3.98	-0.18	0.000	0.75

	SL	0.33	0.18	0.000	
	NL	1.11	0.07	0.006	
	DSI	-1.01	-0.19	0.000	
	RC	0.004	0.08	0.000	
CN: Constant ACS: Accessibility SL: Speed Limit NL: Number of Lanes DI: Density of Signalized Intersection RC: Radius of Curvature RG: Road Gradient LUA: Land Use Agriculture LUCC: Land Use City Center LUP = Land Use Park LUI: Land Use Infrastructure LUMA: Land Use Mixed Activity					

For road type 1, the functions of around 92% of segments are motorways and ramps, and their speed limit is between 70 km/h and 100 km/h, so the most influential factor is the speed limit, and the radius of curvature has a minor effect. In addition, the density of signalized intersections in this road type significantly impacts driving speed compared with other groups. Furthermore, around 93% of road links are flat; therefore, the gradient does not considerably influence driving speed.

Regarding road type 2, almost 85% of the road links are secondary and tertiary in residential areas. Hence accessibility and density of intersections negatively affect vehicle speed.

In road type 3, as with other road types, the speed limit has the most notable effect on driving speed, and the density of the signalized intersections has the most significant negative influence on driving speed. In addition, as with road types 1 and 2, the slope value of 80% segments is 1% to zero, and, as a result, its effect on driving speed was negligible.

For road type 4, the speed limit significantly affects driving speed compared to other factors, including road gradient and the density of signalized intersections.

In road type 5, accessibility has the most significant negative influence on driving speed compared with other road types; as with different road types, the speed limit is the leading player.

In conclusion, the model results show road features' effects on driving speed. In other words, changing the combination of road features changes their influence. For example, lane number, density, and the speed limit in different road types have different values, which shows that their effects on driving speed are varied. In addition, the number of influencing factors is changed as well. For instance, the radius of curvature is an influential factor in all road types other than road type 3. Moreover, in all groups, the speed limit significantly influences driving speed compared

with additional features, which is not surprising. Hence, these results help better understand the existence of contradicting results in the literature.

4.5 Conclusion

This paper aimed to contribute to filling three gaps in the literature on the association between road features and driving speed: generating a new typology of the road segments, proposing an approach to validate clustering results in transportation, and developing models to explain better relationships between road features and driving speed.

First, previous studies mainly categorized road segments according to their functions, while literature shows that driving speed and vehicle emissions are influenced by various road features. Hence, this study developed a data-driven framework using machine learning to identify and categorize road segments according to the most critical features affecting driving speed. Clustering techniques were performed on the Montreal Road network as a case study, and the cluster validation techniques show that the combination of k-means and Gower similarity measure produces the most accurate results, leading to the segments being grouped into five road types.

The main advantage of the proposed approach is its flexibility in terms of the scale of the study area; it can be used to classify city road segments regardless of the street network size. Moreover, the developed framework benefits from conventional machine learning techniques to generate road typologies using various street features; therefore, this framework could be a reference to classify road segments in other cities.

However, the main challenge of this step was data scarcity and the quality of data. Due to the magnitude of the city and the number of segments, the preprocessing step was time-consuming.

The second gap is related to validating the clustering results. The literature revealed that the data type could affect conventional clustering validation techniques like Silhouette. Consequently, an additional validation step was developed for vehicle speed data to increase the results' reliability. Pairing road segments generated five groups of speed data in each road type with speed data. The ANOVA, Levane, and post hoc analysis tests confirmed a significant difference between the speeds of each road type.

This study's last contribution relates to developing models to explain better relationship between road features and driving speed. Developing the base model demonstrated that utilizing entire street

networks cannot explain changes in operating speed very well. Producing a regression model for each road type improves R2.

This study's finding proved that the combination of road features should be considered a significant factor in determining the most critical factors. Furthermore, unlike the previous studies introducing some specific features as the most influential factors, which is the primary source of existing contradicting results in the literature, the present research confirmed that the main influencing features differ from one road type to another.

In conclusion, this study shows that the developed framework can identify different road types according to their road features and suggests that as the driving speeds in the road types are different, the driving cycle in each road type could be different too.

CHAPTER 5 ARTICLE 2: IDENTIFYING OPTIMAL SETS OF DRIVING CYCLES IN DIFFERENT DRIVING CONDITIONS USING MACHINE LEARNING

Written by

Asad Yarahmadi, Catherine Morency, Martin Trepanier

Published in February 16, 2023

Journal: World Conference on Transport Research Society (WCRT 2023)

Abstract

Driving cycle is one of the main inputs of vehicle emission modeling. However, the variability of driving cycles due to fluctuations in weather conditions is one of the primary sources of uncertainty in vehicle emission estimation. This study aims to identify representative driving cycles that can correctly represent driving patterns in diverse weather conditions. First, a multivariate multiple regression model is developed to determine the most important weather factors affecting the driving patterns. Then, similar weather conditions are identified according to these factors using unsupervised machine learning. Next, two driving cycles are constructed for diverse weather types, one for weekdays and one for weekends. Afterward, descriptive analysis and a similarity matrix are employed to determine how similar the generated driving cycles are in different weather types. Finally, 15 driving cycles are identified to represent driving patterns in diverse driving conditions.

Keywords: Driving Cycle; Weather Conditions; Vehicle Emissions; Driving behaviors Variability

5.1 Introduction

Vehicle emissions are a major contributor to air pollution and climate change, making it a critical issue affecting the health and well-being of humans and the environment. As the world becomes increasingly dependence on transportation, it is imperative to find ways to resolve the existing uncertainty of estimation of the harmful emissions produced by cars. The exactitude of tailpipe emissions largely depends on the accuracy of the driving cycle.

The driving cycle represents the average driving behaviors in a region and can be characterized by a set of driving speed, acceleration, and braking patterns. It has been widely employed in designating fuel consumption (Achour and Olabi 2016; Ma et al. 2019), assessing vehicle performance (Degraeuwe and Weiss 2017; Ladubec et al. 2021), optimizing EMS (Energy Management System) (Chen et al. 2019; Wu et al. 2023), and estimating driving style (Rios-Torres, Liu, and Khattak 2019). Driving cycles most importantly associate drivers' behavior with exhaust emissions in vehicle emissions models like MOVES (MOtor Vehicle Emissions Simulator) (Bodisco and Zare 2019; EPA 2014; Muhammad 2022).

Numerous studies have examined the influences of the main weather-based parameters, such as precipitation, temperature, wind, and visibility, on drivers' behaviors (Jägerbrand and Sjöbergh 2016; Kilpeläinen and Summala 2007). Findings indicate that motorists often adapt their speeds in response to changing weather conditions, which in turn affects fuel consumption (Abdi Kordani et al. 2018; Andrey et al. 2013; Chen et al. 2019; Chowdhury 2015; Christian and Jensen 2014; Elsagheer Mohamed et al. 2022; Faria, Baptista, and Farias 2017; Zhang et al. 2018). However, the analysis of weather's impact on the driving cycle parameters has been infrequently explored by researchers. With the increase in global temperatures and climate change, choosing the sets of driving cycles that best reflect the weather conditions and the resulting vehicle emissions is essential.

Generally speaking, there are two types of driving cycle including: Standard Driving cycles (SDC) and Local Driving Cycle (LDC) (Gebisa et al. 2021). A thorough analysis of existing knowledge reveals that the existing driving cycle production approaches are beset with two major flaws. First, previous investigations tend to inadequately incorporate adverse weather conditions, such as heavy rainfall and severe winds, into the process of collecting driving data and constructing driving cycles. For example, SDCs predominantly consider climatic parameters that affect the engine's

thermodynamic efficiency, often disregarding adverse weather conditions that may affect the drivers' behavior. Similarly, Worldwide Harmonized Light-duty Vehicle Test Procedure (WLTP) methodology prioritizes temperature and humidity (United Nations 2014), the Canadian standard driving cycle, 5-cycle test, incorporates a restricted range of temperatures (-7°C, 35°C, and 20°-30°C), the USA FTP-75 test scrutinizes a narrow ambient temperature spectrum 20-30°C (Eckert et al. 2018). Second, prior studies have commonly utilized one or few driving cycles to represent driving behaviors across different meteorological conditions, neglecting the variations in driving behaviors under diverse environmental scenarios. For instance, Ma et al. (2019) synthesized two driving cycles, the off-peak and peak hours, for Beijing. Chauhan, Joshi, and Parida (2020) developed one driving cycle to mimic driving behaviors for Vadodara City. Kaymaz, Korkmaz, and Erdal (2019) constructed one driving cycle to represent driving patterns in Istanbul. This oversimplification potentially engenders uncertainties in the estimation of vehicle emissions.

To consider the influences of variability of weather conditions on driving behavior, there are two principal challenges. Firstly, weather conditions exhibit significant variability over the course of a day and year, and these changes can impact driving patterns. For instance, in the city of Montréal, Canada, the temperature can fluctuate from -30°C to 35°C throughout the year and over 20°C in a single day, while humidity ranges from 20% to 100% annually. On the other hand, the reactions of drivers to these weather conditions depend on a variety of factors, including demographic characteristics (such as age and gender), vehicle attributes (such as type and condition), and driving skills (such as experience and training) (Yeo, Lee, and Jang 2021). Due to the complex interplay of these factors, identifying the number of representative driving cycles that accurately capture the driving behaviors volatility influenced by weather patterns is a challenging task. Secondly, the arduous process of generating driving data to encompass all driving scenarios and a diverse range of drivers is time consuming and cost prohibitive. As a result, the produced LDCs very often rely on inadequate data, obtained within a limited timeframe and involving a restricted number of drivers (Ahmed Al-Samari 2017), as evidenced by studies conducted by (Cui et al. 2022; Zhai et al. 2018).

The principal objective of this study is to devise an approach for identifying the representative driving cycles that capture driving patterns under diverse weather conditions. To achieve the desired objective, this study collected comprehensive driving and weather datasets to cover various drivers’ behaviors and different weather conditions.

As weather conditions affect driving cycle variables, the number of driving cycles in a region can be a function of number of weather types. Therefore, it is crucial to disentangle the effects of weather factors from those of road features on driving cycle parameters (Abele and Møller 2011; Ragione and Giovanni 2016). This is achieved by collecting driving data on a specific road type selected based on the findings of our precedent study (Yarahmadi, Morency, and Trepanier 2023b).

This study examines the following question:

- How many driving cycles are required to represent driving behaviors in different weather conditions?
- What weather parameter(s) have the most significant impact on driving cycle parameters?

It presents three notable contributions aimed to advance the understanding and optimization of vehicular performance:

- Firstly, a systematic approach is proposed to establish a statistically significant relationship between weather conditions and the attributes of driving cycles.
- Secondly, a methodology is proposed to generate and identify the representative driving cycles for diverse driving scenarios, thereby facilitating the development of efficient and effective transportation systems.
- Thirdly, the impacts of weather parameters on driving cycle characteristics are isolated from the influence of road features.

The remainder of this paper is structured as follows. First, previous studies that explored the relationship between driving conditions and driving cycles are discussed, and their arguments are exposed. Then, the methodology section describes the technical steps to find representative sets of driving cycles and their various parameters are elaborated. Finally, the results are discussed, and some perspectives are proposed.

5.2 Related Work

The background section comprises two distinct segments. The initial segment entails a critical evaluation of the prevailing methodologies that incorporate meteorological conditions in the development of driving cycles. Subsequently, the second segment scrutinizes the existing body of research that explores the impact of weather variables on driving behaviors parameters. Collectively, these components enable the discernment of pertinent research lacunae.

5.2.1 Driving Cycle Generation and Weather Conditions

The current section is divided into two subsections to comprehend the research problem and identify the knowledge gap. Firstly, the conventional approach of incorporating the impact of weather conditions in the development of driving cycles is reviewed. Secondly, the current practices for selecting the most representative driving cycle are discussed in detail.

The SDC is a process approved and applied by authorities to evaluate and certify a model of vehicle that complies with all performance standards (relative to emissions) before authorizing it to be introduced to the market (European Commission Joint Research Center, 2013). SDCs are commonly utilized to evaluate and compare vehicle emissions across various regions, whether at a tertiary or global level. In order to ensure their applicability and relevance for different regions, SDCs must be designed with general conditions in mind, encompassing a wide range of driving scenarios, traffic patterns, and environmental factors. This broad approach allows for a more inclusive and representative assessment of vehicle emissions, ultimately contributing to developing effective policies and standards for emission control worldwide. SDCs predominantly concentrate on the parameter of ambient temperature and do not inherently emulate other meteorological conditions like precipitation, wind velocity, humidity, or visibility. The principal justification for this approach is rooted in the necessity of establishing a regulated, reproducible setting for the comparative analysis of various vehicles. The incorporation of additional variables such as precipitation or wind speed into these assessments would introduce further fluctuations, complicating the task of contrasting vehicles under uniform, standardized conditions. Table 5-1 illustrates that the established SDCs often fail crucial weather factors that can impact vehicle emissions.

Table 5-1 SDCs and weather parameters

Name of SCD	Country	Weather parameters
New European Driving Cycle	European Union	Temperature: 20-30°C
WLTC	European Union United Kingdom Turkey, China, India	Temperature: (23°C ± 5°C) Temperature: -7°C Relative humidity: 40-60%

Table 5-1 SDCs and weather parameters

Highway Fuel Economy Test (HWFET)	USA	Temperature: 20-30°C
Japanese JC08 Cycle	Japan	Temperature: 20°C ± 5°C
KDC	Korean	Temperature: 20-30°C
India Drive Cycle (IDC)	India	Temperature: 25°C average
CB	Brazil	Temperature: 20-30°C
ADR 81/02	Australia	Temperature: 20°C ± 5°C

Multiple studies have underscored a discrepancy between emissions estimates derived from real-world data and those obtained through Standard Driving Cycles (SDCs). For example, the International Council on Clean Transportation (ICCT) reported a growing divergence in this disparity in Europe, from 8% in 2008 to 39% in 2017 (Greene et al. 2017). In order to address the limitations, Local Driving Cycles (LDCs) are developed for specific regions, incorporating real-world driving data to synthesize these cycles (Seers, Nachin, and Glaus 2015). As a result, LDCs provide more accurate estimations of vehicular emissions compared to Standard Driving Cycles (SDCs) (Anida et al. 2017; Esteves-Booth et al. 2001; Fotouhi and Montazeri-Gh 2012).

However, acquiring a comprehensive dataset encompassing all variables influencing vehicular conduct for developing a representative local driving cycle necessitates a considerable investment of time and financial resources. Consequently, the researchers very often concentrated their efforts on capturing region-specific traffic attributes to ascertain representative driving cycle parameters utilizing limited data sets. At the same time, exogenous factors such as meteorological conditions are often neglected.

Liu et al. (2020) collected data for 15 consecutive days to establish a representative driving cycle for buses in China, which may not be sufficient to capture the variability in driving conditions. Galgamuwa, Perera, and Bandara (2016) identified a representative driving cycle for an expressway in Sri Lanka by collecting 56 hours of driving data over three days, which may not be comprehensive enough to account for different driving conditions. Pouresmaeili, Aghayan, and Taghizadeh (2018) gathered driving data for two weeks under clear weather and dry road

conditions in Mashhad, Iran, to develop a representative driving cycle, but the limited scope of data collection may not consider the full spectrum of driving variability.

In addition, Simulation-based methodologies, while useful in controlled settings, may not accurately capture the complex interplay of factors present in real-world scenarios, such as the impact of weather conditions on driving dynamics. For example, the study by (Amirjamshidi and Roorda 2013), which utilized simulated data to construct driving cycles for light and heavy trucks, failed to incorporate the influence of meteorological factors on driving behavior, thus limiting the ecological validity of their findings. As a result, the LDCs constructed for the same region have varied drivers' behavior representation.

Furthermore, the second gap is related to the authors' technique for selecting representative driving patterns. Indeed, unlike the first group, some studies gathered the required data in different seasons but eventually picked only one driving cycle as a representative.

Ashtari, Bibeau, and Shahidinejad (2014) believe that insufficient data and ambient conditions are two factors that generate uncertainty in vehicle emissions. Accordingly, they gathered 44 million driving data points for one year to develop a driving cycle in Winnipeg, Canada. Finally, they invented a new snipped selection algorithm to find the best candidate driving cycle. Ma et al. (2019) developed 17000 driving cycles but selected two as the most representative driving cycle for peak and off-peak hours. Günther et al. (2017), over one full year, collected around 27365 hours of driving operation data in Hamburg, Germany, to produce driving cycles for public buses. Based on this vast data, the authors developed a driving cycle.

5.2.2 Combined Effects of Weather Parameters and Driving behaviors Study

The impact of weather conditions on driving behaviors have attracted interest from researchers in numerous transportation domains, such as safety, traffic flow, driver behavior analysis, roadway design, intelligent transportation systems, and climate change adaptation. However, weather-related influences on vehicular emissions have by far received less attention. Specifically, according to the authors' knowledge, the nexus between weather parameters and driving cycle parameters remains underexplored.

Reviewing the existing methods to investigate the influences of weather conditions on driving behaviors revealed that most of the previous research predominantly focused on the effects of

single or limited weather parameters on driving behaviors and should have considered the impact of other external factors like road features in their studies. Neglecting the synergistic interactions of these parameters, referred to as combined effects, may lead to conflicting findings. For instance, Broughton, Switzer, and Scott (2007) classified fog into two levels—light and dense—demonstrating a 19% decrease in driving speed during light fog and a 33% decrease during dense fog. However, it is unclear which road functions and design attributes they considered in their study. Concurrently, Yan et al. (2014) observed negligible speed differences between light and dense fog in straight road segments. While numerous studies indicate that rainfall results in speed reduction, Bijleveld and Churchill (2009) asserted that the impact of precipitation on driving speed varies according to the season. Furthermore, Bakhshi et al. (2022) employed statistical analysis of questionnaires to establish that speed choice during rainfall depends on demographic factors such as marital status and age. Zhang et al. (2018) confirmed that travel speed in snowy conditions is contingent on road function (e.g., highway, collector) and congestion level. Furthermore, Yuhui, Yuan, and Huibao (2018) utilized ANOVA tests to demonstrate that driving speed adaptation in adverse weather, including fog and rain, is influenced by the vehicle type—passengers and truck. Lastly, Jägerbrand and Sjöbergh (2016) emphasized that daylight and posted speed limits are crucial factors in determining driving speed changes under varying weather conditions.

Hence, to precisely ascertain the substantial impacts of weather conditions on operating speed, a comprehensive approach that considers both external factors, such as road characteristics, and the intricate interactions among all weather parameters is imperative.

5.3 Research Methodology

This segment presents the data utilized in the present investigation, elucidating the methodological involved. Additionally,

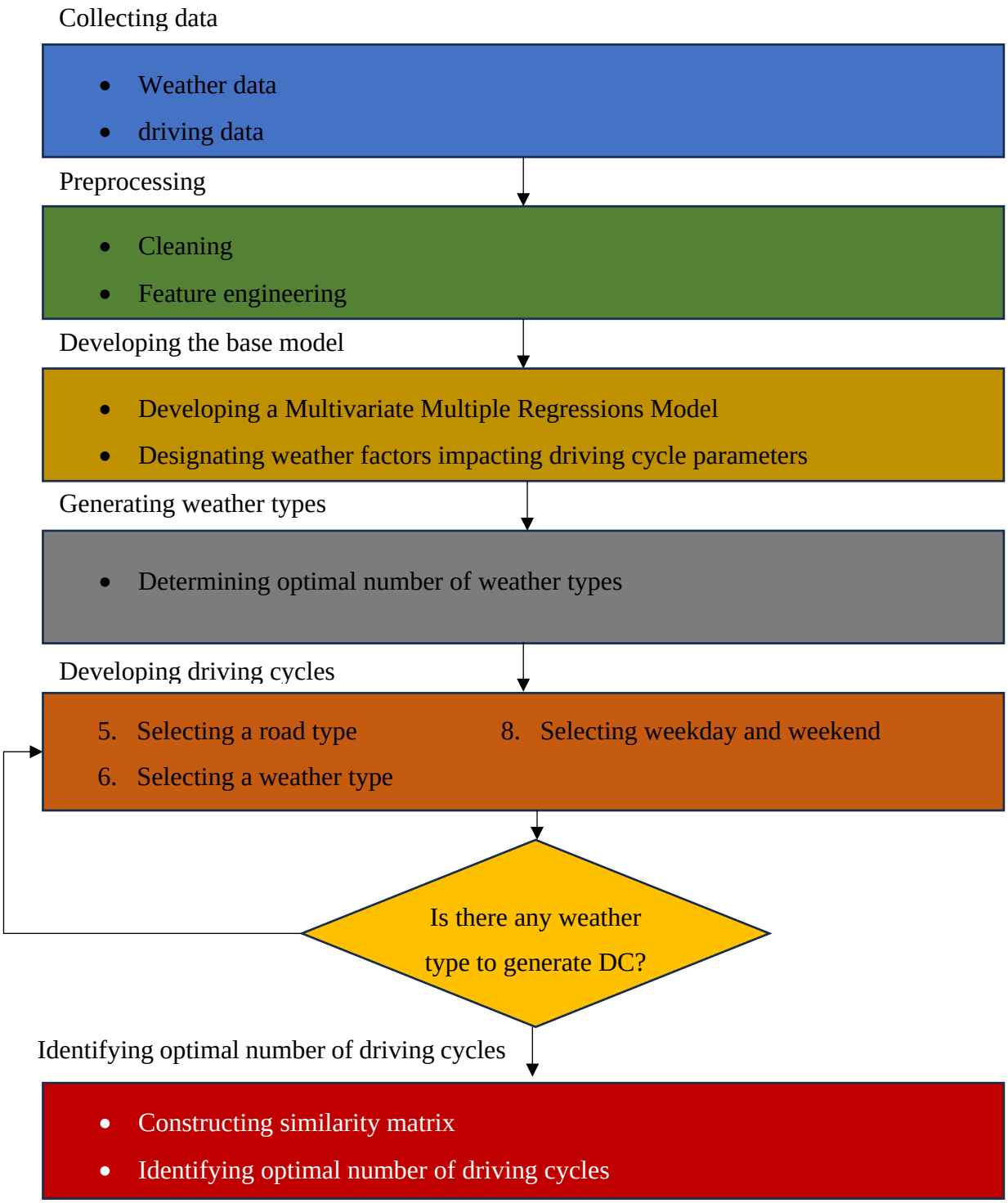


Figure 5-1 visually delineates the stages undertaken during the course of this study. This study encompasses a comprehensive six-step approach, explained in detail in sections 0 to 5.3.5. As articulated, this research aims to determine the representative number of driving cycles pertinent to diverse driving conditions. To that end, crucial data about weather and driving are gathered and prepared in the first and second steps. Subsequently, a sensitivity analysis identifies the critical meteorological factors significantly influencing driving patterns. Weather types are subsequently generated in the next step based on these factors. In the fifth stage, a sensitivity analysis is carried out to explore how the proposed model effectively captures the complex relationship between weather conditions and driving patterns. Finally, two driving cycles are generated for each weather type, corresponding to weekdays and weekends.

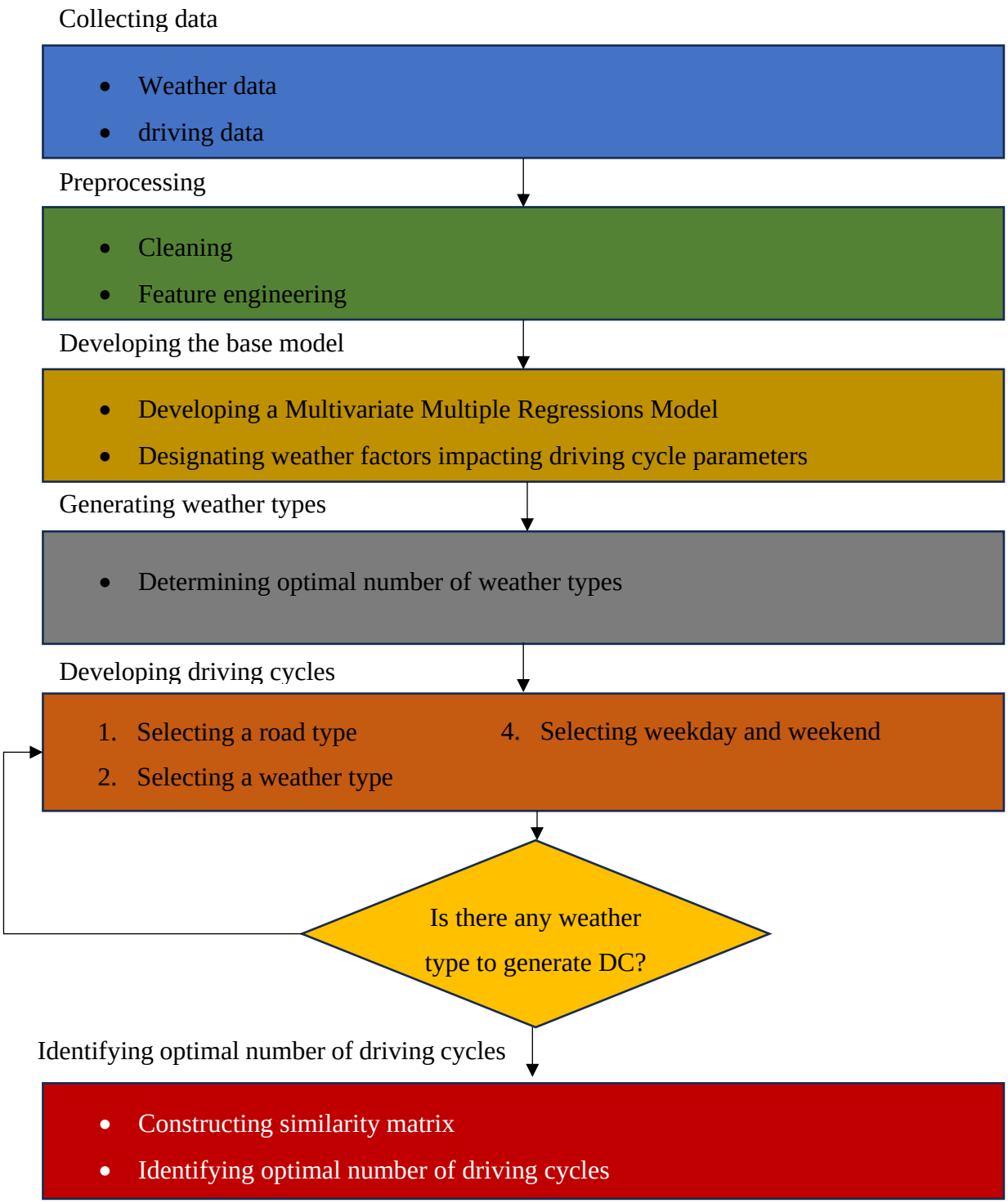


Figure 5-1 Research methodology of the present study

5.3.1 Study Area and Data Description

As Figure 5-1 illustrates, the initial phase of this study, detailed in (Yarahmadi, Morency, and Trepanier 2023b), involved classifying Montreal's road network into five types based of factors affecting driving speed. Subsequently, this paper focuses on one road type, comprising 2608 segments, to examine the impact of various weather conditions on driving cycle parameters. Table 5-2 provides a detailed overview of the defining characteristics of the road segments within this selected category.

Table 5-2 Characteristics of the selected road type

Road Feature	Values	Proportion of values in the database
Road Functions	Motorway	25.3
	Primary	26
	Secondary	43
	Tertiary	3.7
	Ramp	2
Land uses	City Center	4.1
	Industry	25.1
	Mixed Activities	27.7
	Residential area	39.1
	Park	4
Speed Limit (km/h)	50	55
	60	10
	70	35
Number of Lanes	1	1
	2	24.5
	3	69
	4	5.5
Gradient (%)	-1	3.8

	0	41.2
	1	25
	2	18
	3	12
Density of Signalized Intersections*	0	23.2
	1	21
	2	27.7
	3	14.1
	4	8
	5	6
Accessibility	Min: 0.1 Average: 0.42 Max: 1.5	-
Radius of curvatures (m)	Min: 102 Average: 455 Max: 1587	-
The density of signalized intersections illustrates the number of intersections located in a grid which cell size is 500m*500m		

5.3.2 Driving Data

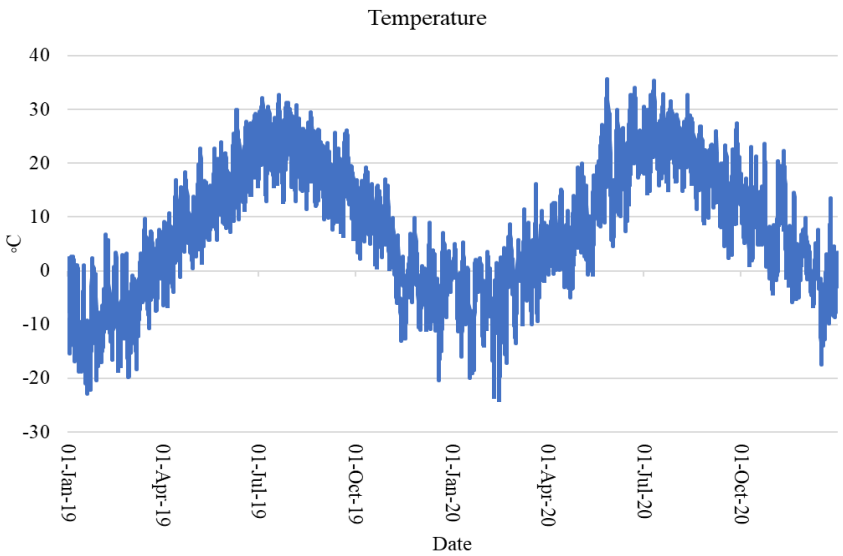
To conduct this study, a fleet of 3,323 connected taxicabs produced a vast dataset consisting of approximately 3,227,000 trips undertaken between April 2019 and December 2020 under genuine driving conditions, including both peak and off-peak periods when the cabs were in service (only trips with passengers onboard are used). The characteristics of the driving data are presented in Table 5-3. Additionally, the driving data was not selectively gathered based on ancillary attributes such as the vehicle type, driver's age, or gender. Thus, this issue lets authors suppose the data likely originated from driving data belonging to various age groups, genders, and vehicle passenger types.

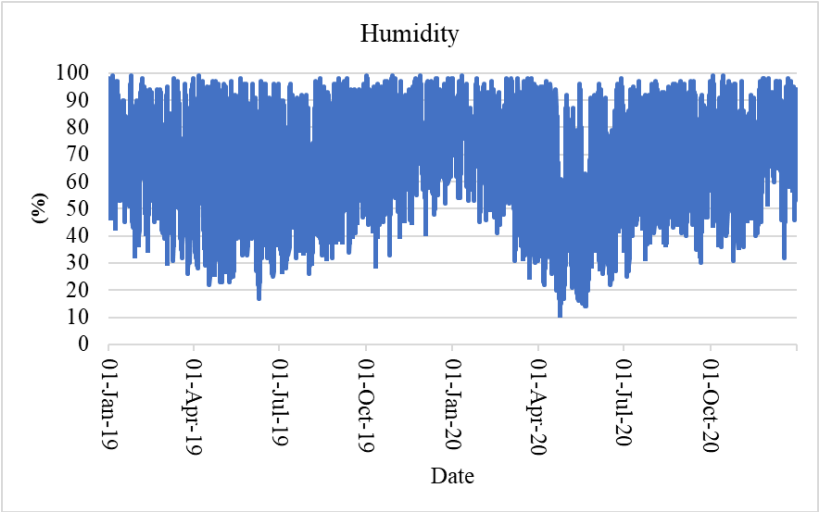
Table 5-3. Driving data features

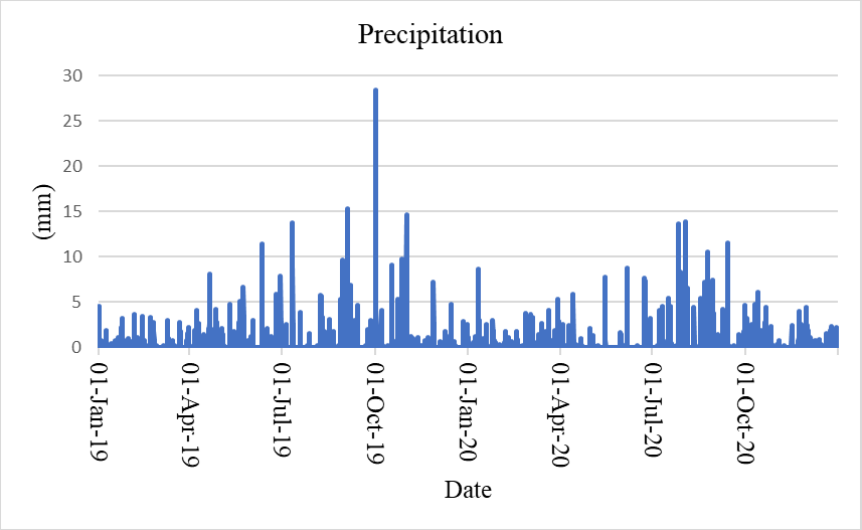
Variable	Timestamp	Speed	Longitude	Latitude
Type	Date	Number	Number	Number

5.3.3 Weather Data

Montreal's climate is continental, with frigid winters and warm summers (Wood 2015). Hourly basis weather data was collected from Environment Canada for 2019 and 2020. The main weather parameters are temperature, precipitation, wind speed, and humidity. Figure 5-2 indicates the variation of weather parameters over 2019 and 2020.







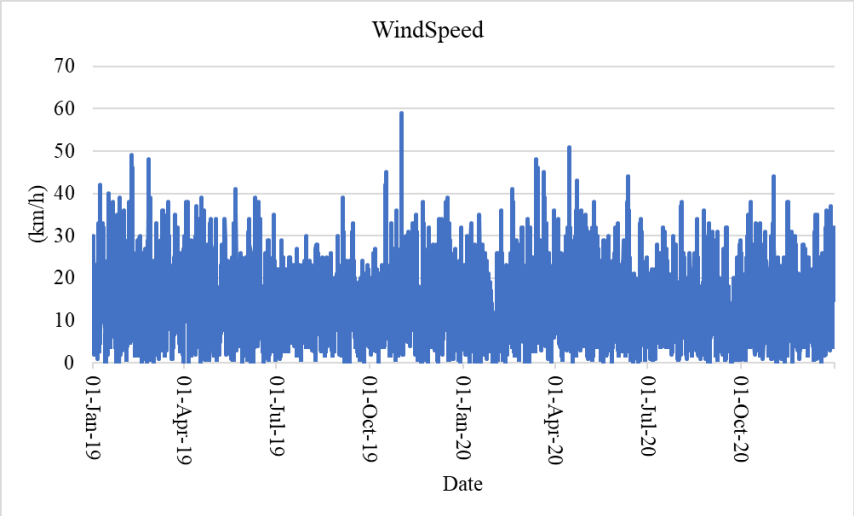


Figure 5-2 Variation of temperature, humidity, precipitation, wind speed over the years (2019-2020)

5.3.4 Data Preprocessing

The second step of this study is data preprocessing. Outlier analysis and data cleaning techniques elucidate unnecessary data in both datasets. Afterward, two steps are taken to prepare the data: generating required driving data attributes (data enrichment) and feature engineering. Each phase is explained in the following sections.

Data Enrichment

Data enrichment is a step to generate new items from existing data to make it ready for analysis. To create a driving cycle, driving data should be generated per second. It is because driving modes are generated based on a one-second interval. Investigating the raw data showed that data are mainly available in 5-second intervals or more. To address this issue, a resampling technique is used to increase the time interval granularity from 5 seconds to 1 second. However, while this adjustment increases the frequency of time stamps, it results in null values for other driving data parameters, such as driving speed, latitude, and longitude, at the newly created time stamps. Literature shows that the Kalman filter is a powerful tool to impute null values in time series data (Ribeiro 2021). To validate the results obtained from the Kalman Filter, the imputed data were examined using QGIS. This process involved a comparative analysis where both the pre-imputation dataset and the post-imputation dataset, resampled at a one-second interval, were loaded into QGIS. By overlaying these datasets in the QGIS environment, I was able to visually assess the accuracy of the imputations. Specifically, I compared the imputed positions against known data points to evaluate the consistency and reliability of the imputed values. This approach allowed for a detailed examination of how well the Kalman Filter preserved the overall patterns and characteristics of the original data while filling in the gaps. In the final step, the distance between each point is calculated using the Geopy library in Python and the Haversine formula. The Haversine formula is a very accurate mathematical technique to calculate the distance between two points on the surface of a sphere.

The most critical weather factors are made available by Canada Environment, and there is no additional feature that the authors can calculate. Although visibility is another vital weather parameter that affects driving patterns, it is not available at the hourly level. Unavailability of some data is one of the limitations faced by this study.

Feature Engineering

Feature engineering combines techniques to make input data compatible with machine learning algorithms and reduce computational complexity (Chen, Zhang, and Wei 2014).

Regarding the weather data, the first feature engineering step is correlation analysis. As its name implies, the correlation between segment variables is explored in this step, and highly correlated variables, referred to as multicollinearity, are deleted. A feature is strongly correlated if its

correlation coefficient is more than 0.7 with another part (Ratner 2009). To perform the correlation analysis, the Dython library in Python is utilized. This library is based on the correlation ratio (Lewandowski and Cooke 2007). The output illustrates that none of the driving and weather data were correlated.

So, the second step of feature engineering was to evaluate the distribution of every segment feature. To do that, the skewness of features was measured. Skewness shows how symmetric data distribution is. Skewness can help identify potential issues in the data, such as outliers, non-normality, or extreme values. In addition, some clustering algorithms assume that the data follow a specific distribution, often a multivariate normal distribution. Assessing skewness can help to verify whether the data meet these assumptions and to select appropriate clustering techniques. If the skewness values of a column exceed 0.5, it is skewed. The SciPy library in Python programming is used to calculate the skewness. The result shows that the precipitation and wind speed were highly and slightly skewed respectively. To treat and analyze skewed data there are three methods comprising: implementing transformations, utilizing nonparametric methodologies, or synergizing robust statistical methods with analytical procedures. In the first step, transformation techniques are used to cope with skewness. The square root transformation yielded superior results for wind speed data, whereas the quantile transformation performed better for precipitation data. Despite this, the distribution of precipitation data remained skewed.

Another area for improvement that could affect the accuracy of machine learning technique output is the uniformity of units of continuous unscaled data. For instance, the unit of precipitation is centimeters, the wind speed is m/s, and so on. To resolve this matter, all weather parameter values were standardized. To achieve this, each value is standardized by subtracting the mean value of that feature and then dividing by the standard deviation. As a result, the standardized values typically have a distribution with a mean of 0 and a standard deviation of 1.

Dimensionality reduction, the final feature engineering technique applied in this study, serves a critical purpose beyond simply identifying and omitting variables with minimal variance. It addresses the curse of dimensionality, where high-dimensional spaces can lead to sparse data distributions and make clustering analyses less effective due to increased computational complexity and difficulty in identifying meaningful patterns. Principal Component Analysis (PCA), a widely recognized dimensionality reduction technique, is particularly valuable in this

context. PCA transforms the original variables into a new set of uncorrelated features, known as principal components, which are linear combinations of the original variables weighted by their contribution to the total variance. By constructing a covariance matrix and determining its eigenvectors and eigenvalues, PCA quantifies the variance each feature contributes to the dataset. The eigenvectors represent the directions of maximum variance, while the corresponding eigenvalues indicate the magnitude of variance along these directions. The selection of principal components is based on the cumulative explained variance ratio derived from these eigenvalues, ensuring that the reduced dataset retains the most significant variance within the data. This process not only streamlines the feature space, enhancing computational efficiency, but also improves the interpretability of clustering by focusing on the most informative aspects of the data. By mitigating noise and reducing overfitting risks, PCA allows for a more robust and meaningful clustering outcome that can reveal hidden structures within the data, which might otherwise be obscured in a higher-dimensional context. It is confirmed that the performance of PCA is reduced if data is not normally distributed. On the other hand, Robust PCA (RPCA), one of the extensions of PCA, shows better performance in skewness cases (Candes et al., 2009). In the present study, PCA is used for parametric clustering techniques, while RPCA is used for non-parametric clustering techniques.

Multivariate Multiple Regression Model (MMRM)

It is necessary to benefit from regression techniques to provide insight into the relationship between weather parameters and the driving cycle. In addition, its result can be interpreted to understand how the generated model can describe the relationship between independent and dependent variables. In addition, its result can be interpreted to understand how the generated model can describe the relationship between independent and dependent variables. There are various regression models, but the main reasons for selecting MMRM are as follows:

Simultaneous analysis: Given the presence of 13 numerical dependent variables in the data, MMRM permits a simultaneous examination of the relationships between the multiple predictor and outcome variables. This allows us to holistically understand the interactions within the datasets.

Revealing complex relationships: MMRM is capable of uncovering intricate relationships and interactions among variables, which may be overlooked if separate univariate analyses are conducted. This results in a more nuanced understanding of the data.

Combined effects analysis: MMRM facilitates the investigation of combined or collective effects of predictor variables on the dependent variables. This provides insights not just into individual influences, but also into how these predictors might work together to influence the outcomes.

The weakness of the multivariate multiple regression model is that it is sensitive to outliers and collinearity. As described in section 0, the features engineering techniques were employed to cope with these issues.

Cluster Analysis

Clustering, a form of unsupervised learning, constitutes a robust analytical technique aimed at categorizing data into distinct groups based on shared attributes. By maximizing intra-cluster similarities and concurrently accentuating inter-cluster dissimilarities, this approach effectively delineates homogeneous subsets within heterogeneous datasets, providing valuable insights into the underlying structure and relationships among observations.

Clustering techniques can be tailored to different data types, such as parametric, nonparametric, and mixed datasets. Nonparametric approaches, such as Affinity Propagation make no distribution assumptions and can handle diverse data types, including non-Gaussian or skewed distributions. Although k-means is categorized as a nonparametric technique, its performance can be less than optimal for skewed data due to its assumptions about cluster shapes, sensitivity to outliers, and reliance on minimizing variance. For heterogeneous datasets comprised of both skewed and non-skewed attributes, two approaches are recommended for effective clustering: (1) employing robust preprocessing techniques to prepare the data, followed by clustering, and (2) utilizing nonparametric clustering methods. To do that, the three following approaches are taken:

- Preparing data by robust PCA and then performing K-means.
- Employing Affinity Propagation (AP)
- Agglomerative Hierarchical Clustering (AHC)

In this study, weather types serve as the foundational basis for generating driving cycles corresponding to each weather category. The accuracy of these generated weather types significantly impacts the precision of the derived driving cycles. To ensure a comprehensive understanding of the relationship between weather conditions and driving cycles, this research employs various techniques to generate diverse weather types, acknowledging that each method

has its unique advantages and disadvantages. Utilizing multiple approaches allows for the generation of diverse weather types, which can be compared and contrasted to identify the most accurate and representative categories. Subsequent cluster validation helps ascertain the most reliable results, ensuring the credibility and relevance of the generated driving cycles.

Agglomerative Hierarchical Clustering (AHC) and Affinity Propagation (AP) clustering methods that do not require pre-specifying the number of clusters. k-means, on the other hand, is a fast, user-friendly algorithm that performs well with continuous variables. However, its primary limitation is the need to determine the number of clusters a priori, which can often result in incorrect assumptions by the user. To address this issue, various techniques can be employed to determine the cluster count. The following section provides a concise overview of existing methods for this purpose, offering an understanding of the approaches available to determine cluster numbers.

Determining the Best Compromise Number of Clusters

Establishing the ideal number of clusters is a critical aspect of clustering analysis, as it significantly influences the outcome's quality and the algorithm's ability to discern patterns. In addition, selecting the appropriate number of clusters contributes to mitigating overfitting and underfitting, increasing the comprehensibility of results, and optimizing computational efficiency. Striking this balance is essential for revealing concealed structures and patterns within the dataset, ultimately ensuring a reliable and robust model performance. Xu et al. (2016) documented various techniques to determine the number of clusters.

The well-acknowledged techniques to determine the number of clusters are as follows:

Elbow method: This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and selecting the number of clusters at the "elbow" of the curve where adding more clusters no longer significantly reduces the WCSS.

Gap statistics: This method compares the within-cluster dispersion for different numbers of clusters to a reference distribution and selects the number of clusters that maximizes the gap between them.

A dendrogram is a tree-like diagram that illustrates hierarchical clustering by depicting the relationships among data points. It enables an intuitive understanding of the data's organization and helps determine the number of clusters by examining linkage distances.

In the current research, the optimal number of distinct weather classifications is dependent on the ideal number of clusters; hence, all previously mentioned methodologies are employed to ascertain this number of clusters. Additionally, in order to determine the 'K' parameters, the authors took into account various criteria such as the size of the dataset, the nature of the data, the objectives of the analysis, and pre-existing knowledge in the field.

5.3.5 Driving Cycle Construction

Driving cycle are constructed using variation of the driving speed over time. There are four main steps to generate a driving cycle: microtrip generation, clustering of the generated microtrips, constructing candidate driving cycles, and selecting the most representative driving cycle. These steps are further detailed below.

Microtrip Generation

A microtrip is a small part of a trip, and the sum of all microtrips builds a trip. In the microtrip generation step, driving pattern parameters are calculated for each microtrip and are used to compare all microtrips.

There are various approaches to generating a microtrip. The most famous ones are as follows:

- Trip-based microtrip: microtrip is bounded by two stops.
- Fixed time interval microtrip: a trip is divided into a time intervals.
- Fixed distance microtrip: a trip is divided based on a fixed distance.
- Intersection microtrip: a microtrip is bounded by two intersections.

Nouri and Morency (2017) generated driving cycles using the above-mentioned methods and determined that the fixed distance of 250 meters provides superior accuracy compared to other techniques. Thus, for the purposes of this study, the driving cycle was generated according to this method.

Microtrip Clustering

After generating microtrips, the next step is to find similar driving patterns. To do that, by comparing microtrips together, a similarity matrix is generated. Then, the matrix is clustered using the clustering technique to identify and group similar driving patterns into a cluster.

Constructing Candidate Driving Cycles

To construct a driving cycle, Markov chain modeling is performed to generate a driving cycle from produced microtrips using a transition matrix. The main output of this step is several driving cycles.

Selecting the Most Representative Driving Cycle

All candidate driving cycles are ranked based on assessment criteria to select the most representative driving cycle. The assessment criteria used in this study are presented in Table 5-4Table 5-4 Driving cycle assessment criteria.

Table 5-4 Driving cycle assessment criteria

Assessment criteria	Denoted	Unit
Difference in SAFD	NA	(%)
The average speed of the entire driving cycle	V	(km/h)
Average running speed	Vr	(km/h)
Maximum speed	Vm	(km/h)
The average acceleration of all acceleration phases	Acc	(m/s2)
Average deceleration of all deceleration phases	Dcc	(m/s2)
The average number of acceleration-deceleration changes	Acc_std	NA
Root means square acceleration.	Acc2	(m/s2)
Road power	NA	(kW)
Time proportion of idle mode	Idle_p	(%)
Time proportion of acceleration mode	Acc_p	(%)
Time proportion of cruising mode (−0.1 m/s2 < acceleration < 0.1 m/s2; average speed > 5 m/s)	Cru_p	(%)
Time proportion of deceleration mode (acceleration ≤− 0.1 m/s2)	Dcc_p	(%)

Table 5-4 Driving cycle assessment criteria

Time proportion of creeping mode (−0.1 m/s2< acceleration < 0.1 m/s2; average speed < 5 m/s)	Cre_p	(%)

The assessment criteria can be categorized into three groups: velocity indicators, acceleration and deceleration indicators, and driving mode indicators. Driving mode refers to the manner in which a vehicle is being driven at a given moment, as determined by its speed, acceleration, and deceleration patterns. Commonly recognized driving modes include a percentage of idling (when the engine is running, but the vehicle is stationary), acceleration (when the vehicle is speeding up), deceleration (when the vehicle is slowing down), cruise (when the vehicle is moving at a steady speed), and creeping (when the vehicle is moving at a very slow speed). Monitoring and analyzing driving modes can provide valuable insights into a range of factors related to vehicle performance, fuel efficiency, and environmental impact. To generate the driving cycle in this study, the application developed by Roy and Morency (2020) is used.

5.4 Results and Discussion

In this section, the results of the developed regression models are assessed and discussed, and clustering approaches are implemented to recognize various weather types.

5.4.1 Multivariate multiple regression analysis before generating weather types

In pursuit of discerning the influence of weather conditions on driving cycle parameters, attributes specified in Figure 5-2 and Table 5-3 are employed for a sensitivity analysis. As producing driving cycles for all hourly data is not possible, the initial step necessitates judicious use of sampling techniques to extract a subset from the weather data.

The rationale for applying the method to a subset of weather data stems from the computational and practical challenges associated with generating driving cycles for each hour of the two-year period. This process is not merely data-intensive but also computationally demanding due to the complexity and granularity of the driving cycle generation process. Given these constraints, a representative sample was necessary to facilitate a feasible analysis while still capturing the variability of weather conditions over the selected period.

To do that, the FTestAnovaPower function from Statsmodels library in Python, which is "*based on Cohen's f as effect size measure*", is leveraged, setting effect size, alpha, and power at 0.25, 0.05, and 0.8 respectively. In fact, these parameter values were chosen to detect a medium effect size

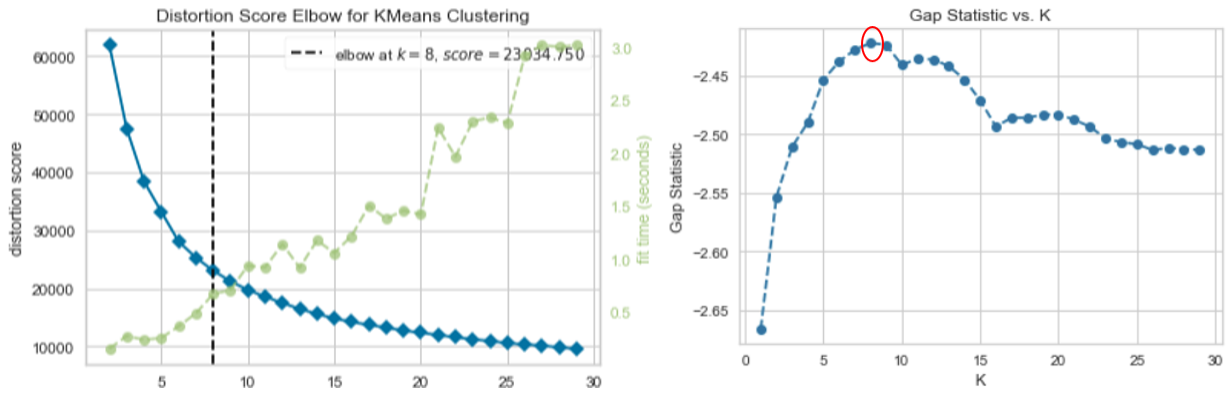
with a conventional risk of Type I error and a high probability of avoiding a Type II error. We used this test just to determine the number of sample sizes. To ensure that selected samples offer a fair representation of the dataset throughout the year, the stratified sampling approach is adopted. This method facilitates breakdown of the population into homogeneous subsets, termed strata, and consequently allows for random selection of samples from each stratum. Thus, weather data is segregated into four distinct strata, each embodying a particular season. Consequently, for the longer seasons of winter and summer, we selected more samples. For the shorter seasons of spring and fall, fewer samples were chosen. In total, we selected 472 samples. Subsequent to this, all prerequisites of the MMRM approach, including the assumption of the normality of residuals, are meticulously examined. Following these confirmatory steps, driving cycles are synthesized based on the sampled weather data dates. Then, developed driving cycles and weather sample data are integrated for further analysis. Eventually, MMRM model was created, and deterministic factors were designated. The R^2 of the generated model is 0.28.

5.4.2 Generating Weather Types

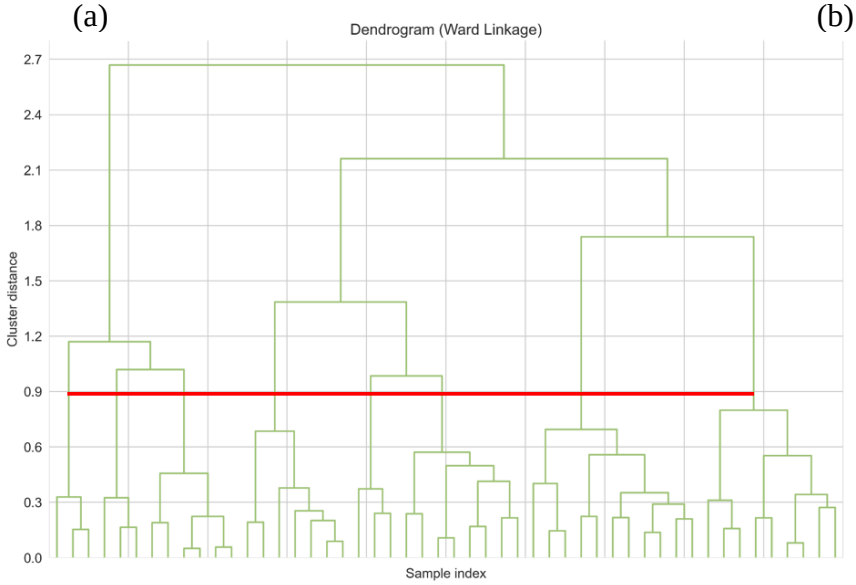
Clustering methods consider the influential factors outlined in 5.4.1 to categorize similar weather conditions. As elaborated in section 5.4.1, three distinct strategies are utilized to cluster weather variables, thereby producing diverse weather typologies.

Determining the Best Compromise Number of Clusters

Three prevalent techniques—Elbow, Gap Statistics, and Dendrogram—were employed to ascertain the number of clusters for the weather dataset. The Elbow and Gap Statistics methods were applied to the k-means algorithm, while the Dendrogram approach was utilized for AHC. As illustrated in Figure 5-3, all three techniques consistently identified 8 clusters as the ideal number for this dataset. In the Elbow method (a), the blue line represents the cluster numbers, and the dashed green line demonstrates the training time required for the clustering model per cluster number. In the Gap Statistics method (b), the number of clusters is highlighted by a red circle. Lastly, in the Dendrogram method (c), the red line indicates the cut line utilized to determine the cluster count.



(a) (b)



(c)

Figure 5-3 Determination of cluster numbers using elbow (a), gap statistics (b), and Dendrogram

In the dendrogram produced by the AHC, the placement of the red line for the cutoff is determined by identifying the largest gap. Gap means distance between the clusters being merged.This is done by examining the vertical distances between successive linkage levels, which represent the dissimilarity measure between clusters being combined. The largest gap method selects the point where the increase in this distance is the most significant. At this point, the clusters below the line are distinctly different from each other, as evidenced by the large jump in cluster distance. Positioning the cutoff at this largest gap ensures that the derived clusters are as homogeneous as possible internally, while the dissimilarity between different clusters is maximized. This method provides a data-driven and objective criterion for deciding the number of clusters, thereby

minimizing subjectivity in the clustering process. In our analysis, the red line is placed at the height corresponding to this largest increase in dissimilarity, suggesting a natural division in the dataset and indicating a number of clusters that best captures the inherent structure of the data.

The elbow method is applied by plotting the k-means distortion scores—sums of squared distances within clusters—against a range of k values. The number of clusters is indicated by the 'elbow' point, where the graph bends and further increases in k yield diminishing returns in decreasing distortion. This point reflects a balance between complexity and clustering quality. In the provided graph, the elbow is at k = 8, suggesting it as the number of clusters based on the method's criteria for parsimony and sufficient data fit.

Approach 1: k-means and RPCA

As discussed in section 0, among weather features, precipitation is skewed yet, and RPCA addresses this issue by using a robust estimation method to identify the principal components. It decomposes the data matrix into a low-rank matrix (capturing the main structure of the data) and a sparse matrix (representing the outliers or noise). This decomposition helps to separate the true underlying structure of the data from the effects of the skewed distribution and outliers, leading to a more accurate representation of the principal components. To perform the RPCA, a library, namely “pypropack,” is used in Python. The output of RPCA is a NumPy array that keeps 95% variability in the original dataset. Table 5-5 shows numbers of observations in each cluster.

Table 5-5 Number of observations in each cluster

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**
1509	2542	1707	2277	2082	2433	1877	3059

The descriptive analysis of the generated clusters revealed the main attributes of clusters as follows:

1. In weather type 1 most temperatures are distributed from 12 to -5 ºC, humidity is between 20 and 50 percent, wind speed is from 28 to 41 km/h, and precipitation is not observed.
2. Weather type 2 mainly represents driving conditions in March, April, October, and November. Approximately 90 percent of the hour’s temperature were between 6 to -12

ºC, humidity spread from 25 to 50 percent, one millimeter of precipitation on average presence, and wind blow between 34 to 39 km/h.

3. Weather type 3 represents summer months with 40 percent, has the highest share. Highest temperature, lowest wind speed, and humidity are found in this cluster. In addition, 94 percent of hours have zero precipitation.
4. In cluster 4, the temperature is distributed from -1 to 11 ºC, humidity spreads from 32 to 65 percent, precipitation was between 0 and 0.7 millimeters. Maximum wind speed, which is 35 to 59 km/h, is observed in this weather type.
5. In weather type 5 share of January, February and December are 27%, 30%, and 28%. Temperature is between -2 and -23 ºC, humidity is distributed from 31 to 62 percent, wind speed is 35 to 52 km/h, and precipitation is between 0.1 and 1.5 millimeters.
6. Cluster 6 composed of October, November, and March, shows the highest precipitation with temperatures between 10 and -5 ºC.
7. Cluster 7 contains temperatures between 2 and -10 ºC and humidity between 35 to 60 percent.
8. January and February have the highest shares in cluster 8. It has the lowest temperature, highest humidity, and precipitation between 0.3 to millimeters.

Approach 2: Affinity Propagation

As mentioned in section 5.3.4 Affinity Propagation (AP) is a nonparametric clustering method that, unlike K-means, does not require the initial specification of the number of clusters. Upon examining the outcomes, it was found that AP identified nine distinct weather patterns. As Table 5-6 presents the main drawback of the generated weather types, is the unbalanced number of observations in each cluster. For example, cluster 2 has around 3000 members, while cluster 9 includes fewer than 500 objects. Moreover, some clusters contain a wide range of values. For example, in clusters 4 and 5, temperature values start from -23 and -15 ºC respectively and reach 30 ºC. Similarly, clusters 1, 2, 8, and 9 humidity comprise both lowest and highest values, from 10 to 96%. Furthermore, other than clusters 3, 5, and 6, all clusters have zero precipitation values.

Table 5-6 Number of observations in each cluster in AP method

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
1831	2993	2603	1062	2886	2099	1464	2129	419

Approach 3: Agglomerative Hierarchical Clustering

AHC (Agglomerative Hierarchical Clustering) is effective with skewed data due to its bottom-up, hierarchical clustering approach, which exhibits less sensitivity to the distribution of data points. This attribute makes AHC more resilient to outliers and data skewness. However, it is essential to note that AHC might not always be the best choice for all types of skewed data. More specifically, it can suffer drawbacks, such as being computationally expensive for large data sets.

As Figure 5-3 shows, based on the analysis of the dendrogram, it has been determined that eight clusters provide an appropriate categorization for the various weather conditions.

Analyzing clusters shows that the AHC does not identify weather types accurately. For example, in clusters 1, 4, 6, and 8, the temperature is distributed from -20 to 20 ºC, humidity is between 10 and 100 percent, and wind speed starts from minimum at 10 km/h, to the maximum speed at 59 km/h. On the other hand, in clusters 3 and 5, temperature spreads from -15 to 30 ºC, precipitation is between 0 and 10 millimeters, and wind speed is 0 to 50 km/h. Table 5-7 shows the number of observations in each cluster.

Table 5-7 Number of observations in each cluster in AHC method

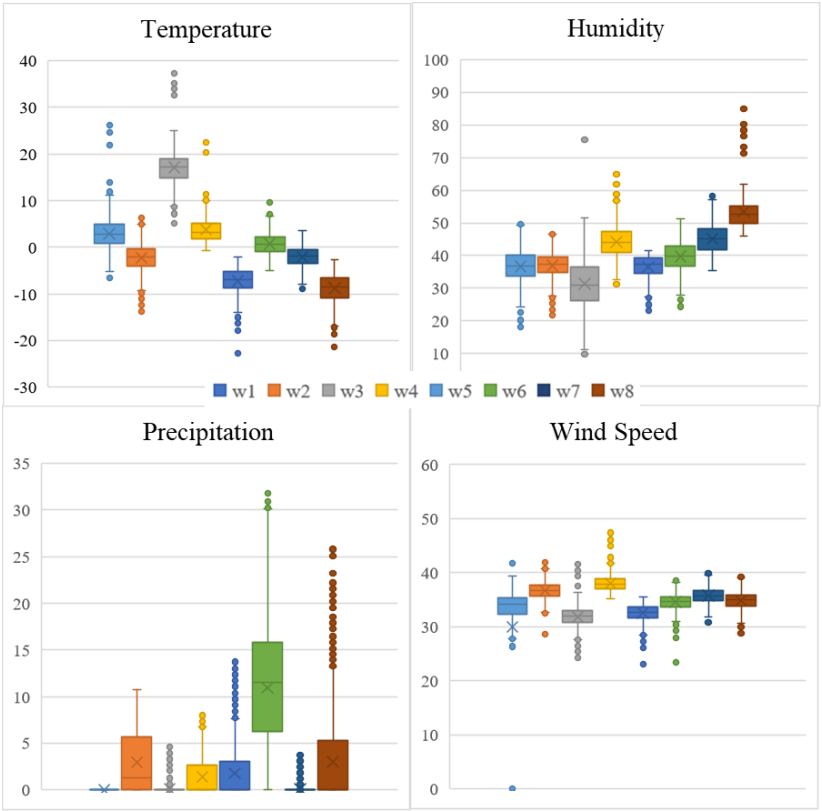
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
2118	2661	3050	2196	2354	1558	2579	970

Cluster Validation

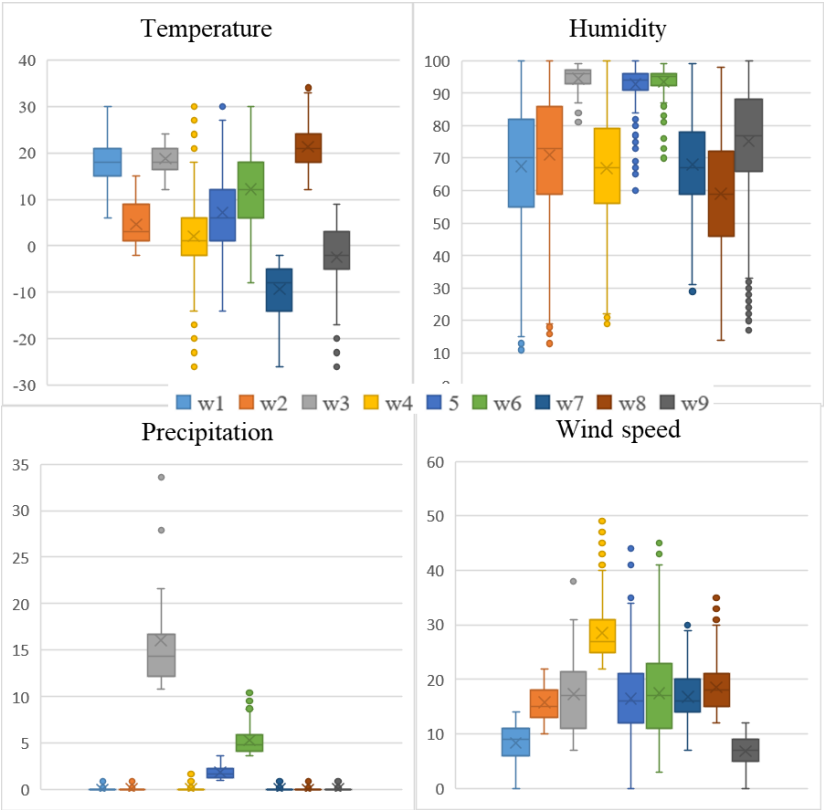
After generating clusters, it is essential to verify the accuracy of the results. One effective method of validating clustering outputs is through content-wise visualization techniques, such as boxplots, which statistically analyze and compare the generated clusters. A boxplot provides a clear and concise way to understand the spread and distribution of data by summarizing the minimum and

maximum values, interquartile range (IQR), and potential outliers. In addition, this tool helps gain insight into how parameter values are distributed among different clusters.

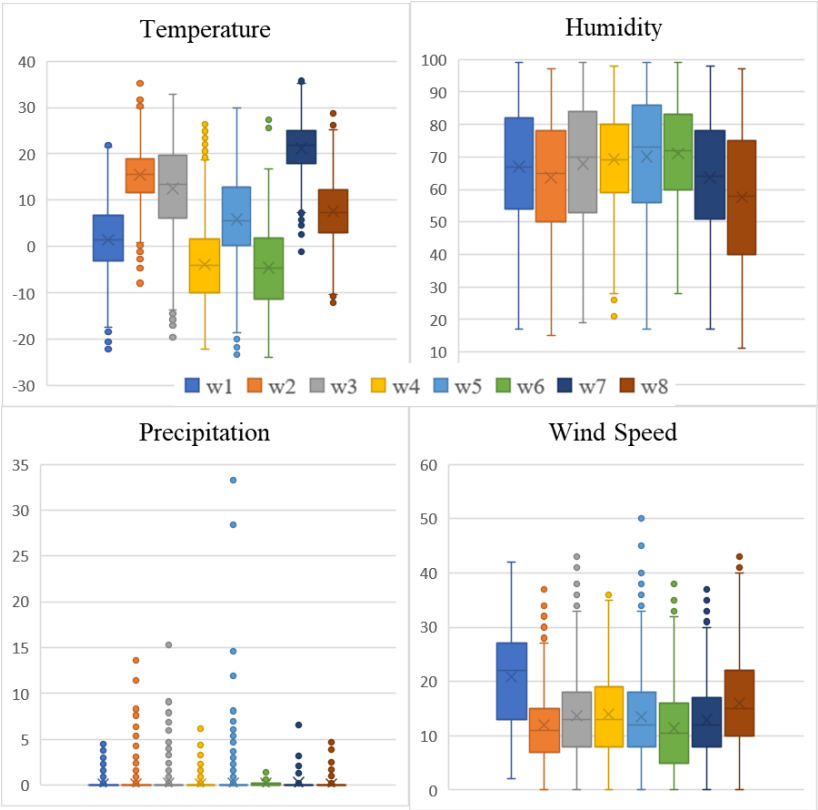
Figure 5-4 compares cluster analysis for three different approaches using boxplots. By examining the distribution of data in various clusters and the range of variations for different clusters, for each weather parameter, it can be concluded that approach 1 provides more accurate weather types.



(1)


Temperature
Humidity
Precipitation
Wind speed
w1 w2 w3 w4 5 w6 w7 w8 w9

(2)


Temperature
Humidity
Precipitation
Wind Speed
w1 w2 w3 w4 w5 w6 w7 w8

(3)

Figure 5-4 Comparing clustering results of approach 1, approach 2, approach 3

The clustering results indicate that the first approach generates better clustering result. Therefore, in the next step of the study, authors use the result of this approach..

5.4.3 Multivariate multiple regression model after generating weather types

By grouping similar weather conditions, clustering methodologies offer a clearer depiction of how weather conditions relate to changes in driving cycle parameters. Applying regression to this clustered data enhances this understanding. To achieve this, sample weather data and constructed driving cycles, as detailed in section **Error! Reference source not found.**, are organized based on the clustering outcomes. It is worth mentioning that cluster 5 contains the least number of observations, a total of 50, while cluster 1 hosts a maximum of 80 observations.

R^2 serves to measure the extent to which independent variables can explain variations in dependent variables. Table 5-8 compares the R^2 values of the model developed for various clustered weather types against a base model. This base model corresponds to the one formulated for the sample data prior to clustering.

As illustrated, except for weather type 6, all other types exhibit higher R^2 values than the base model. This suggests that the model with clustered weather types provides a more accurate representation of the relationship between variations in weather conditions and changes in driving cycle parameters.

Table 5-8 Comparison R-square between clustered weather types and base

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Base**
R^2	0.43	0.53	0.56	0.33	0.53	0.28	0.30	0.35	0.28

Table 5-9 outlines the results of the developed MMRM models. This table presents standardized coefficients to illustrate the impact of weather parameters on driving cycle attributes, along with p-values that indicate the significance of the associations between weather parameters and driving cycle characteristics. Cells shaded in gray within the table indicate that their p-values are greater than 0.05. This suggests that the respective weather characteristics do not have a significant association with the driving cycle parameters. In the table T stands for temperature, H stands for

humidity, W stands for wind speed, and P denotes precipitation. The primary insights drawn from this table include:

1. The influence of weather types on driving cycle parameters is not uniform, likely due to the combined effects of different weather elements.
2. The number of driving cycle parameters which are influenced by weather factors varies across weather types. Specifically, weather type 7 has a lesser impact on driving cycle characteristics compared to other types.
3. Precipitation typically has a significant effect on driving cycle attributes. However, it is absent in weather types 1 and 7, and thus does not influence the driving cycle parameters in these categories.
4. Temperature has a negative impact on driving cycle attributes for all weather types, with the exception of weather type 3, where its influence is not negative.
5. The effects of humidity and wind speed on driving cycle parameters differ across weather types. As an illustration, humidity positively affects velocity in weather type 2 but negatively in weather type 3.

Table 5-9. MMRM models output for the developed weather types

DCP/ Weather		**Cluster 1**		**Cluster 2**		**Cluster 3**		**Cluster 4**		**Cluster 5**		**Cluster 6**		**Cluster 7**		**Cluster 8**	
		Coef.	**P-Value**	**Coef.**	**P-Value**	**Coef.**	**P-Value**	**Coef.**	**P-Value**	**Coef.**	**P-Value**	**Coef.**	**P-Value**	**Coef.**	**P-Value**	**Coef.**	**P-Value**
V	T	-0.88	0.00	-0.78	0.00	0.65	0.00	-0.80	0.00	-0.78	0.00	-0.84	0.00	-0.79	0.00	-0.88	0.00
	H	-0.09	0.11	0.21	0.02	-0.33	0.00	-0.06	0.39	-0.03	0.74	-0.10	0.20	-0.15	0.19	0.05	0.63
	P	-0.02	0.65	-0.18	0.03	-0.03	0.80	-0.15	0.05	-0.12	0.05	-0.64	0.00	-0.05	0.05	-0.68	0.00
	W	-0.09	0.12	-0.03	0.00	0.26	0.01	-0.12	0.01	0.09	0.36	0.07	0.38	0.00	0.99	0.00	1.00
Vr	T	-0.87	0.00	-0.77	0.00	0.63	0.00	-0.79	0.00	-0.69	0.00	-0.79	0.00	-0.76	0.00	-0.88	0.00
	H	-0.08	0.14	0.21	0.02	-0.32	0.01	-0.06	0.41	-0.06	0.59	-0.12	0.17	-0.14	0.25	0.04	0.69
	P	-0.02	0.76	-0.20	0.03	-0.07	0.56	-0.16	0.05	-0.03	0.04	-0.72	0.00	-0.02	0.05	-0.82	0.03
	W	-0.10	0.10	-0.02	0.79	0.26	0.01	-0.12	0.10	0.08	0.51	0.05	0.62	0.00	0.99	0.00	0.98
Vm	T	-0.76	0.00	-0.68	0.00	0.60	0.00	-0.38	0.01	-0.64	0.00	-0.59	0.00	-0.24	0.04	-0.47	0.00
	H	0.00	0.98	0.23	0.04	-0.27	0.03	-0.18	0.18	0.04	0.75	-0.08	0.60	-0.01	0.94	-0.05	0.80
	P	-0.07	0.36	-0.19	0.07	-0.01	0.96	-0.20	0.05	-0.04	0.03	-0.25	0.13	-0.17	0.03	-0.68	0.03
	W	-0.11	0.20	0.12	0.30	0.02	0.82	-0.26	0.05	0.00	0.98	0.06	0.71	0.00	0.99	-0.04	0.71
V-std	T	-0.17	0.00	-0.17	0.04	0.15	0.00	-0.28	0.02	-0.18	0.04	-0.14	0.01	-0.17	0.03	-0.08	0.05
	H	-0.23	0.53	0.25	0.02	-0.23	0.35	-0.16	0.51	-0.11	0.24	-0.07	0.51	-0.05	0.21	0.08	0.33
	P	-0.11	0.42	-0.38	0.28	-0.23	0.40	-0.07	0.05	-0.09	0.03	-0.14	0.01	0.03	0.04	-0.18	0.00
	w	-0.18	0.23	-0.02	0.00	0.31	0.21	-0.02	0.04	0.19	0.32	0.11	0.41	0.2	0.85	0.00	0.09
Acc	T	-0.21	0.04	-0.42	0.00	0.64	0.00	0.41	0.01	-0.15	0.33	0.43	0.00	0.23	0.13	-0.40	0.00

	H	-0.10	0.41	0.22	0.14	-0.05	0.72	-0.22	0.15	-0.13	0.42	-0.02	0.86	0.03	0.89	-0.25	0.04
	P	-0.02	0.88	-0.08	0.05	-0.10	0.46	-0.09	0.05	-0.03	0.04	-0.57	0.00	0.33	0.10	-0.48	0.04
	W	-0.08	0.55	0.05	0.76	-0.10	0.43	-0.26	0.05	0.08	0.63	-0.04	0.78	0.00	0.99	0.07	0.54
Dcc	T	-0.50	0.00	-0.79	0.00	-0.62	0.00	0.41	0.01	-0.79	0.00	0.31	0.05	0.25	0.05	0.48	0.00
	H	0.26	0.02	0.01	0.95	-0.17	0.20	-0.17	0.27	0.05	0.59	-0.09	0.52	-0.25	0.17	0.16	0.40
	P	-0.11	0.31	0.13	0.02	-0.08	0.55	0.05	0.04	0.08	0.03	0.47	0.00	0.04	0.84	0.32	0.04
	W	-0.03	0.78	0.09	0.40	0.01	0.92	-0.13	0.38	0.09	0.37	-0.02	0.90	0.00	0.99	0.07	0.52
Acc2	T	0.51	0.00	0.62	0.00	0.69	0.00	0.01	0.02	0.92	0.00	0.09	0.05	0.11	0.46	0.11	0.38
	H	-0.22	0.04	0.07	0.63	0.06	0.65	0.09	0.60	0.20	0.01	0.11	0.47	0.06	0.76	0.17	0.45
	P	0.08	0.48	0.00	0.99	0.31	0.02	-0.01	0.05	-0.04	0.63	-0.25	0.02	-0.16	0.43	-0.22	0.03
	W	-0.04	0.73	-0.19	0.17	-0.11	0.32	-0.06	0.04	-0.03	0.66	0.02	0.88	0.00	0.99	0.03	0.82
Acc_ std	T	-0.51	0.00	-0.63	0.00	-0.56	0.00	-0.29	0.07	-0.68	0.00	0.11	0.01	-0.13	0.40	-0.52	0.00
	H	0.17	0.11	0.07	0.60	-0.31	0.02	-0.10	0.50	0.13	0.26	-0.06	0.67	-0.03	0.89	0.23	0.22
	P	-0.12	0.26	-0.05	0.71	0.00	1.00	0.01	0.95	0.06	0.60	-0.28	0.04	0.16	0.43	0.19	0.29
	W	-0.07	0.54	0.14	0.30	0.04	0.71	-0.11	0.46	0.04	0.72	0.09	0.57	0.00	0.99	0.03	0.75
Idle _p	T	-0.54	0.00	-0.75	0.00	-0.63	0.00	-0.25	0.0	-0.76	0.00	0.54	0.00	-0.21	0.01	-0.44	0.00
	H	0.17	0.10	0.10	0.33	-0.24	0.05	-0.19	0.22	0.06	0.56	-0.06	0.66	-0.15	0.44	0.11	0.57
	P	-0.12	0.25	0.13	0.02	-0.01	0.96	0.06	0.05	0.08	0.04	0.38	0.00	0.03	0.87	0.73	0.05
	W	-0.03	0.77	0.18	0.10	-0.01	0.96	-0.17	0.26	0.13	0.19	0.00	0.98	0.00	0.99	0.05	0.64
Acc _p	T	0.81	0.00	0.79	0.00	0.64	0.00	0.59	0.00	0.78	0.00	-0.57	0.00	0.66	0.00	0.73	0.00
	H	0.03	0.66	-0.10	0.30	0.36	0.00	0.05	0.66	0.47	0.05	-0.10	0.42	0.34	0.01	-0.06	0.67
	P	0.01	0.85	-0.20	0.04	0.16	0.13	-0.22	0.05	-0.14	0.04	-0.02	0.05	0.10	0.44	-0.16	0.03
	W	0.00	0.99	-0.10	0.33	-0.32	0.00	0.12	0.29	-0.16	0.11	-0.07	0.58	0.00	0.99	-0.01	0.91
Dcc _p	T	-0.26	0.02	-0.67	0.00	-0.69	0.00	0.16	0.03	-0.85	0.00	0.14	0.03	-0.17	0.23	-0.12	0.03
	H	0.24	0.04	-0.11	0.40	0.24	0.08	0.09	0.58	-0.11	0.20	0.15	0.28	-0.43	0.02	0.02	0.93
	P	0.02	0.86	-0.08	0.05	-0.20	0.13	-0.20	0.02	-0.01	0.04	-0.18	0.02	-0.07	0.71	-0.02	0.02
	W	0.08	0.55	0.23	0.09	-0.04	0.72	0.18	0.24	-0.10	0.27	-0.20	0.17	0.00	0.99	0.03	0.77
Cru _p	T	-0.42	0.00	-0.59	0.00	-0.80	0.00	0.02	0.90	-0.85	0.00	-0.25	0.09	-0.15	0.03	-0.09	0.04
	H	0.22	0.04	-0.33	0.02	0.05	0.67	-0.07	0.67	-0.02	0.84	0.19	0.19	-0.15	0.44	-0.40	0.07
	P	-0.03	0.82	0.14	0.02	-0.16	0.17	0.04	0.04	-0.08	0.03	0.05	0.03	0.04	0.85	-0.36	0.04
	W	0.08	0.49	-0.01	0.94	0.03	0.79	0.02	0.88	0.12	0.14	-0.22	0.14	0.00	0.99	0.00	0.97
Cre _p	T	-0.79	0.00	-0.70	0.00	-0.70	0.00	-0.88	0.00	-0.53	0.00	-0.80	0.00	-0.88	0.00	-0.86	0.00
	H	-0.15	0.03	0.21	0.06	-0.14	0.21	0.00	1.00	0.17	0.19	-0.12	0.16	0.18	0.09	0.11	0.29
	P	0.00	0.98	-0.03	0.04	0.18	0.09	0.03	0.05	0.13	0.03	-0.72	0.00	0.01	0.93	0.12	0.02
	w	-0.05	0.48	-0.10	0.37	0.00	0.98	-0.04	0.53	-0.05	0.72	0.15	0.09	0.00	0.99	-0.01	0.87

5.4.4 Driving Cycle Characteristics Analysis

In the next step, driving cycles are constructed for each weather type. In addition, researchers revealed that driving behaviors are different between weekdays and weekends (Agarwal 2004; Harris and Webber 2012; Zhong, Hunt, and Lu 2008). To understand the discrepancies and their connections with weather conditions, separate driving cycles were designed for weekdays and weekends. This resulted in a total of sixteen distinct driving cycles: eight for weekdays and eight

for weekends. By using this methodology, the dynamics of traffic patterns can be better understood, and the effect of meteorological factors on driving behaviors can be more accurately quantified. The following sections provide more details about different characteristics of the generated driving cycles.

Velocity analysis

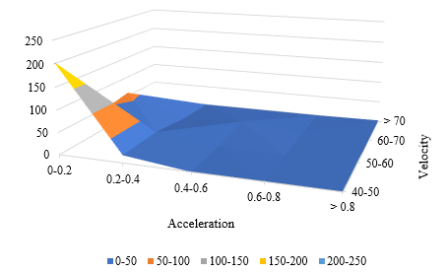
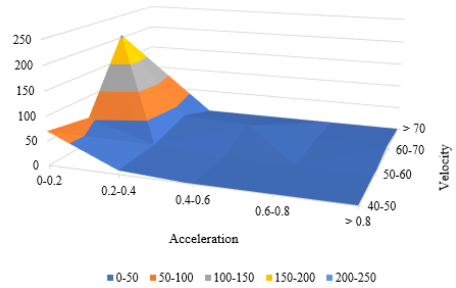
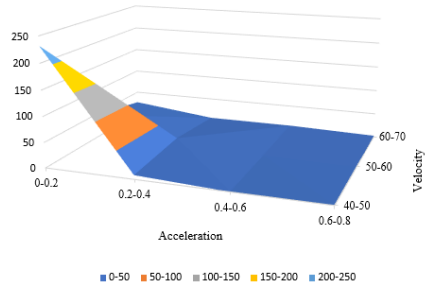
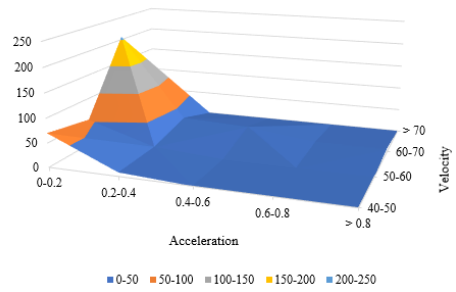
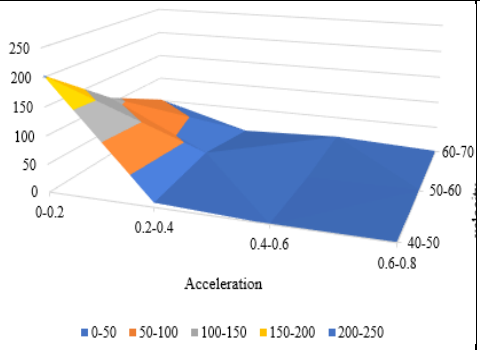
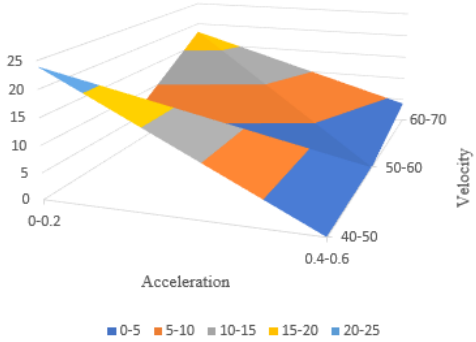
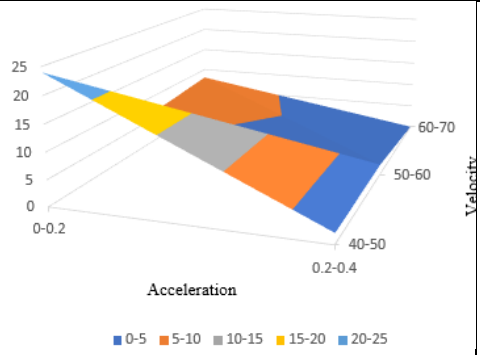
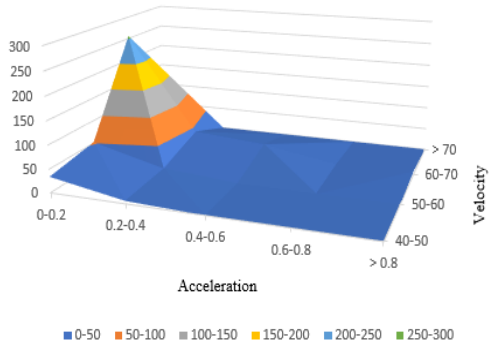
Velocity is a crucial driving cycle indicator that significantly impacts both vehicle emissions and fuel consumption. The data indicates that, in general, drivers tend to drive faster on weekends. On weekdays, the highest average speed of 43 km/h is observed under weather condition type 3, while the lowest average speed of 26 km/h is recorded in weather condition type 8. On weekends, the picture is somewhat similar, with the highest average speed of 48 km/h occurring under weather type 3, and the lowest average speed of 28 km/h being observed under weather conditions classified as type 8.

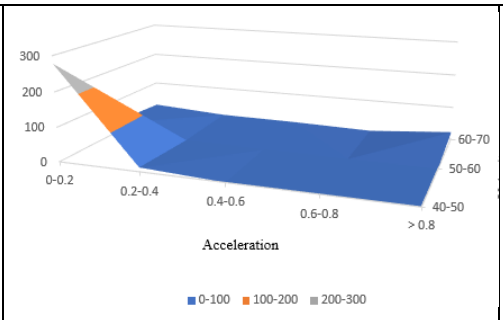
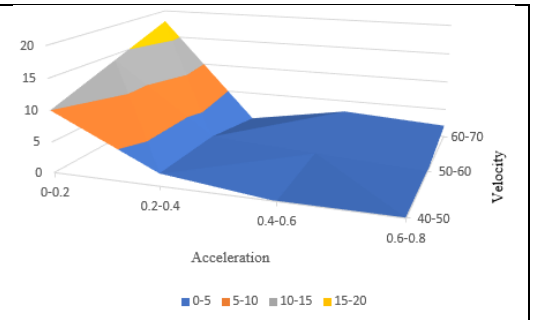
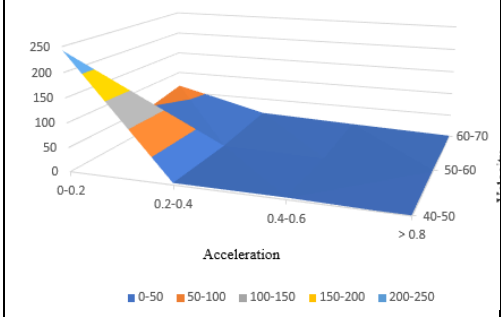
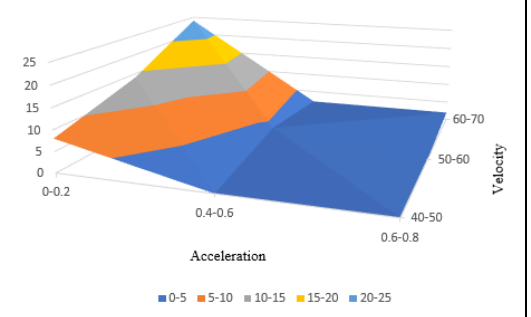
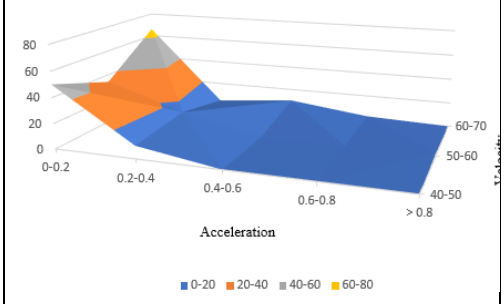
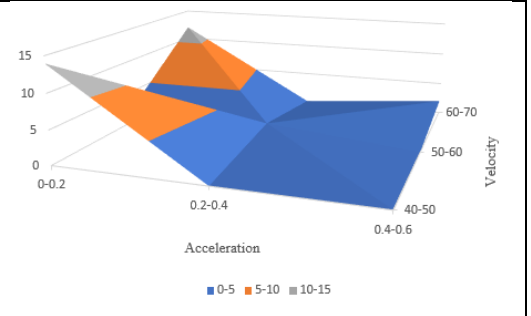
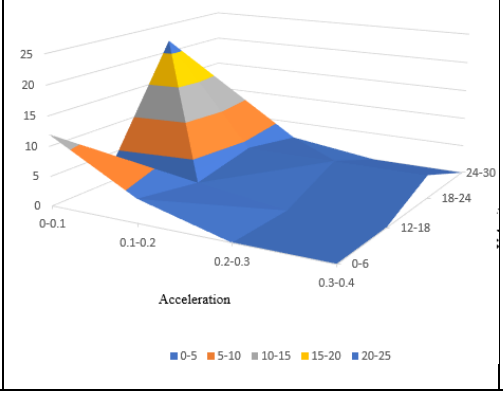
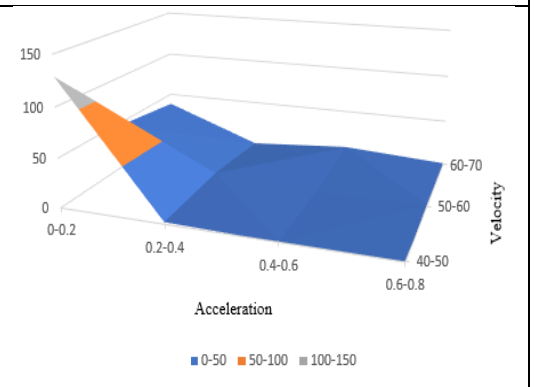
Acceleration analysis

Acceleration and deceleration are key factors that control how vehicles move and stay stable on the road. Vehicle acceleration and deceleration refer to the rates of change in the vehicle's velocity over time, measured in units of meters per second squared (m/s^2) or kilometers per hour per second (kph/s), and represent speed up or slow down, respectively.

Table 5-10 shows the Speed Acceleration Frequency Distribution (SAFD) of the generated driving cycles. SAFD shows the frequency and magnitude of changes in speed and acceleration during a driving session. An analysis of the SAFD reveals that driving patterns differ between weekdays and weekends. The frequency and magnitude of changes in speed and acceleration vary depending on the day of the week. In addition, in all weather types except weather type 4, accelerations are distributed from 0 to more than 0.8 meters per second squared (m/s^2).

Table 5-10 Comparison between SAFD of weekdays and weekend

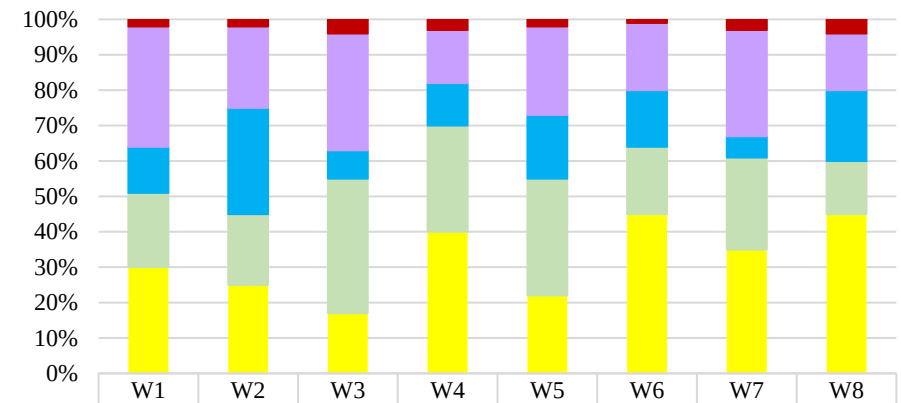
WT	Weekdays	Weekend
1		
2		
3		
4		

5		
6		
7		
8		
WT: Weather Type		

Share of Driving Modes

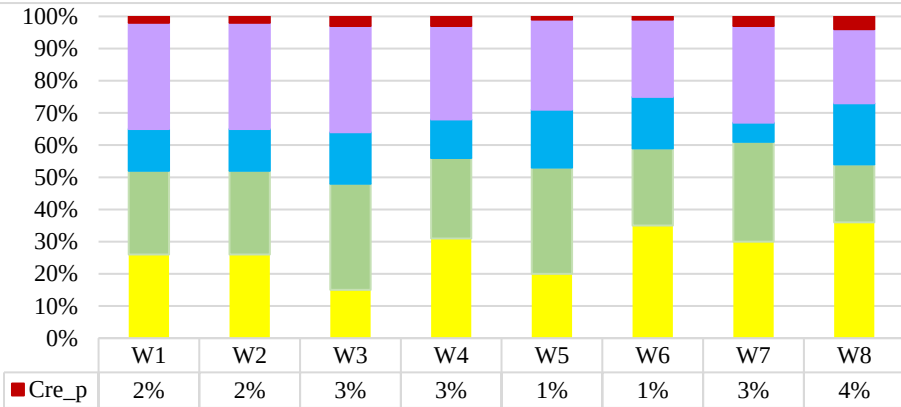
Figure 5-5 portrays the distribution of various driving modes across weekdays and weekends under different weather conditions. Analyzing the mode distribution on different days, except for weather types 1 and 2 on weekends, reveals that they are dissimilar, indicating varying driving patterns in

diverse driving scenarios. During weekdays, the time proportion of acceleration values exhibit oscillations ranging from 15 to 38%, while on weekends, they range from 18 to 33%. Additionally, the average deceleration values range between 6% to 30% on weekdays and 6% to 19% on weekends. In nearly all cases, except for weather type 8 on weekends, acceleration percentages exceed deceleration percentages. The idling time percentages vary on weekdays, with weather types 3 exhibiting 17 percent, while weather type 8 has almost 45%. The share of cruising mode is from 15 to 34 on weekdays, but on weekends, it ranges from 23 to 33%.



	W1	W2	W3	W4	W5	W6	W7	W8
Cre_p	2%	2%	4%	3%	2%	1%	3%	4%
Cru_p	34%	23%	33%	15%	25%	19%	30%	16%
Dcc_p	13%	30%	8%	12%	18%	16%	6%	20%
Acc_p	21%	20%	38%	30%	33%	19%	26%	15%
Idle_p	30%	25%	17%	40%	22%	45%	35%	45%

(a)



	W1	W2	W3	W4	W5	W6	W7	W8
Cre_p	2%	2%	3%	3%	1%	1%	3%	4%
Cru_p	33%	33%	33%	29%	28%	24%	30%	23%
Dcc_p	13%	13%	16%	12%	18%	16%	6%	19%
Acc_p	26%	26%	33%	25%	33%	24%	31%	18%
Idle_p	26%	26%	15%	31%	20%	35%	30%	36%

(b)

Figure 5-5 Share of modes for all weather types on weekdays (a) and weekends (b)

5.4.5 Comparative Analysis of Newly Constructed Driving Cycles Versus Established Models

After creating the driving cycles, the first step involved comparing these newly constructed cycles with existing ones, such as the Montreal Local Driving Cycle (Nouri and Morency 2017) and the Standard Driving Cycle (SDC) from Canada's 5-Cycle test. To illustrate their differences, as presented in Table 5-11, the authors used descriptive statistics of the constructed driving cycle parameters' values alongside those from previous studies. This comparison included the minimum and maximum values of driving cycle parameters on weekdays and weekends. The results confirmed that the values of the former driving cycles mostly do not fall within the range of the newly generated driving cycle values. For instance, the minimum and maximum velocities were 26 and 48 km/h, respectively, compared to 25 and 34 km/h in the previous local and standard driving cycles.

Additionally, a One Sample t-test was performed to determine whether the mean values of the new driving cycle parameters significantly differed from those in the previous local and standard cycles. The null hypothesis posited that the new driving cycle's mean value, μ, for each parameter, would equal the corresponding values in the local and standard driving cycles (μ0). Conversely, the alternative hypothesis suggested that the driving cycle's mean value for each parameter would not equal the corresponding ones in the local and standard driving cycles. The hypotheses are formulated as follows:

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

To perform this test, the study utilized the SciPy library in Python. The equation (5.1) shows t-test statistic formula:

$$t = \frac{\mu - \mu_0}{\frac{s}{\sqrt{n}}} \tag{5.1}$$

Where standard deviation as *s* and the number of newly constructed driving cycles as *n*.

Table 5-11 presents the results of the t-test, indicating significant differences between the generated driving cycles and the two former ones. The table clearly shows that the p-values of the driving cycle parameters are less than 0.05, suggesting that the proposed approach significantly differs from the pre-existing ones.

This highlights that a single driving cycle cannot adequately represent driving behaviors under varying conditions, emphasizing the importance of considering the influence of weather when constructing driving cycles.

5.4.6 Determining the Representative Sets of Driving Cycles

Descriptive analysis revealed that the proportion of different modes of transport varies with weather types on weekends and weekdays, indicating that the driving cycles differ. This study specifically employs the similarity matrix technique, which is adept at facilitating the pairwise comparison of the multidimensional arrays representing driving cycles. The construction of the similarity matrix is a nuanced process that involves the normalization of parameter data to a common scale to ensure comparability and then employing a pairwise Euclidean distance computation across all possible pairs of driving cycles. The similarity between driving cycles is quantified with a value between 0 and 1, which in this case, is derived from calculating the Euclidean distance between vectors of the 13 parameters for each pair of driving cycles. Euclidean distance between two points is calculated using the equation (5.2).

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{5.2}$$

Where pi and qi are the values of the ith parameter for driving cycles P and Q respectively, and n is the total number of parameters. After calculating the Euclidean distances, these are then transformed into similarity scores. This transformation is typically achieved by inverting the distances so that a smaller distance (indicating higher similarity) results in a higher similarity score.

Table 5-11 differences between the developed driving cycles local and standard driving cycles

		V	Vr	Vm	V_std	Acc	Dcc	Acc 2	Acc_std	Idle_p	Acc_p	Dcc_p	Cru_p	Cre_p
Proposed model	**Min-WDs**	26	26	73	2	0	-0.6	0	0.2	17%	15%	6%	15%	1%
	Max-WDs	43	43	91	6	0.8	0	1.5	1.2	45%	38%	30%	34%	4%
	Min-WE	28	28	73	2	0.1	-0.2	1	0	15%	18%	6%	23%	1%
	Max-WE	48	48	92	10	0.3	-0.1	0.3	0.6	36%	33%	19%	33%	4%
(Nouri P. 2015)		25	32	104	-	0.43	-0.5	-	0.29	19.9%	34.5%	12.5%	31.2%	1.6%
5-cycle test (city)		34	-	90	-	5.3	-	-	-	18%	-	-	-	-
One sample T-test (Nouri P. 2015)	T-statistic	6.92	2.3	-16.86	-	-5.1	10.6	-	2.1	4.3	-5.2	2.21	-3.18	3.2
	P-value	0.0	0.04	0	-	0	0	-	0.04	0	0	0.04	0	0
One sample T-test 5-cycle test (city)	T-statistic	2.32	-	-7.2	-	-110	-	-	-	5.1	-	-	-	-
	P-value	0.03	-	0	-	0	-	-	-	0	-	-	-	-

One common method is to use the formula $\frac{1}{(1+d(P,Q))}$ which bounds the similarity score between 0 and 1. The closer the score is to 1, the more similar the driving cycles are with respect to the parameters considered.

As is presented in Figure 5-6, the similarity matrix visually encodes the level of resemblance, with darker shades representing higher values and thus greater similarity. The findings from this matrix are pivotal, as they demonstrate that weekends in weather types 1 and 2 exhibit a 0.92 similarity, which is a substantial overlap, suggesting that they could be effectively classified as a single driving cycle for the purposes of modeling and analysis. This high degree of similarity may reflect a commonality in driving behavior patterns during these periods, perhaps due to reduced variability in driving conditions or consistent driver responses to these conditions.

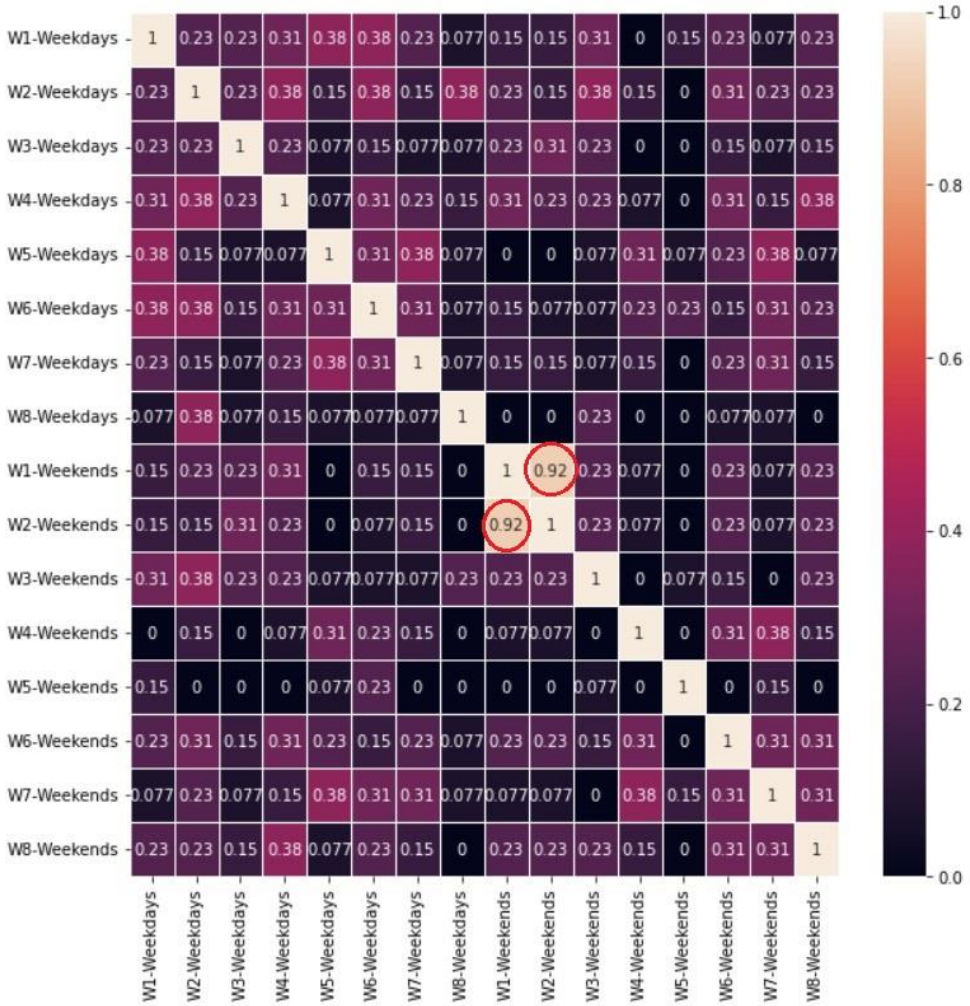


Figure 5-6 Similarity matrix of the generated driving cycles in different weather conditions

Such a matrix not only provides insights into the inherent grouping among the driving cycles but also aids in identifying unique driving cycles that do not conform to the general patterns observed. This has significant implications for the design and testing of automotive systems, where a nuanced understanding of driving behaviors across different conditions is crucial for optimizing performance and safety.

To conclude, based on the obtained results, it can be inferred that the representative number of driving cycles to accurately portray driving behaviors in all driving situations for the chosen road category is 15.

5.5 Conclusion

This study mainly targets three concerns that are associated with the volatility of the driving cycle under various weather circumstances, utilizing scientific investigation and analysis. 1) limitations in representing real-world driving conditions with current driving cycle development methodologies; 2) insufficient analysis of the impact of weather characteristics on driving cycle parameters; 3) scarcity of discussions on a systematic approach to identify different weather types in vehicle emissions context.

Firstly, the best compromise number of driving cycles is determined to represent the driving patterns in various driving conditions. The proposed model effectively identified 15 driving cycles that represent driving behaviors under various conditions, but it's important to note that these findings are currently specific to a certain type of segments. Further validation is necessary to determine if these results can be generalized to other segment types. Of the 15 driving cycles generated in this study, 8 represent weekday driving behaviors and 7 represent weekend driving behavior. The aforementioned finding underscores the crucial need to incorporate weather conditions, weekday and weekend driving behaviors in developing fuel consumption testing protocols. Doing so is essential to ensure accurate guidance for the automobile industry and effective decarbonization planning.

Secondly, a MMRM approach is developed to identify the most essential weather variables playing a significant role in the variability of driving behaviors. The model revealed that temperature exerts a more significant influence on deceleration compared to other variables, whereas precipitation impacts velocity, wind speed affects acceleration, and humidity displays minimal effects on driving cycle parameters. To enhance their understanding of the influence of these factors on driving

parameters, the researchers opted to isolate the effects of weather conditions by selecting a road type prior to performing the regression analysis. The model's results shed new light on the association between weather conditions and driving cycles, offering insights that may prove useful for researchers exploring the impact of climate change on driving behavior.

Lastly, an unsupervised machine learning method was employed to cluster weather conditions based on factors affecting driving cycle parameters. The resulting clusters revealed the existence of 8 distinct weather types in the study area.

A major constraint of this study is the limited availability of data. Numerous factors can characterize weather conditions, yet the prevailing weather data compiler, Environment Canada, presently only collects a restricted set of weather parameters on an hourly basis. Notably, visibility, a key significant parameter in transportation, is not included in the hourly data.

The following points are suggested for future research:

- There is mounting evidence that weather variables are expected to undergo significant changes in the future as a result of climate change. To better inform mitigation strategies, decision-makers may benefit from a model that can predict how changes in weather parameters will affect driving cycle parameters. In addition, such a model could help ensure that legislative and regulatory measures are updated in a more informed and proactive manner.
- This study utilizes taxi data to generate driving cycles, and applying the developed model to passenger cars can potentially improve the accuracy of driving cycle predictions.
- Given that Montreal has a continental climate, extending the proposed methodology to other urban areas with different climate types can enhance the comprehensiveness of this study.
- Validation of the proposed model across different types of segments. This would not only test the robustness of the model but also potentially broaden its applicability in understanding driving behaviors under a variety of driving conditions.

CHAPTER 6 ARTICLE 3: A METHODOLOGICAL FRAMEWORK TO DETERMINE THE OPTIMAL SET OF DRIVING CYCLES FOR A REGION

Written by

Asad Yarahmadi, Catherine Morency, Martin Trepanier

Published in Octobre 2, 2023

Journal: Transportation Research Board (TBR 103)

Abstract

Driving cycles are essential for estimating vehicle emissions by reflecting regional driving patterns. While factors like weather and road types profoundly impact these patterns, many studies tend to neglect this diversity, opting for limited driving cycles. This research introduces a comprehensive framework to determine the best driving cycles, using Montreal as an exemplar. With machine learning, we categorized weather and road types. From these categories, we constructed a multilayered framework, culminating in a mixed model that examines how road and weather factors affect driving cycles. We then generated 80 unique driving cycles, tailored to different conditions, differentiating between weekdays and weekends. Next, a two-stage method was employed: first, a matrix to measure similarities between cycles, and then, a method using information loss metrics to finalize the ideal cycle number. Our findings suggest that Montreal's driving conditions are best captured with 76 distinct cycles.

Keywords: Vehicle emissions, Driving cycle, Information loss, Variability, Weather condition, Road types

6.1 Introduction

The challenge of curbing greenhouse gas emissions from global transportation systems is amplified due to the persistent escalation in passenger activity, which could potentially eclipse the beneficial impacts of all deployed mitigation strategies (IPCC 2022). According to the International Council on Clean Transportation (ICCT), a significant obstacle in mitigating vehicle emissions lies in improving the accuracy of effective emissions measures to ensure that all stakeholders in the transportation sector adhere to the stipulated legislative guidelines for emissions control (Z. Yang, Muncrief, and Bandivadekar 2017). The driving cycle serves as a pivotal measure extensively used in the modeling of vehicle emissions.

The prevalent driving cycles are SDC (Standard Driving Cycle) (Knez et al. 2014) and LDC (Local Driving Cycle) (Andrade et al. 2020; Yang Yang et al. 2020). The SDCs represent driving behaviors at the national (e.g., FTP-75 and J10-15) or international level (e.g., WLTC) to certify a vehicle's emissions. While LDCs represent driving patterns at a local scale (city, road type …). The majority of pre-existing driving cycles, formulated at the urban or national levels, may fail to address the fluctuations in emissions, potentially leading to significant inaccuracies in emission estimations (Mayakuntla and Verma, 2018; Liu et al., 2020; Zhang et al., 2021; Kazemi Miyangaskari et al., 2023).

A key challenge in our study is determining the "optimal" number of driving cycles, which is crucial to capture the variability of driving behaviors across all types of roads and driving conditions. Recent research has shown that factors such as precipitation, temperature, road gradient, and speed limits significantly contribute to the variability of driving cycle parameters (Yarahmadi, Morency, and Trepanier 2023b; Zhang et al. 2021). As a result, driving behavior vary across an entire city road network, changing as weather conditions change.

As shown in Figure 6-1, the number of unique driving patterns can range from a single cycle to a count equal to the total number of drivers in an urban area. Opting for a singular driving cycle necessitates a highly aggregated level of driving data, a process which results in substantial information loss and a marked reduction in behavior variability representation. Consequently, the derived driving cycle often fails to accurately represent driving behaviours, engendering a notable discrepancy between actual and projected vehicle emissions. It is worth mentioning that information loss refers to the reduction in the diversity, granularity, and representation of specific

driving characteristics when transitioning from the collected driving data to the representative driving cycle(s) (Liu et al. 2020).

Figure 6-1 Driving cycle number spectrum

Conversely, generating individual driving cycles for each driver heightens computational complexity and magnifies both the consumption of time and resources. This approach also escalates the risk of encountering outliers and data redundancy. Against this backdrop, the central objective of our current study is to devise a novel machine learning methodology with the aim of precisely determining the “optimal” number of driving cycles within an urban context.

In this study, “optimal” refers to a compromise between a set of values. The “optimal” number of driving cycles is the minimum number of driving cycles that still can capture the maximum variability with the least information loss, effectively representing driving behaviors across different driving conditions, various road types, and in different types of days. The number of driving cycles in an urban setting like Montreal is unlimited and undefined. In fact, trip characteristics are function of exogenous factors like road, weather, and type of day, therefore, there are indefinite number of driving cycles in an urban area. As a result, this study uses the word 'finite' to indicate a practical number of driving cycles that can realistically be generated, considering the limitations of data availability, computational resources, and the importance of minimizing the impact of anomalies. A crucial component of our proposed strategy is its intended adaptability - the methodology should be sufficiently versatile to be applicable across diverse urban areas, irrespective of the size and complexity of a city's road network structure, or its prevailing weather conditions.

A review of existing literature underscores the trend of previous studies predominantly selecting one, or a limited number of driving cycles to represent driving patterns in urban contexts. Our study aims to significantly expand upon this base knowledge, proposing an optimized, adaptive solution that is capable of accurately representing the driving behaviors variability in diverse urban contexts.

The present investigation offers several significant contributions to the understanding and analysis of driving cycles within urban environments:

- The study delineates a methodology for determining the “optimal” number of driving cycles for an urban area, contributing to a more accurate representation of driving behaviors.
- The research introduces an approach for defining the appropriate level of data amalgamation for the formation of driving cycles, thereby mitigating the challenge of information loss and ensuring the representation of behavior variability.
- Furthermore, the investigation provides a robust model designed to identify and quantify the exogenous factors of most significant influence on driving cycle parameters. This addition to the field enables a more nuanced understanding of the variables that shape driving behaviors and contributes to the precise generation of representative driving cycles.

This study addresses the following critical inquiries:

- What is the required number of driving cycles necessary to accurately represent the typical driving patterns in Montreal?
- Which exogenous factors rank as the most significant contributors to the variability observed in driving conditions and vehicle emissions in urban settings?

The remainder of this paper is structured as follows. First, previous studies that explored the impacts of exogenous factors on driving cycle parameters and different approaches to determine “optimal” number of driving cycles are exposed. Then, the methodology section provides a background about techniques used to conduct this study and describes the steps to find the “optimal” number of driving cycles. Finally, the results are discussed, and some perspectives are proposed.

6.2 Literature review

The development of a driving cycle can be predominantly delineated into three phases: accumulation of driving data, creation of a driving cycle, and recognition of a representative driving cycle. To enhance the precision of the designed driving cycle, prior research has leveraged precise data collection methodologies, which range from smartphone sensor data (Rechkemmer et al.

2021), to Onboard Recording Devices and OBD-II systems (Ma et al. 2019), connected Vehicle Data (Khattak and Wali 2017), and observation of Naturalistic Driving behaviors (Esser and Rinderknecht 2020). Previous research also have introduced diverse techniques as Microtrips-based (Yang et al. 2020), Markov Chain-based (Roy and Morency 2020; Yang et al. 2018), and Fuel-based (Huertas et al. 2018; Quirama et al. 2020) methods to generate driving cycles (Czégé, Vámosi, and Kocsis 2020; Huertas et al. 2019).

Although considerable strides have been made in creating driving cycles, researchers have scarcely explored the link between fluctuations in driving cycle parameters and external elements like road attributes and weather conditions. As a result, a variety of road segments are selected for data collection during the route selection phase without considering their impact on driving behavior. The data was gathered without accounting for driving conditions, and in the end, all data was amalgamated and a single or a few driving cycles were selected to typify driving behaviors within a city area.

6.2.1 Route selection and road features

The selection of a route is a crucial step that can significantly influence the precision of a driving cycle. During this phase, road links are designated for the data collection process. Therefore, if the chosen segments fail to accurately mirror the characteristics of an urban road network, the driving cycle might not appropriately depict driving behaviors.

A review of related literature indicates that authors commonly focus on traffic characteristics (Tong and Ng 2021; Yang et al. 2020), experiential knowledge (Knez et al. 2014), road functions (Arun et al. 2017), or a limited set of factors such as traffic flow, population density, squares, junctions, and public transport systems (Pouresmaeili, Aghayan, and Taghizadeh 2018) when identifying representative road segments. However, the consideration of road features, like gradient, curvature, intersection density, road width, and built environment that can influence driving behaviors, is seldom observed (Azam, Bhaskar, and Haque 2022; Cui and Levinson 2018).

In their study, Salihu, Demir and Demir, (2023) analyzed the effect of road gradient on driving cycle parameters, amassing around 10 hours of driving data. Their findings revealed that acceleration and deceleration are contingent on the incline of the road. They concluded that driving cycles differ in road segments with varying slopes and to increase the accuracy of vehicle emissions estimates, road segments with gradients ranging from -2 to 2 should be incorporated in driving

cycle computations. However, they did not elaborate on the relationship between road gradient and changes in driving mode.

In another study, Abele, Møller and Mette, (2011) examined the influence of road design features, including road shoulders, their presence, and horizontal geometry, on driving speed and lateral position as indicators of driving behaviors. Their research demonstrated that participants drove slower and closer to the centerline on curved roads compared to straight ones. The presence of road shoulders was linked with faster speeds and changes in lateral positions on all road segment types. Participants generally displayed higher speeds when road shoulders were available. Additionally, their findings contradicted previous studies (Antonson et al. 2009; Sagberg 2003) by suggesting that the presence of trees has no significant effect on driving speed. However, their study was limited as it was conducted in a simulator environment with limited data and did not consider the influence of traffic conditions on driving behavior.

Further research by Wolfermann, Alhajyaseen and Nakamura, (2011) confirmed that signalized intersections affect driving speed profiles, while Khanfar et al., (2023) demonstrated that flashing green lights encourage more conservative driving behaviors through an analysis of vehicle kinematic parameters. Supiyono, Djakfar and Wicaksono, (2022) found that among all traffic light phases, the yellow light has a significant impact on driving behavior.

6.2.2 Collecting driving data and driving conditions

The process of collecting driving data typically involves gathering real-world data over a period that ranges from three days to two weeks in most studies. Alternatively, researchers may use simulation techniques to generate representative driving cycles. In both scenarios, the consideration of driving conditions may not be adequately addressed. Existing literature evidences that weather parameters, such as precipitation, temperature, wind speed, and humidity, can impact driving behaviors characteristics. Faria et al., (2020) demonstrated that average speed during heavy rain is 22% lower than in clear weather, and acceleration and deceleration reduce by approximately 10%. Ahmed et al., (2018) further illustrated that weather conditions like snow, fog, and rain compel drivers to significantly reduce their speed compared to clear weather conditions.

The failure to consider weather conditions in the generation of driving cycles may result in these cycles not adequately representing the diversity of driving behaviors. Consequently, the Local Driving Cycles (LDC) may be less representative than the Standard Driving Cycles (SDC) (Gebisa

et al. 2021). Some studies have addressed these limitations by collecting vast driving datasets across various road segments and weather conditions. However, they tend to aggregate all collected data and select either a single or a limited number of driving cycles.

To improve the accuracy of the LDCs, it is essential to consider the aggregation level of data and generate driving cycles for various driving conditions and road types. Yarahmadi, Morency and Trepanier, (2023b), using nearly 3 million trips collected over two years, confirmed that driving cycles under different driving conditions are distinct and proposed considering a separate driving cycle for each driving condition. Furthermore, Zhang et al., (2021) generated a driving cycle for different road functions and provided compelling evidence that driving cycle parameters differ among road types. They ultimately suggested that having a road type-based driving cycle could improve the estimation of vehicle emissions.

In Table 6-1, we present a comparative analysis between the methodology proposed in this study and those of prior investigations. This table delineates the distinctions between the current research and pre-existing literature. The column labeled "External Factor(s) considered in DC generation" elucidates the exogenous variables taken into account in driving cycle creation by various researchers. The "Aggregation level" column indicates the granularity at which driving data have been aggregated for the development of the driving cycle. The column “Number of DCs” presents the number of driving cycles were selected to represents driving behaviours in an area. Column “Duration of data gathering” shows how much driving data are collected to construct driving cycle. Additionally, the column titled " Approach(s) to measure accuracy of methodology " enumerates the metrics that authors have deemed important for assessing information degradation and variability. It should be highlighted that the acronym "NP" is indicative of "Not Presented", while "NC" denotes "Not Considered".

In conclusion, road characteristics and weather parameters significantly influence driving cycle parameters. However, the majority of prior studies have overlooked this aspect. They have instead aggregated all data collected under varying road and weather conditions to construct one or a limited number of driving cycles meant to characterize driving behaviors, which may not provide a sufficiently accurate representation. The development of a method to determine the “optimal” number of driving cycles for various driving conditions and road types could address this gap and enhance the estimation of vehicle emissions.

Table 6-1Comparison of conducted studies and proposed approach

Authors	Objectives	External Factor(s) considered in DC generation	Aggregation level	Number of DCs	Duration of data gathering	Approach(s) to measure accuracy of methodology
(Ragione and Giovanni 2016)	To probe road gradient impacts on DC features	Road slope	Road slope	4	NP	NC
(L. Zhang et al. 2021)	To improve vehicle emissions estimation	Road function	Highway, arterial, and other roads	3	167 hours	NC
(Valera et al. 2014)	To predict real-time DC for road grades	Road slope	NP	NP	NP	NC
(Seers *et al.*, 2015)	To develop DCs for suburban road vehicles	NP	All data	2	2 months	NC
(Ho *et al.*, 2014)	To construct a new SDC for road in Singapore	Traffic conditions	All data	2	40 hours	NC
(Yarahmadi, *et al.*, 2023a)	To generate and identify representative DCs in all driving conditions	Weather parameters	Weather type	15	19 months	NC
(Tharvin et al. 2018)	To construct a new LDC for road in Malaysia	Traffic conditions	peak and off-peak hours	1	NP	NC
(Arun et al. 2017)	To build new LDCs for passenger cars in Chennai, India	Traffic conditions	peak and off-peak hours	2	6 months	NC
(Chauhan, Joshi, and Parida 2020)	To generate new LDCs for an arterial	Road function/ intersections	Arterial urban roads	9	NP	NC
Proposed Methodology	To determine “optimal” number of DCs	Road types and weather types	Road type/ weather types/ weekday and weekends	76	19 months	Information loss and variability

6.3 Methodology

The primary challenge in determining the “optimal” number of driving cycles for a region is quantifying the variability of driving behaviors across the entire road network and under all driving conditions. Figure 6-2 illustrates various steps in this study to develop a methodological framework for identifying the “optimal” number of driving cycles. As shown in Figure 6-2, two steps (developing typology of the road network and typology of weather conditions) highlighted in gray, have already been completed in previous stages, and their results are utilized in this study.

The first step of this study involves selecting a road type and a weather type, and subsequently extracting the driving data generated under these selected conditions. Next, based on weekdays and weekends, two driving cycles are generated for the chosen road type and weather condition. This is an iterative process and is repeated for all weather conditions across all road types.

Subsequently, a similarity matrix is used to compare the similarity between the constructed driving cycles. Finally, the “optimal” number of driving cycles is calculated using two indicators: information loss and variability of driving cycles.

6.3.1 Driving data description

This study utilized a total of 6292 taxis, amassing a substantial dataset of approximately 10,567,000 journeys carried out from April 2019 to December 2020 under authentic driving conditions. This incorporated both high-demand and low-demand service periods. A detailed overview of the primary characteristics of this driving data is provided in Table 6-2. In addition, the data collection did not favor secondary factors such as vehicle model, driver's age, or gender. This allows researchers to hypothesize that the data likely represents driving patterns across a diverse range of age demographics, genders, and passenger vehicle types. The data was gathered from the entire Montreal Road network, encompassing all road functions (arterial, primary, local, etc.) with approximately 9,400 segments spanning a total length of 130 km, under all driving conditions.

Table 6-2 Driving data features

Variable	**Ride ID**	**Timestamp**	**Taxi ID**	**Speed (km/h)**	**Longitude**	**Latitude**
Type	Text	Date and Time	Text	Number	Number	Number

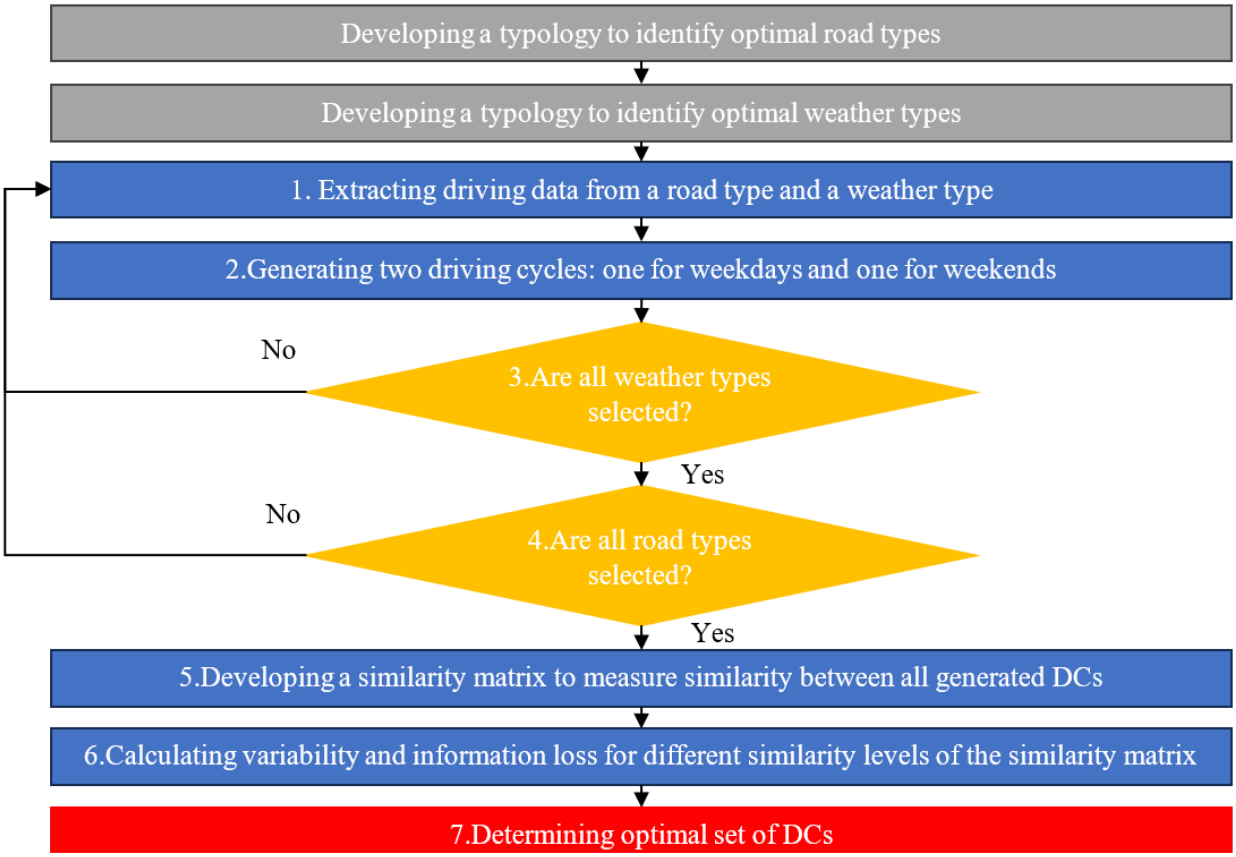


Figure 6-2 Conceptual model of the research methodology

6.3.2 New typology of road network

Yarahmadi, Morency and Trepanier, (2023b) constructed a typology to categorize road segments based on their impact on driving speed. They initially identified the most significant road features that have a significant relationship with driving speed. The key features include road grades, curvature, density of signalized intersections, number of road lanes, surrounding land uses, posted speed limits, and road functions. Subsequently, they utilized machine learning to categorize the road segments, leading to the identification of five distinct road types. Their results demonstrated that the newly created typology significantly enhances the explanation of the association between changes in driving speed and road features. In addition, a statistical analysis of speed distributions within each road type revealed that driving speeds are more homogenous within each road type than between different road types.

The developed typology enables the authors to collect driving speed data for each road type, generate driving cycles, and analyze and compare the characteristics of the constructed driving cycles.

6.3.3 Weather typology

The significant variability in weather conditions throughout the day and year presents a significant challenge in identifying a representative driving cycle for a specific condition. For instance, a descriptive analysis of hourly weather data from Montreal over a year exhibits the following oscillations in factors influencing driving behavior: temperatures range from -35 to 35 degrees, humidity fluctuates between 20 to 100%, wind speed varies from 0 to 70 km/h, and precipitation (snow and rain) ranges from 0 to 28 mm. As a result, to accurately represent driving behaviors throughout the year, it is crucial initially to develop a framework that can comprehend and quantify the variability of driving conditions. An important prerequisite for this is the isolation of the impact of weather from road features, given that literature suggests that road features can influence the effects of weather on speed (Chen et al. 2019; Chiou and Chen 2010).

Addressing this, Yarahmadi, Morency and Trepanier, (2023b) developed a weather typology using factors affecting speed to quantify the variability of weather conditions. Their findings suggest that weather conditions in Montreal can be categorized into eight distinct weather types, each corresponding to distinct driving behaviors. To validate their results, they selected a road type, as established in the previous step, and generated driving cycles for the different weather types. As the road type remained constant while weather conditions varied, it was possible to isolate the impact of weather conditions on driving cycle parameters and ascertain if the generated weather types exert different influences on driving behavior. Subsequent statistical analysis revealed that the generated driving cycles under different weather conditions were notably distinct, validating the results of the new weather typology. Moreover, the developed typology provides a better representation of changes in driving cycle parameters.

This study leverages the established weather typology to generate driving cycles under diverse driving conditions.

6.3.4 Driving cycle generation

The process of generating a driving cycle can be broken down into three stages: data collection, construction of candidate driving cycles, and selection of the representative cycle.

Concerning data collection, various methodologies exist including the chase car method, connected vehicle, OBD Data collection, driving simulator, fleet data collection, and mobile application. This study capitalizes on the fleet data collection technique as it presents unique advantages. Firstly, data loggers (i.e., GPS) are permanently installed on a large number of vehicles such as taxis, which in contrast to other techniques, results in the collection of extensive datasets. Secondly, a wide array of drivers traverses various road types and driving conditions, during peak and off-peak hours, in order to transport passengers to their destinations. Consequently, due to the large volume of data that showcases driving behaviors in specific road types and driving conditions, the repeatability of data analysis becomes feasible, leading to more robust conclusions. In this study, the fleet data collection technique has been employed to gather driving data from April 2019 to December 2020.

Upon data collection, map matching is utilized to partition the data according to road types. Following that, the driving data for each road type are subdivided into weekday and weekend groups. These groups are further partitioned into eight subgroups in accordance with weather types. For each weather type, distance-based microtrips are generated. Research by (Nouri and Morency 2015) demonstrated that a 250-meter length for microtrips is optimal for generating a driving cycle. Subsequently, Kmeans clustering is employed to categorize microtrips and the Markov chain technique is implemented to construct candidate driving cycles.

Ultimately, as illustrated in Table 6-3, thirteen parameters are used to evaluate the generated driving cycles and choose the representative one. However, diverging from the traditional approach in existing literature, in this methodology two driving cycles are selected to represent driving behaviors under a specific driving condition - one for weekdays and one for weekends. Thus, in total, 16 driving cycles are chosen to reflect driving behaviors for each road type. Consequently, roughly 80 driving cycles are generated, encompassing the five road types and eight weather conditions.

Table 6-3 Driving cycle assessment criteria

Assessment criteria	Denoted	Unit
Difference in SAFD	NA	(%)
The average speed of the entire driving cycle	V	(km/h)
Average running speed	Vr	(km/h)
Maximum speed	Vm	(km/h)
The average acceleration of all acceleration phases	Acc	(m/s2)
Average deceleration of all deceleration phases	Dcc	(m/s2)
The average number of acceleration-deceleration changes	Acc_std	NA
Root means square acceleration.	Acc2	(m/s2)
Velocity standard deviation	NA	(kW)
Time proportion of idle mode	Idle_p	(%)
Time proportion of acceleration mode	Acc_p	(%)
Time proportion of cruising mode (−0.1 m/s2 < acceleration < 0.1 m/s2; average speed > 5 m/s)	Cru_p	(%)
Time proportion of deceleration mode (acceleration ≤− 0.1 m/s2)	Dcc_p	(%)
Time proportion of creeping mode (−0.1 m/s2< acceleration < 0.1 m/s2; average speed < 5 m/s)	Cre_p	(%)

6.3.5 Mixed models

After constructing the driving cycles, it is necessary to examine the interplay between the generated road types, weather types, and parameters of the driving cycles that were constructed. Mixed models, also known as multilevel models, are employed for this purpose. These models blend fixed and random effects to account for hierarchical, complex structures within data. They prove especially useful in dealing with data that is grouped or clustered, as they consider the interdependencies typically found in such datasets. Fixed effects are parameters that remain

constant across individuals, while random effects allow these parameters to exhibit variation. In mixed-effects models, both types are incorporated, granting a higher degree of flexibility and applicability. This is particularly beneficial when dealing with clustered or grouped data because they efficiently account for the intra-group correlations by incorporating random effects for the groups.

The model is crafted using the Python programming language, utilizing the Statsmodels library. As depicted in Figure 6-3, there exists a hierarchical data structure where road types, weather types, and divisions of weekdays and weekends are treated as fixed effects. Drivers in this study are treated as random effects. 30 drivers are selected randomly, and driving cycles are generated for each individual driver in different driving conditions and totally 2400 driving cycles are constructed. It should be noted that Table 6-2 shows that the driving data includes a unique Taxi ID for each taxi. We utilized these IDs to collect driving data for taxis under various weather conditions and on different road types. Therefore, in this study, the driver—identified by the taxi ID—is considered a random effect. Ultimately, equation 6.1 is developed in the final step to examine the interplay of fixed effects and to determine their influence on the random effects.

$$DC_p = \boldsymbol{GI_p} + \sum_{i=1}^{n} \beta_{1i,p}\, RT_i + \sum_{j=1}^{m} \beta_{2j,p}\, WT_j + \sum_{k=1}^{l} \beta_{3k,p} DT_k + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} \beta_{ijk,p}\left(RT_i * WT_j * DT_k\right) + \boldsymbol{\varepsilon}_p \quad (6.1)$$

Where:

DC_p: Dependent variable for the p-th driving cycle parameter, where p ranges from 1 to 13. Each number from 1 to 13 represents one of the 13 driving parameters that are being analyzed.

n: Total number of road types,

m: Total number of weather conditions, and

l: Total number of day types.

$GI_{\mathbf{p}}$: Global intercept of the chosen parameter.

$\boldsymbol{\varepsilon}_p$: Random error term for the p-th parameter.

$\beta_{1i,p}$: Coefficient for the effects of the i-th road type on the p-th parameter.

$\beta_{2j,p}$: Coefficient for the effects of the j-th weather condition on the p-th parameter.

$\beta_{3k,p}$: Coefficient for the effects of the k-th type of day on the p-th parameter.

$\beta_{ijk,p}$: Coefficients for the three-way interaction effects among the fixed factors for the p-th parameter.

RT_i, WT_j, DT_k: Indicator variables correspond to each level of the categorical fixed effects.

The equation 6.1 is broken down into two components: a summation and a product. The summation is used to investigate the impact of external factors on driving cycle parameters. The product is used to examine the interaction between road types and weather conditions during weekdays and weekends.

6.3.6 "Optimal" number of driving cycles

To decide on the "optimal" number of driving cycles, this study employed a two-stage process. Initially, information loss is used to set the level of data aggregation. After that, the study utilized variability and information loss indicators to determine the "optimal" number of driving cycles.

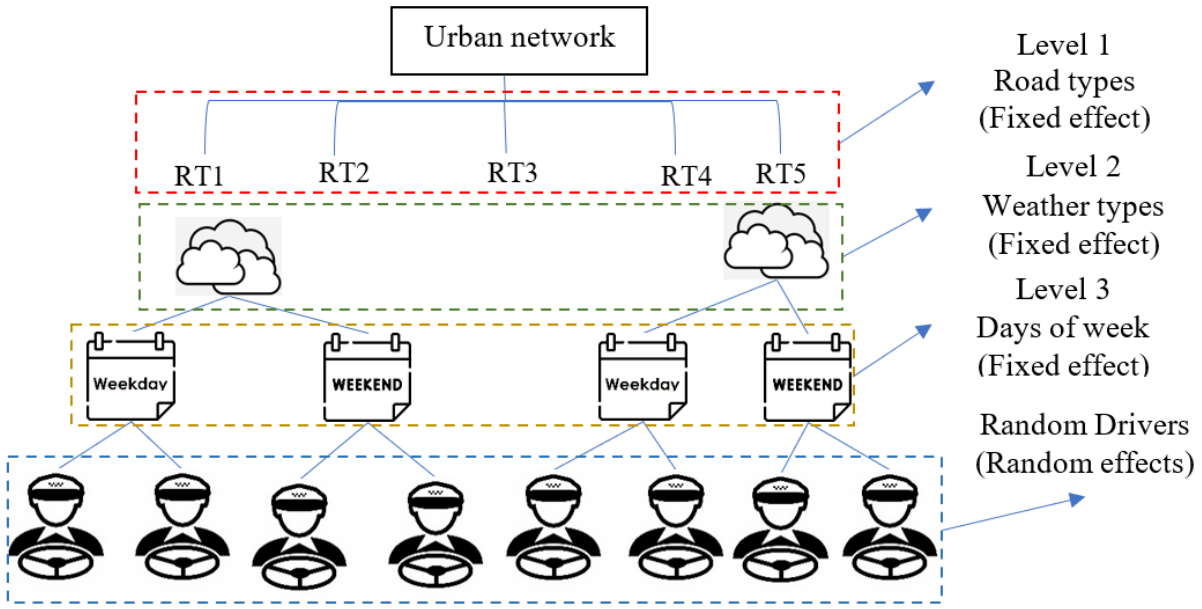


Figure 6-3 Hierarchical structure of data of the present study

Step 1: Determining aggregation level of driving data

One of the primary hurdles in identifying the "optimal" number of driving cycles, as discussed in Section 6.2.2, is setting the aggregation level of driving data. This is accomplished using information loss as a tool to statistically investigate and contrast the distribution of driving cycle parameters against the distribution of all driving data. Information loss could happen if the chosen driving cycles do not satisfactorily represent all driving behaviors. Hence, it is critical to confirm that the method proposed for aggregating driving data leads to minimal information loss.

This research utilizes two metrics to quantify information loss. The first metric, expressed in equation (6.), is proposed by (Liu et al. 2020). This metric computes the ratio of the range of selected driving cycle parameters to the range of driving cycle parameters for all drivers and is referred to as the ratio of sampled to actual range.

Estimating driving cycle parameters for every driver is not feasible, so to overcome this hurdle, all trips (the complete dataset) are uploaded into the driving cycle tool and the microtrips are generated for them. Since a driving cycle is composed of linked microtrips, the range of microtrip parameters can be assumed as the actual range of driving cycle parameters.

$$\text{Ratio of selected DC to all DCs} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{R_{dcp}^{s}}{R_{mtp}^{a}}\right) \quad \textbf{(6.2)}$$

where

i: set of driving cycle parameter, i ∈ {1,2, …, N}

$R_{dcp}^{s} = |pmax_i^s - pmin_i^s|$, range of ith selected driving cycle parameter (s) in an aggregation level

$R^{a}_{mtp} = |pmax_i^a - pmin_i^a|$, actual range (a) of corresponding microtrip parameter

The second indicator, equation (6.3), calculates the deviation of quartiles of the selected driving cycle parameters from quartiles of microtrip parameters.

$$\text{Distribution-based information loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M}\sum_{j=1}^{M}\frac{\left|PQ_{ij}^{s} - PQ_{ij}^{a}\right|}{\left|PQ_{ij}^{a}\right|}\right) \quad \textbf{(6.3)}$$

where

i: set of driving cycle parameter, i ϵ {1,2, …, N}

j: set of quarter, j ϵ {1,2, 3}

PQ_{ij}^{s} = The jth quartile of ith selected driving cycle (s)parameter in an aggregation level

PQ_{ij}^{a} = The jth quartile of actual range (a) of corresponding driving cycle parameter

Step 2: Determining "optimal" number of driving cycles

In the subsequent phase, both variability and information loss indicators are utilized to ascertain the "optimal" number of driving cycles. To do this, initially, a similarity matrix is constructed to gauge the likeness between all the created driving cycles. If two driving cycles exhibit a high degree of similarity, it implies that they are exhibiting analogous driving behaviors and therefore should be merged. Hence, the SciPy library in Python is employed to construct a similarity matrix accompanied by a dendrogram. The matrix delivers a pairwise comparison of driving cycles and the dendrogram represents a similarity level of driving cycles within the matrix.

Next, it is crucial to aggregate driving cycles exhibiting the highest similarity and then measure the impact of this procedure on both the variability of driving cycles and information loss. Variability reflects the diversification of the driving cycles; therefore, a decrease in variability signifies a decrease in the representativeness of the selected driving cycles. In simpler terms, the "optimal" number of driving cycles is represented by the smallest number that exhibits maximum variability and minimum information loss.

To measure the variability of driving cycles, this study employs the Proportional Variability (PV) indicator, as described in equation 6.4 and developed by (Heath and Borowski 2013). The domain of PV is on a truly proportional scale [0,1].

$$\mathbf{PV} = \frac{\sum_{\mathbf{p=1}}^{\mathbf{M}}\left(\frac{\mathbf{2} * \sum_{\mathbf{comb.}} \mathbf{D}(\mathbf{z_{pi}}, \mathbf{z_{pj}})}{\mathbf{n(n-1)}}\right)}{\mathbf{M}} \qquad \textbf{(6. 4)}$$

Where i,j be indices such that i,j ∈ {1, 2, ..., n} and i≠j. These indices are used to calculate the relative differences between the parameters of driving cycles. For a given parameter p, z_{pi} represents the value of p in the i-th driving cycle, and z_{pj} represents the value of p in the j-th driving cycle. The function $D(z_{pi}, z_{pj}) = \frac{|z_{pi} - z_{pj}|}{Max(z_{pi}, z_{pj})}$ is defined to quantify the relative difference between

these two values. n is the total number of driving cycles and M is total number of driving cycles parameters.

6.4 Result

This section examines the results derived from implementing the established framework methodology, which is aimed at determining the “optimal” number of driving cycles.

6.4.1 Mixed models to study the influences of fixed effects on driving cycle parameters

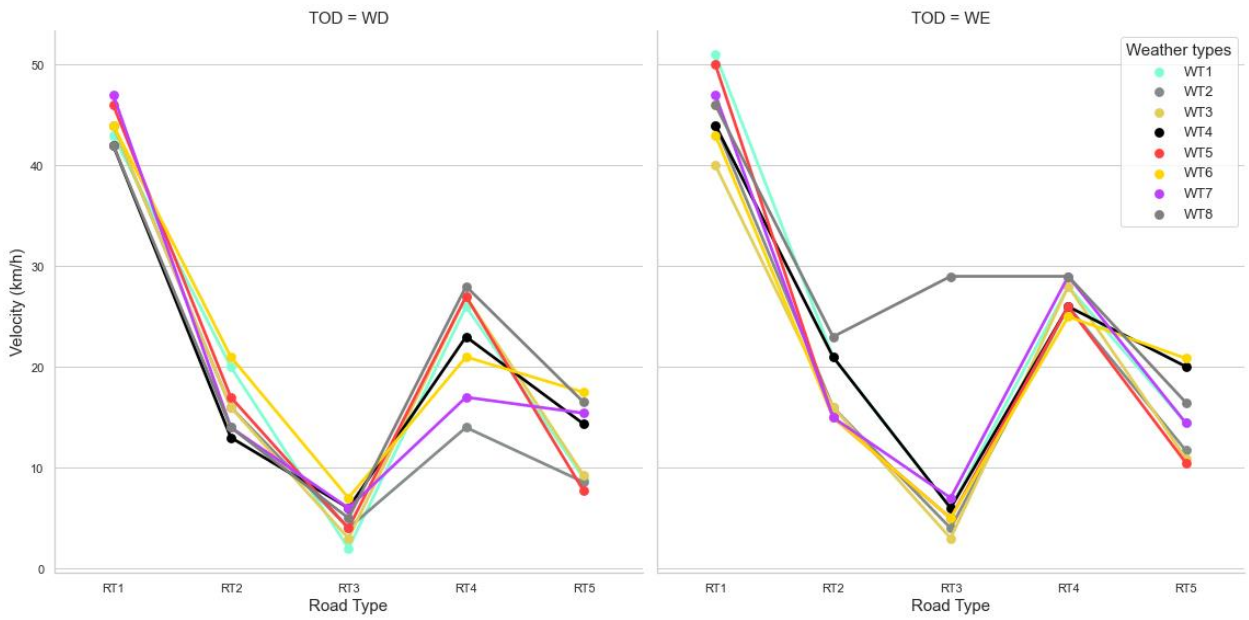
As previously discussed, we have formulated a mixed model to investigate the interaction between external factors. This interaction between fixed effects represents the actual driving conditions because in the real world, driving behaviors are influenced by roads, weather, and day of the week. As a result, considering the interaction between these exogenous factors can lead to a better understanding of driving behaviors under realistic driving conditions.

Figure 6-4 displays the results of this model for all driving cycle parameters. The x-axis represents the road types, and the y-axis shows the driving cycle parameters for different weather conditions and weekdays/weekends.

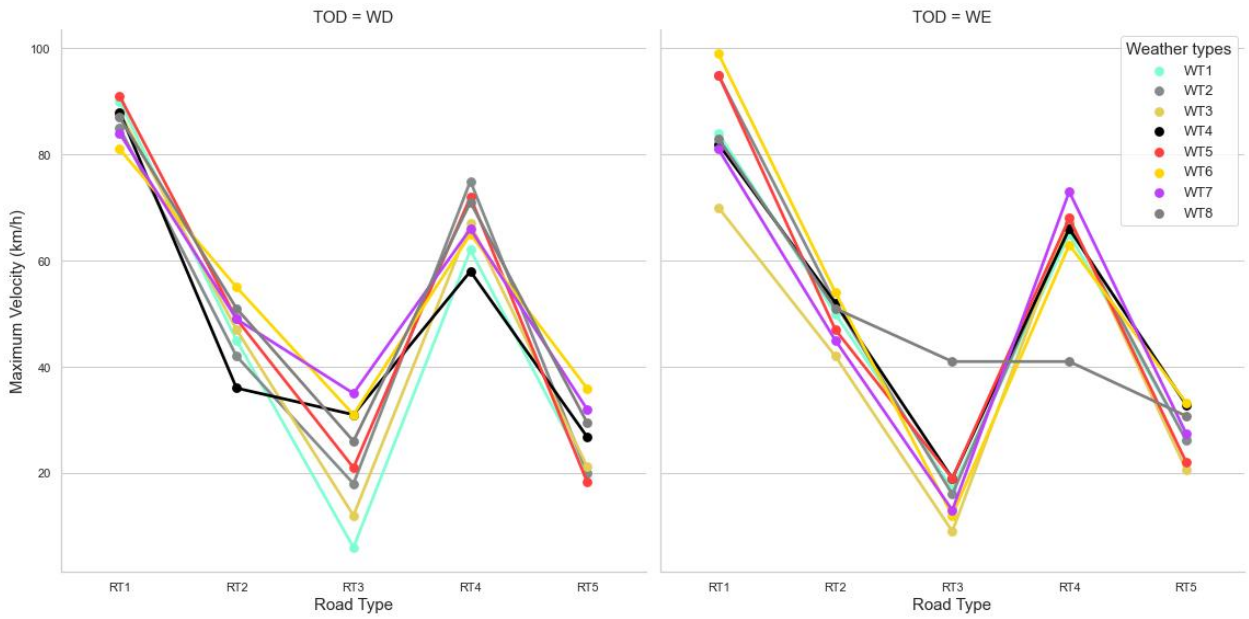
In all graphs, we can see that the driving cycle parameter values change with variations in fixed effects. This supports our premise that driving cycle parameters are functions of road type, weather type, and weekdays/weekends.

Additionally, the slope of the lines in these graphs can indicate the direction and magnitude of the relationship between the road type and driving cycle parameters, given a specific weather condition and time of day. For instance, a steep upward slope signifies a strong positive relationship, suggesting that the creeping time increases as the road type changes from one category to another. A flat line indicates no relationship, whereas a downward slope signifies a negative relationship.

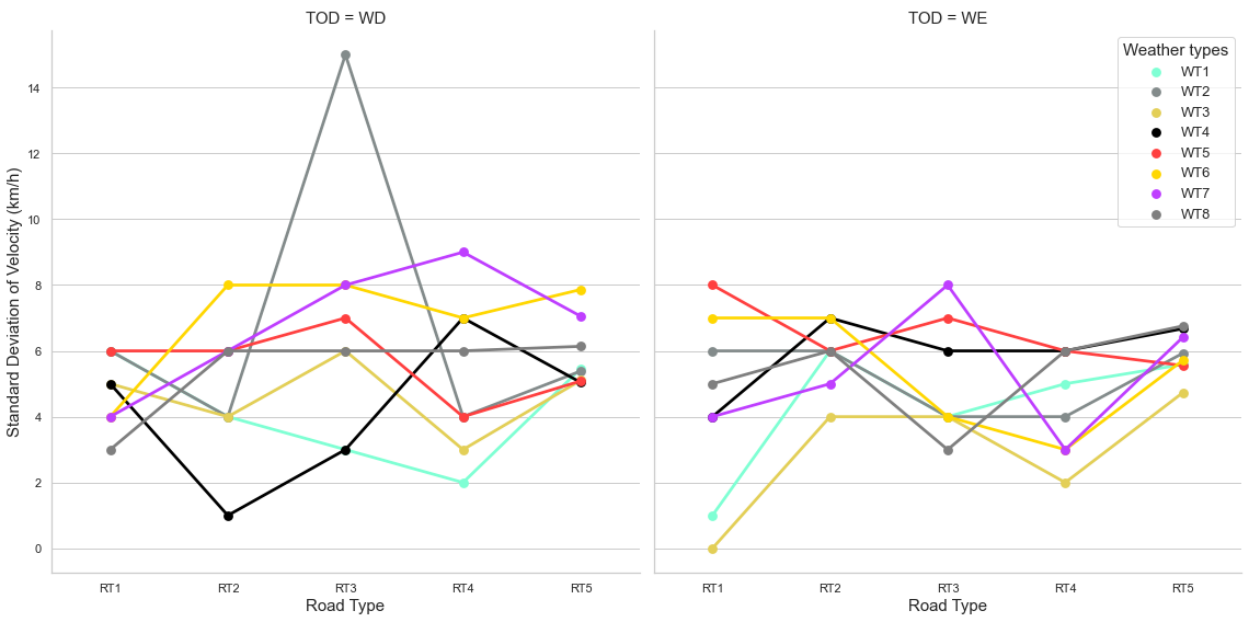
Upon examining the generated graphs, it's clear that driving cycle parameters change for all road types, weather conditions, and days of the week. For example, velocity decreases significantly from road type 1 to road type 2 in all weather conditions.



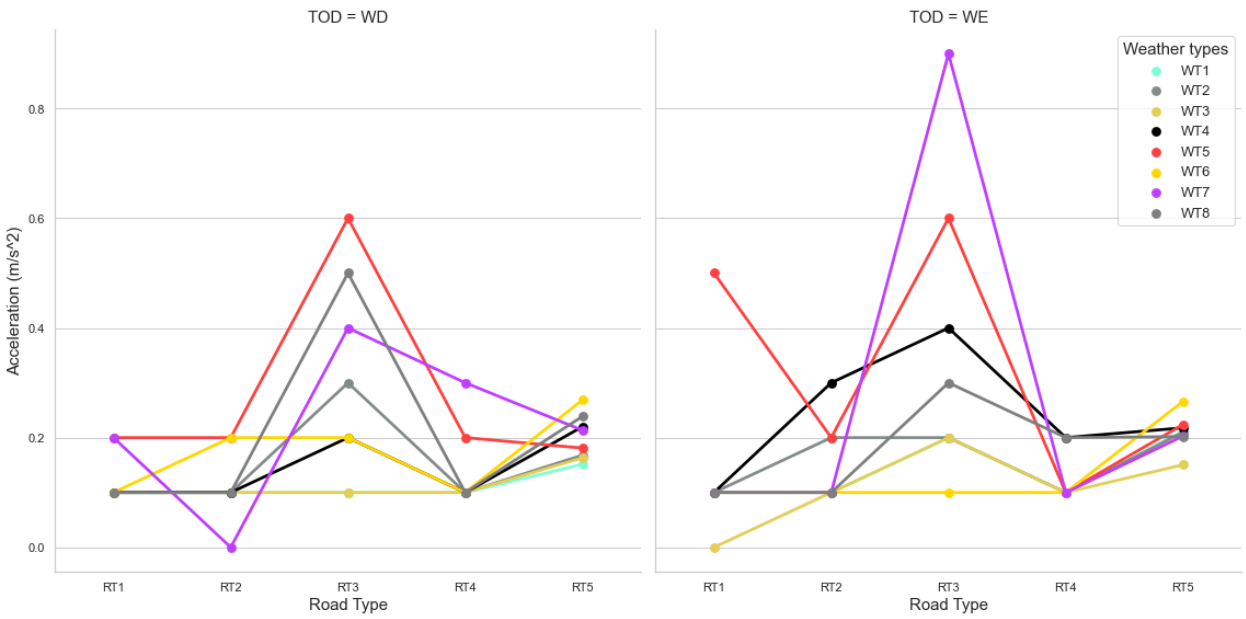
(Velocity)



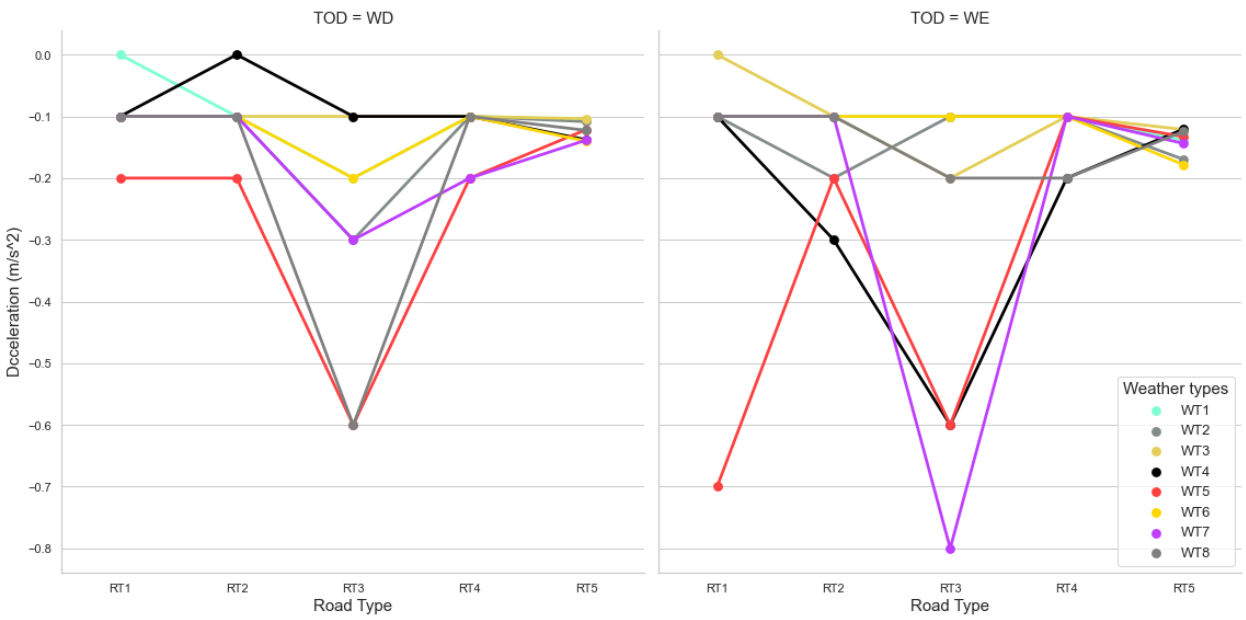
(Maximum velocity)



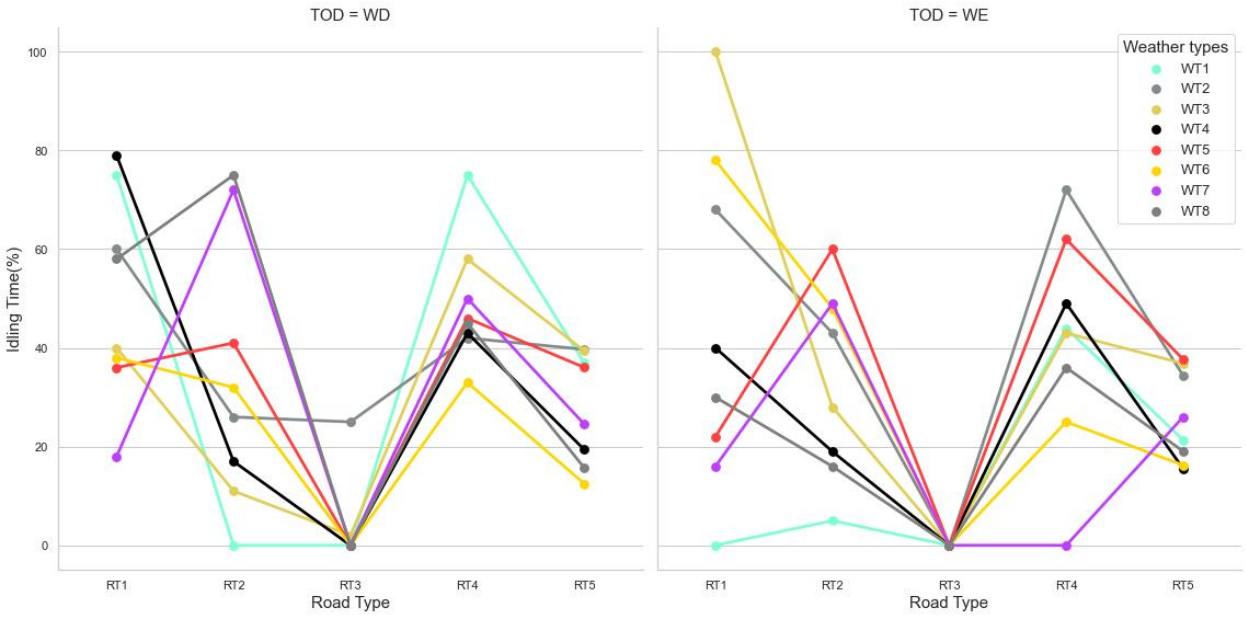
(Standard Deviation of velocity)



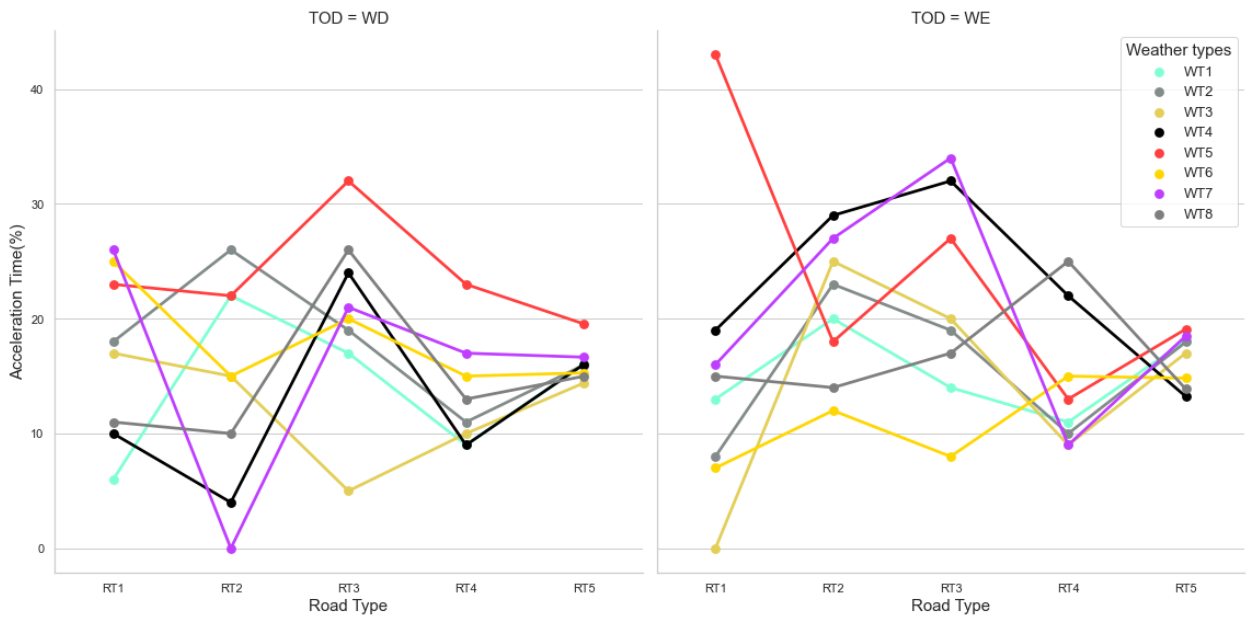
(Acceleration)



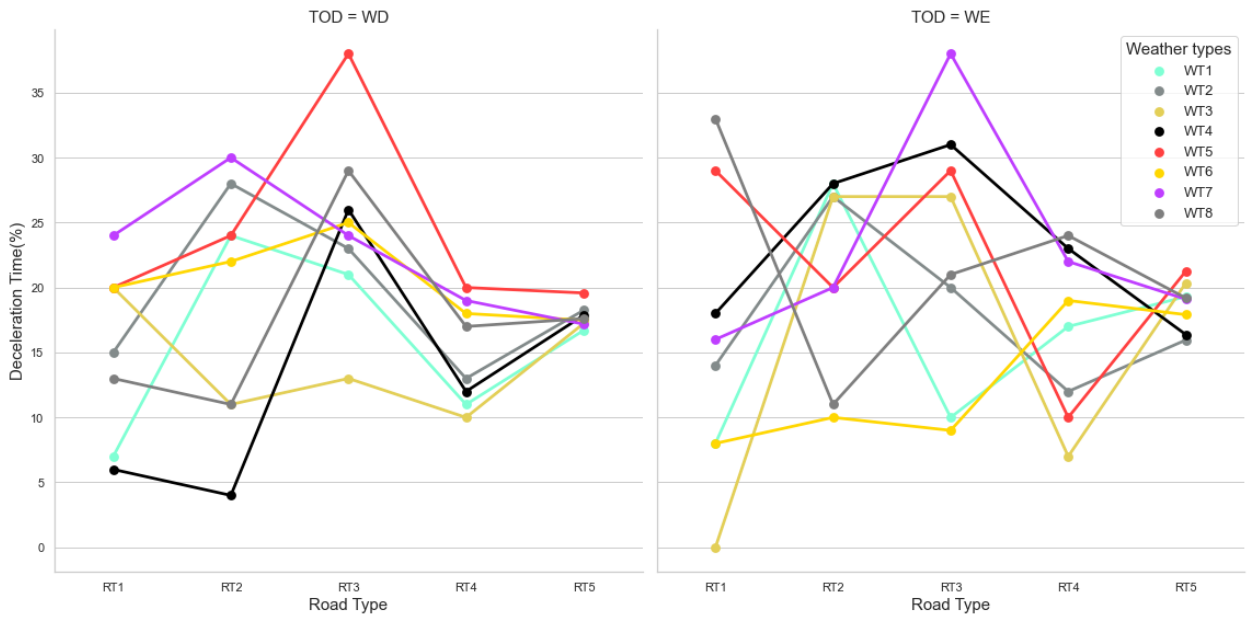
(Deceleration)



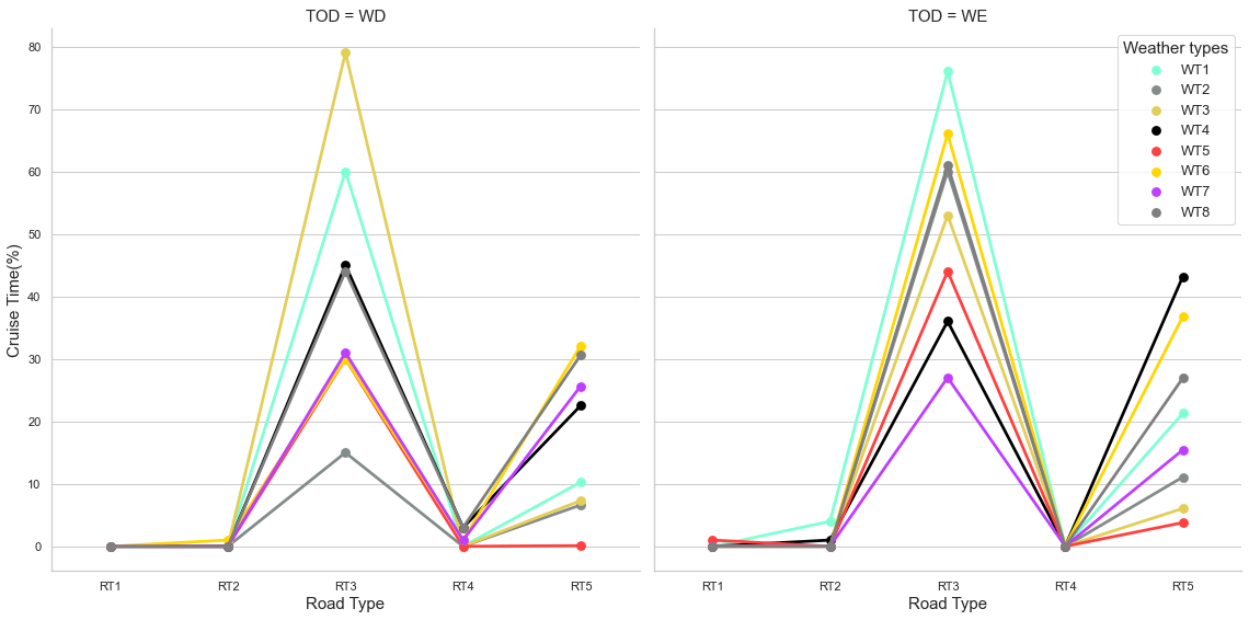
(Idling Time)



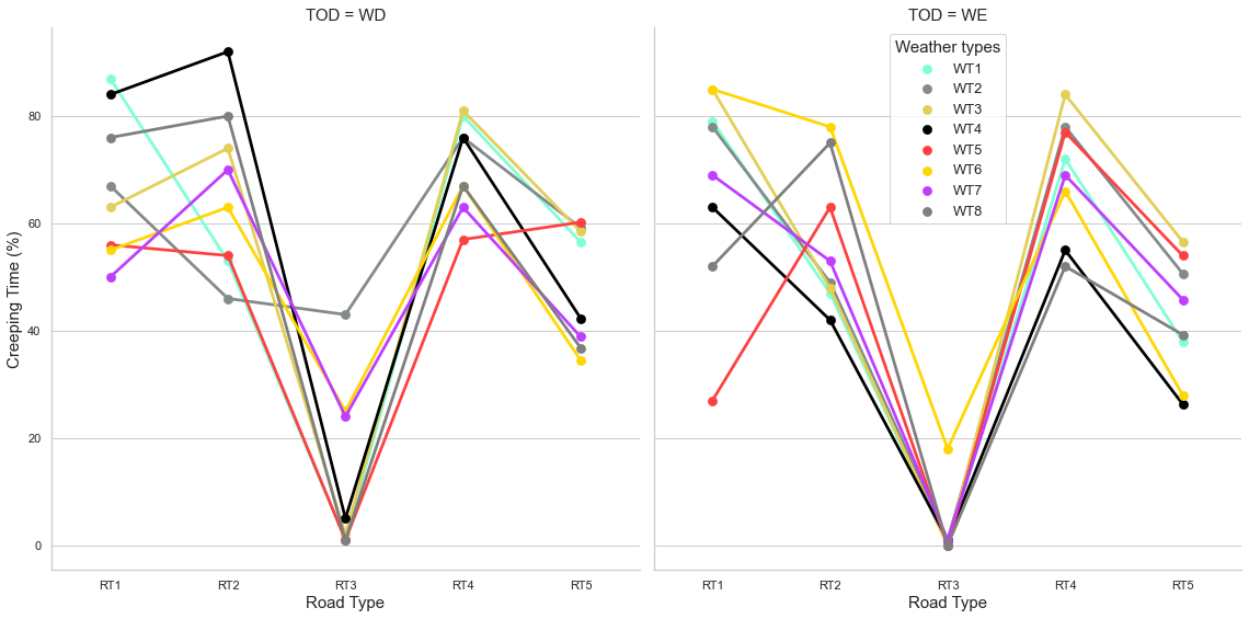
(Acceleration Time)



(Deceleration Time)



(Cruise Time)



(Creeping time)

Figure 6-4 Changes of driving cycle parameters in different road types, weather types in weekend and weekdays

Table 6-4 presents the most significant fixed effect for each driving cycle parameter. Notably, the p-value for all factors was less than 0.05, indicating a statistically significant association between the driving cycle parameters and fixed effects.

As equation 6.1 demonstrates, the constructed mixed model examines the relationship between fixed effects and driving cycle parameters independently. This enables the model to explore the impact of a single factor while controlling for the effects of the others. Furthermore, it assesses the impact of interactions between fixed effects on driving cycle parameters.

Table 6-4 The most important exogenous factors for each driving cycle parameters

Driving cycle parameters	The most important influencing factor according to highest coefficient			
	Fixed effects			Real driving conditions
	Road type	Weather type	Days of week	Combination of fixed effects
V	RT3	WT7	WE	RT3WT8WE
V-std	RT4	WT8	WD	RT3WT2WD
Vm	RT5	WT6	WD	RT3WT6WD
Acc	RT5	WT5/WT7	WE/WD	RT3WT7WE
Dcc	RT3	WT5	WD	RT3WT4WD RT3WT7WE
Idle-p	RT2, RT3	WT7	WD	RT3WT3WE
Acc-p	RT2	WT7	WE	RT2WT7WE
Dcc-p	RT2	WT7	WE	RT3WT3WE
Cru-p	RT3	WT3	WD	RT3WT3WD
Cre-p	RT3	WT7	WE	RT3WT3WE

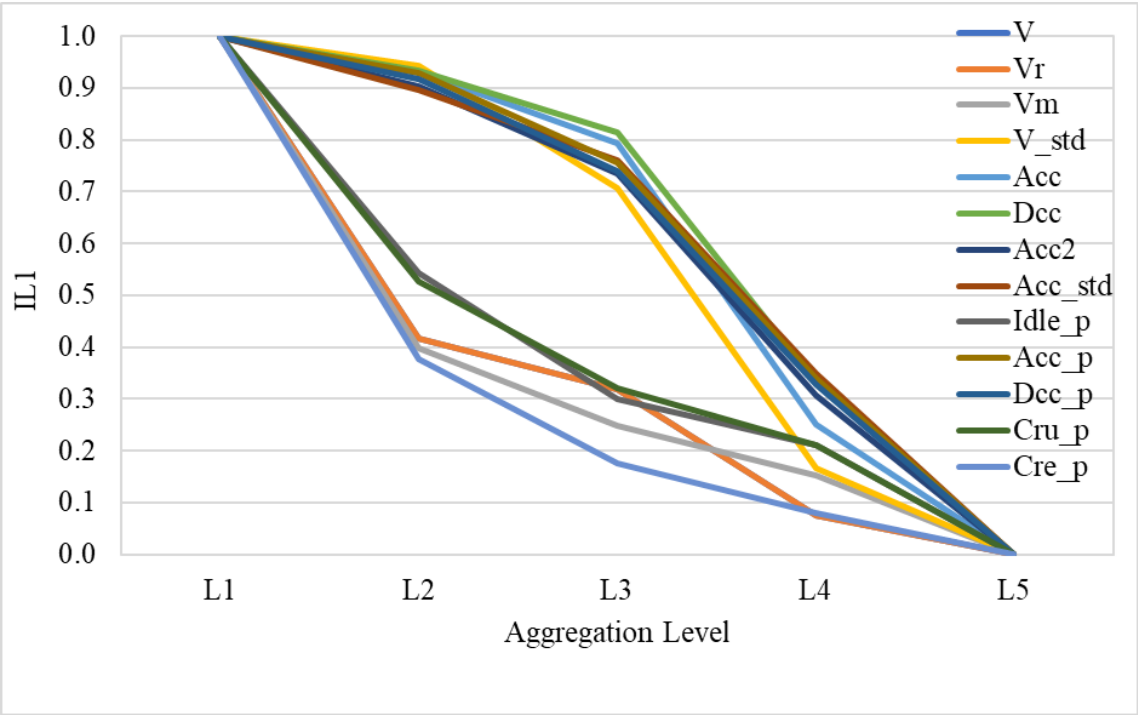
The results underscore the importance of considering all fixed effects simultaneously when analyzing driving cycle parameters. For instance, regarding velocity, the most influential isolated factors are RT3, WT7, and WE, while in real driving conditions, the interaction RT3WT8WE has the most significant impact.

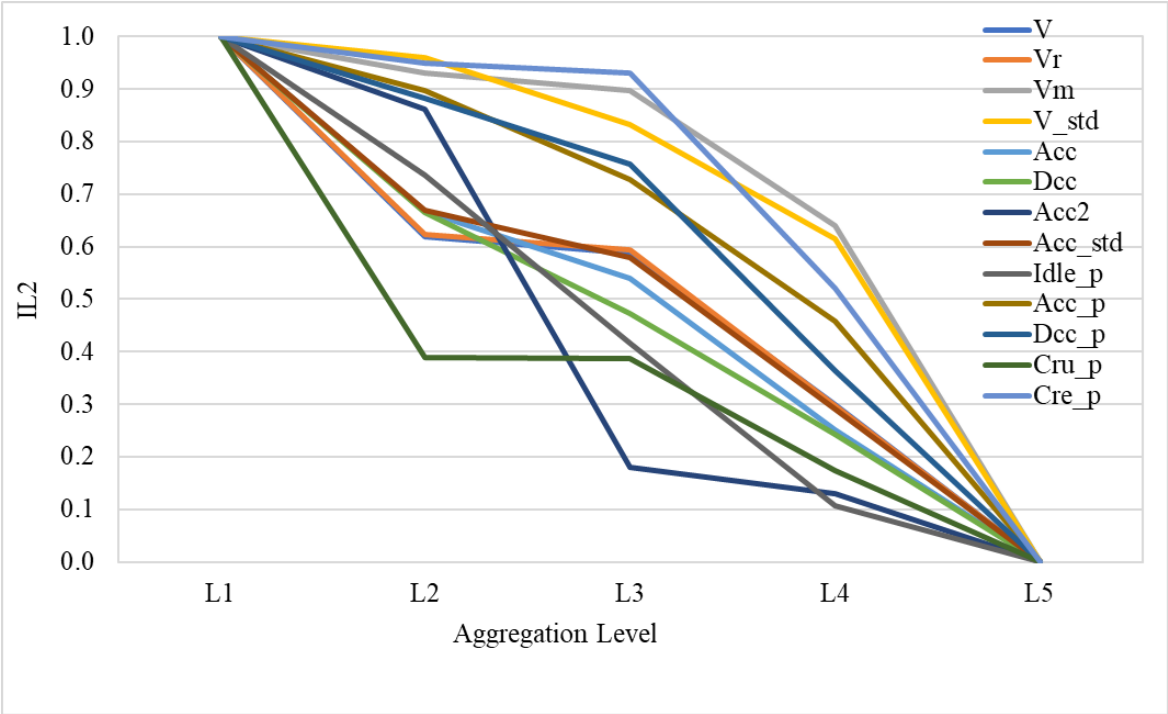
6.4.2 Determining aggregation level for the developed methodology

As elaborated in section 6.3.6, the initial step to identify the “optimal” number of driving cycles involves using information loss indicators to ascertain the ideal level of data aggregation. This involves following the methodology delineated in section 6.3.4 for five different levels of data aggregation. These are:

1. The highest level of aggregation where all driving data are amalgamated and only one driving cycle is produced.
2. The second level where data are grouped according to road types, resulting in 5 driving cycles.
3. The third level where data are aggregated based on different driving conditions in various road types, resulting in a total of 40 driving cycles.
4. The fourth level uses the recommended methodology of this study where data are grouped according to road types, weather conditions, and weekdays or weekends, resulting in 80 driving cycles.
5. The fifth level involves the generation of microtrips for each trip.

Information loss is calculated using equations 6.2 and 6.3. Figure 6-5 visually presents the variation of information loss for different driving cycle parameters at the defined levels. The information loss values on Figure 6-5 range from 0, signifying no information loss, to 1, representing the highest level of information loss. It is clear that the highest and lowest information loss levels are observed for levels L1 and L5 respectively. The results corroborate that the methodology presented in this study results in a minimum possible level of information loss. It is worth noting that Figure 6-5 display information loss for indicators 6.2 and 6.3, and the overall information loss of all driving cycle parameters, respectively. The latter allows us to determine the average information loss using average values for all driving cycle parameters for different levels of aggregation.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
IL1
L1
L2
L3
L4
L5
Aggregation Level
V
Vr
Vm
V_std
Acc
Dcc
Acc2
Acc_std
Idle_p
Acc_p
Dcc_p
Cru_p
Cre_p


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
IL2
L1
L2
L3
L4
L5
Aggregation Level
V
Vr
Vm
V_std
Acc
Dcc
Acc2
Acc_std
Idle_p
Acc_p
Dcc_p
Cru_p
Cre_p

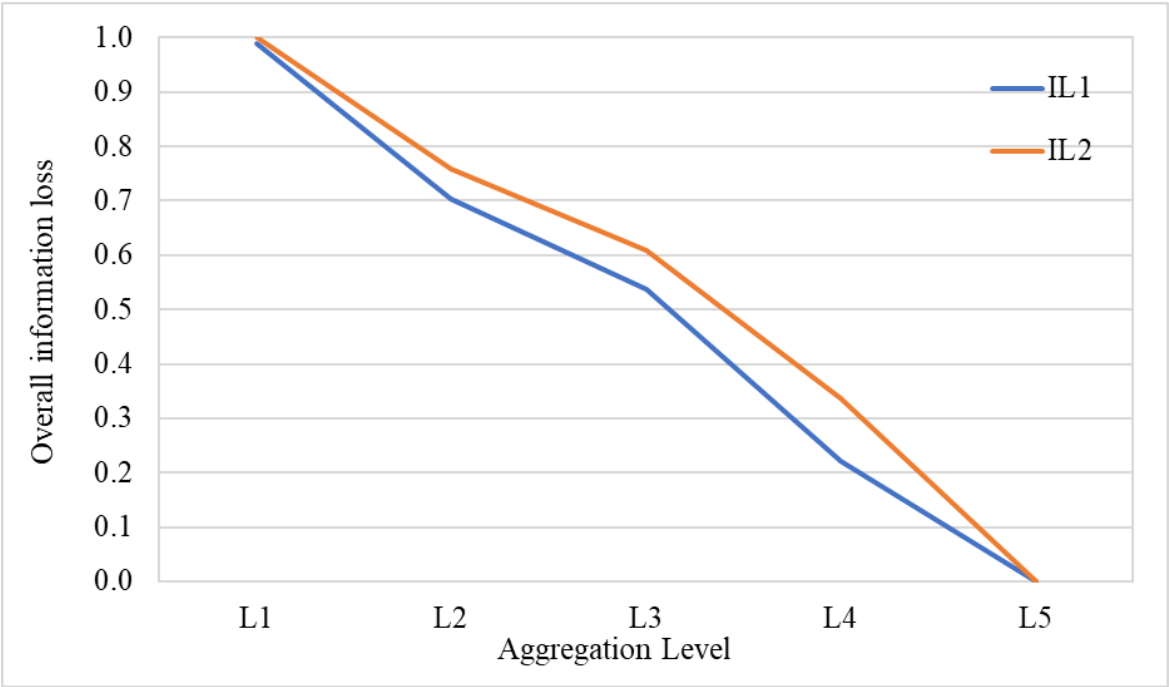


Figure 6-5 Comparison of information loss in different aggregation level for information loss indicators and overall information loss

6.4.3 Determining the "optimal" number of driving cycles

In order to determine the "optimal" number of driving cycles, it is critical to assess the similarity of the produced driving cycles. This is accomplished by creating a similarity matrix, where each array illustrates the similarity between two driving cycles. Figure 6-6 presents a heatmap of the similarity matrix, with similarity intensifying from dark to light shades. Moreover, the associated dendrogram indicates levels of similarity. The dendrograms in both sides of the matrix are same and used to have clearer presentation of similarity level. The value in each cell represent similarity between two driving cycles. The highest similarity observed is 92%, between RT4WT1WE and RT4WT2WE, while the lowest similarity is 0%. Figure 6-7 provides a breakdown of the different levels of similarity. The similarity levels of less than 40% have the largest share which means that most of the generated driving cycles are representing different driving behaviours.

To identify the "optimal" number of driving cycles, the authors utilized the similarity matrix to pinpoint and merge driving cycles exhibiting the highest similarity at each similarity level. Essentially, within each level of similarity, driving cycles that share the highest degree of similarity are merged into one representative cycle. For instance, at the 69% similarity level, four driving

cycles are present, hence each pair of similar driving cycles their driving data are aggregated, and a new driving cycle is calculated as a substitute for them, yielding a total of two derived cycles. Subsequently, information loss and variability are calculated using equations 6.2, 6.3 and 6.4 respectively. This recursive process continues until the variability reaches its maximum and information loss reaches its minimum.

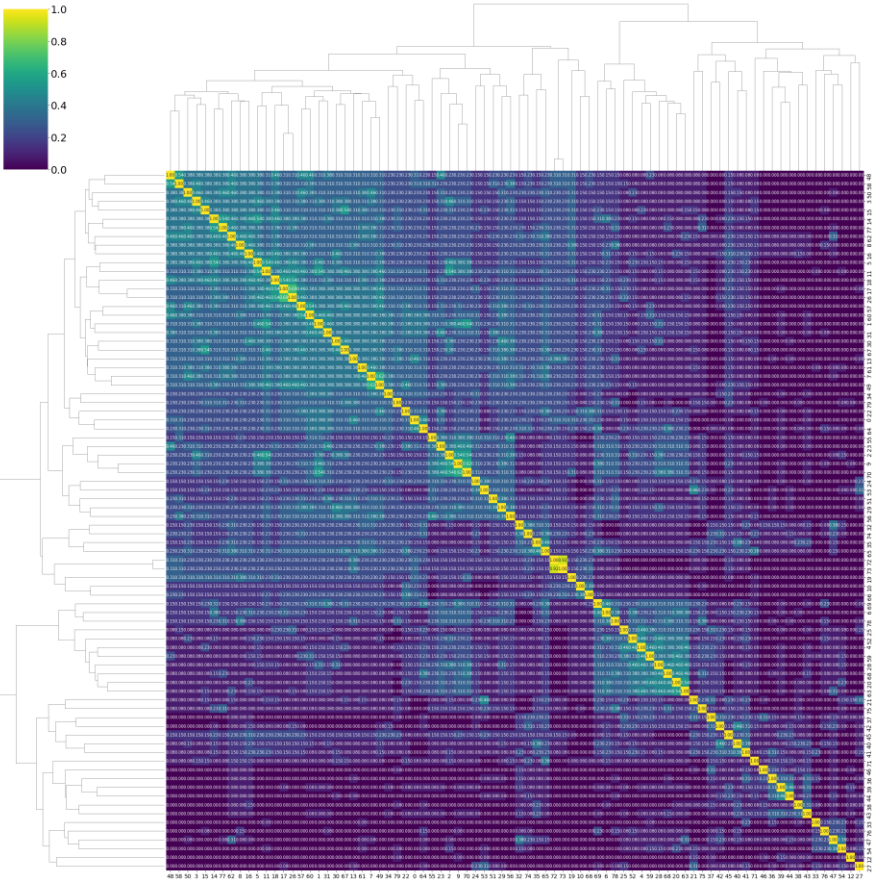


Figure 6-6 Similarity matrix of the generated driving cycles

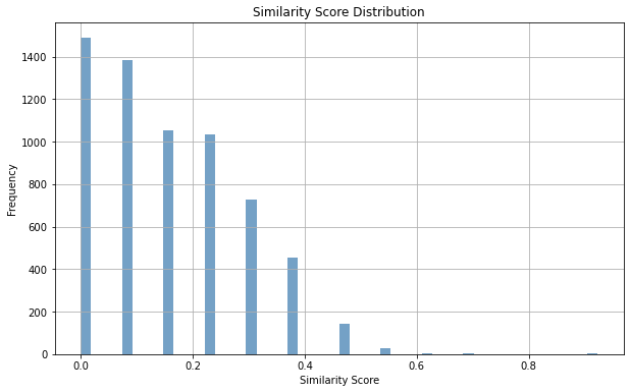


Figure 6-7 Share of similarity level in the similarity matrix

Figure 6-8 presents the changes in information loss, according to average of both indicators, and variability across different similarity levels. The blue graph illustrates the fluctuation of information loss across various similarity levels. Notably, as the similarity level decreases from 92% to 62%, information loss reaches its minimum. Subsequently, as merging continues, information loss begins to increase. Moreover, the red graph displays an increase in variability as the similarity level decreases from 92% to 62%, moving from 0.6734 to its peak at 0.6896, and then decreasing to 0.6667, maintaining this trend at lower similarity levels. The red circle highlights the point of minimum information loss and maximum variability, occurring at the 62% similarity level.

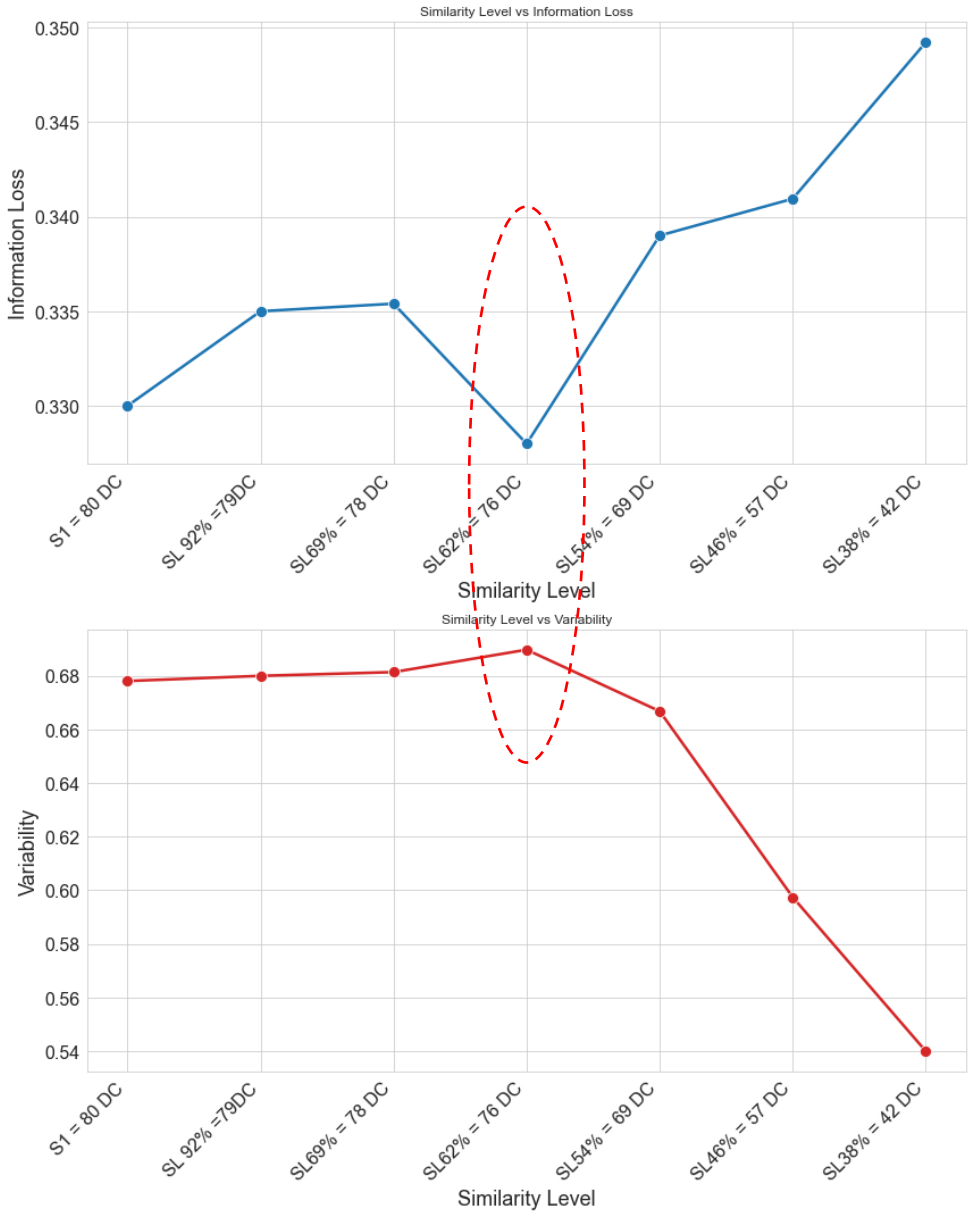


Figure 6-8 Similarity level, driving cycle number vs. information loss and variability

In conclusion, by considering the results of both information loss and variability, it can be determined that the “optimal” number of driving cycles required to represent driving behaviors across all road types and all driving conditions in Montreal is 76.

6.5 Conclusion

This research paper aimed to address three key gaps in the literature on driving cycle generation: determining the “optimal” number of driving cycles, proposing a novel methodology for the

aggregation of driving data, and developing a mixed model to examine the influence of exogenous factors on driving cycle parameters.

Firstly, a driving cycle is a dynamic construct with parameters that change over time and location. However, in the existing literature, typically one or a few driving cycles are used to represent driving behaviors. In response to this, this study developed a novel methodological framework to create and determine the “optimal” number of driving cycles that adequately represent driving behaviors in all driving conditions across an urban street network. The results suggest that 76 driving cycles can capture the maximum variability in driving behaviors, thereby minimizing the uncertainty arising from information loss.

Secondly, data aggregation causes difference between real driving behavior and generated driving cycle which can be a source of uncertainty in the process of generating driving cycles. Many studies do not take into account the impact of data aggregation on information loss and the accuracy of the generated driving cycle, often aggregating all driving data collected from all types of roads and weather conditions. This study proposed an approach that uses two indicators to determine the ideal level of data aggregation for generating driving cycles. This methodology seeks to minimize information loss and enhance the representation of the generated driving cycles.

Thirdly, this study pioneered the development of a multilevel mixed model to probe the influences of road types and weather conditions on driving cycle parameters. The mixed model successfully demonstrated these intricate relationships, enhancing our understanding of the dynamic nature of driving cycles.

The primary constraints of this study relate to time and data. Concerning time, there exist other external elements such as vehicle type, driver gender, road conditions, and traffic flow, all of which can interact with roads and weather, thus altering driving behaviors. However, the time constraints of this study did not permit the inclusion of these factors. Furthermore, this study relies on taxi data to generate driving cycles, while access to passenger car data could potentially enhance the representativeness of driving cycle. The study's findings must therefore be interpreted within the context of these limitations.

CHAPTER 7 GENERAL DISCUSSION

In this chapter, the first section contributes to the general objective of the thesis and follows by reviewing the main objectives, and different results.

7.1 Recall of the general objective of the thesis

A driving cycle reflects average driving behaviors and is vital for linking driving behaviors with vehicle emissions, especially in emission models like MOVES. It is a specific sequence of speed-time data essential for assessing vehicle performance, including emissions and fuel consumption. They mimic real-world driving scenarios, allowing manufacturers and regulators to assess vehicle emissions and fuel efficiency in a controlled, representative setting. Thus, driving cycles play a significant role in vehicle design, regulation, consumer preferences, and environmental policies. However, the process of creating driving cycles faces uncertainties, sometimes resulting in cycles that do not accurately reflect driving behaviors.

Sections 2.1 and 6.2 describe various methods developed to increase the accuracy of driving cycle creation, but uncertainties remain. It appears the ongoing issue stems from researchers' perspectives on transportation systems. Many past efforts focused on creating new, more intricate algorithms for analyzing driving data. However, drivers, affected by weather and road conditions, behave differently across diverse road segments and weather situations. These elements influence traffic, vehicle performance, and drivers' driving behaviors. To bridge the gap between actual driving behaviors and the driving cycles created, it is crucial to consider these factors when constructing driving cycles. Essentially, a key step in reducing uncertainty in vehicle emission estimates is understanding how many driving cycles are needed to accurately represent driving behaviors in city settings.

This thesis proposes a data-driven framework to enhance the accuracy of generating driving cycles by considering these external influences. It presented a methodological approach to determine the ideal sets of driving cycles. Here, "optimal" means the minimum driving cycles that show the most diversity in driving behaviors with the least loss of information.

7.2 Discussions of results

The current thesis unfolded in three primary phases: categorizing road network segments, utilizing machine learning techniques to group hourly weather data, and determining the "optimal" number

of driving cycles. Each phase possesses distinct objectives and offers unique contributions. The subsequent subsections provide a detailed discussion on each of these aspects.

7.2.1 Developing a new approach to categorize road network segments

The main objective of this phase was to develop a new typology to categorize road network segments according to the features affecting driving speed. A road segment can be characterized by various elements such as gradient, number of lanes, signalized intersections, and so on. Road features have different impacts on transportation systems from affecting driving behaviors, safety, travel time, traffic flow, fuel consumption and vehicle emissions. This phase sought to address three gaps in research regarding the relationship between road features and driving behaviors.

Initially, most studies categorized road sections based on their primary function. However, evidence suggested that driving speed and vehicle emissions are influenced by numerous road characteristics. In response, this research introduced a data-driven system, utilizing machine learning, to classify road segments based on key features impacting driving speed. The Montreal Road network was chosen for analysis, and the results indicated that using k-means clustering with the Gower similarity measure generated better result, dividing the segments into five road types.

A notable strength of this method was its adaptability to different study area sizes, making it suitable for city roads, irrespective of the street network's size.

The next challenge concerned the verification of clustering outcomes. Research showed that specific data types might skew traditional clustering validation methods, like the Silhouette measure. As a result, an extra validation step, specifically tailored for vehicle speed data, was introduced to enhance result accuracy. After grouping the road segments, five speed data sets collected for each road category. Tests like ANOVA, Levane, and post hoc analysis highlighted noticeable speed differences across these road categories.

The final contribution of this study centered on crafting models to elucidate the connection between road characteristics and driving speed. The modelling revealed that using the entire street network cannot explain accurately association between road characteristics with speed variations. However, constructing a regression model tailored to each road category led to a more reliable R^2 value.

The study's outcomes highlighted that a blend of road characteristics played a crucial role in driving speed. Unlike past research that pinpointed specific features as the primary influencers (leading to

mixed findings in existing literature), this research underscored that influential factors vary depending on the road type.

7.2.2 Generating a weather typology

The intricate interplay between meteorological conditions and driving behaviors have been rarely investigated. The second phase primarily addresses two main concerns related to driving cycle variations under different weather conditions:

1. The need for a more thorough examination of how weather affects driving cycle parameters.
2. The lack of a systematic approach to classify various weather types concerning vehicle emissions.

First, a MMRM approach was employed to discern the primary weather factors influencing driving cycle characteristics. The findings revealed that temperature affects deceleration, while precipitation influences velocity. Wind speed impacts acceleration, and humidity has a less pronounced effect on driving cycle parameters. For a clearer picture, the effects of these weather conditions were analyzed for a specific road type before the regression analysis. The insights from this model provided a new perspective on how weather conditions interacted with driving behaviors, which will be especially valuable for studies looking at the implications of climate change on driving.

Next, an unsupervised machine learning approach was used to categorize weather conditions based on their impact on driving cycle parameters. This analysis identified 8 unique weather types within the study region.

Lastly, a MMRM approach again was developed and showed generated weather types in comparison with the before clustering explains better association between generated weather types and driving cycle parameters.

7.2.3 Determining “optimal” number of driving cycles

The main objective of this phase was to develop a new machine learning approach to accurately identify the ideal number of driving cycles in an urban setting. This phase sets out to tackle three primary gaps in the driving cycle generation literature: pinpointing the ideal number of driving cycles, introducing a fresh methodology for driving data aggregation, and crafting a mixed model to study the effects of external factors on driving cycle elements.

To begin with, a driving cycle is not static; its parameters shift depending on time and location. Yet, many studies in existing literature often have been relied on just a single or a handful of driving cycles to depict driving behaviors. Addressing this, present study introduced a novel method to both create and identify the best number of driving cycles that truly represent driving behaviors under all conditions within an urban street network. The findings pointed out that 76 driving cycles can encompass a broad range of driving behaviors, thereby reducing the uncertainty linked to information loss.

Next, the way data is aggregated can influence the reliability of driving cycles. Many research efforts seemed to overlook the effects of data aggregation on the clarity of information and the accuracy of the resulting driving cycle. They often group all driving data, regardless of road types or weather conditions. To refine this process, this study suggested a method that utilizes two indicators to designate the best level of data aggregation for crafting driving cycles. The aim here was to lessen information loss and boost the accuracy of the resulting driving cycles.

The third part of this study ventured into territory by creating a multilevel mixed model to delve deep into how road types and weather conditions affect driving cycle elements. The results highlighted the importance of understanding the ties between changes in driving cycle elements and external influencers, especially when considering the interplay between road types and weather conditions. The mixed model showcased these complex connections, enriching our grasp of driving cycle dynamics.

However, the study had its fair share of challenges, mainly tied to data. There are other influential factors, such as vehicle model, driver's gender, road surface conditions, and traffic flow, that could impact driving patterns when combined with roads and weather. Yet, data unavailability on features such as driver’s gender or vehicle types of driving data prevented these factors from being included in this study. Additionally, the study hinged on taxi data to sketch out driving cycles. Having data from regular drivers might have fine-tuned the driving cycle estimates. Therefore, it is essential to view the results of this study in light of these constraints.

7.2.4 General discussion on the contributions

One key point when using the method from this thesis in another area is the connection between different papers. In the first paper, we identified several types of roads. The accuracy and the number of these road types were crucial for the following papers. This is because we used several

methods in sections 4.4.3 and 4.4.4 to check the road type classifications were correct. The number of road types and the number of segments in each type were important because they affected how we picked samples, how long it took to process the MMRM, and how many driving cycles we needed for each weather type in the second paper. More road types meant more processing time and more driving cycles needed, and fewer road types meant less.

The accuracy of how we grouped road segments also influenced which driving data we chose for each weather type, which could change the driving cycle parameters.

Additionally, the results from the first and second papers were necessary before starting the third paper. In other words, it could not be possible to proceed with the third paper without the road and weather classifications from the first two papers. The first and second papers also affected how similar the data was in the matrix in section 6.4.3 and the results of the mixed model in section 6.4.1. Any errors, in terms of cluster numbers, in the road classifications in section 4.4.2 or the weather classifications in section 5.4.2 could lead to a significant difference from the actual driving behaviors for specific road and weather types. To prevent this, the thesis used different validation techniques, as discussed in sections 4.4.1, 4.4.3, 4.4.4, 5.4.1 to ensure the quality of the road and weather classifications.

7.3 Summary of limitations

However, like all scientific endeavors, this thesis is not free from limitations. The most prominent being data availability. Availability of data is related to all three datasets discussed in section 3.1.

7.3.1 Data

A significant limitation in this study is related to data availability, as detailed in section 3.1. For instance, the finest resolution of the Digital Elevation Model (DEM) at our disposal was 30 meters. Utilizing a higher resolution, such as 15 meters, would likely enhance the precision in assessing road gradients for specific road segments. Similarly, the data sourced from Environment Canada amalgamates both rainfall and snowfall under a singular "precipitation" column. Bifurcating this column into distinct entries for rainfall and snowfall would furnish a more granulated understanding, potentially enabling a more nuanced interpretation of their respective impacts on driving cycle parameters.

Furthermore, several pivotal factors which could potentially influence driving cycle parameters were conspicuously absent from our dataset. This includes variables like vehicle types, traffic flow dynamics, road surface conditions and driver's age. Access to such data would have permitted a more comprehensive analysis, shedding light on the correlations between these variables and the variability of driving cycle characteristics.

Such data limitations may have constricted our ability to capture the full spectrum of factors affecting driving cycle variability. This, in turn, could impact the generalizability and applicability of our findings to broader scenarios or contexts.

7.3.2 Geographical Specificity

The approach developed in this research is tailored to the unique urban layout, traffic patterns, and climatic conditions of Montreal Island. Although the developed framework is flexible enough to be employed in other geographic region, the deduced “optimal” number of driving cycles may not directly extrapolate to other urban regions with distinct geographical configurations, driving habits, infrastructure considerations, or climatic variances. It would be prudent for future investigations to evaluate the methodology's adaptability across diverse urban environments to ensure its broader applicability.

7.3.3 Methodology

The methodology devised in this research inherently necessitates a substantial volume of data to effectively determine the “optimal” number of driving cycles, given its comprehensive scope spanning various driving conditions and road types. As corroborated by existing literature, the procurement of such expansive driving data is both time-intensive and financially demanding.

Moreover, the execution of the established methodology demands robust computational resources. This heightened requirement is attributable to the intricate processes involved: namely, the spatiotemporal filtering requisite for isolating pertinent data corresponding to each specific road and meteorological condition, and the subsequent generation of driving cycles. The computational intensity of these operations can pose challenges in terms of efficiency and feasibility, especially when processing vast datasets.

The aforementioned constraints underscore the importance of ensuring adequate data acquisition strategies and computational infrastructure when attempting to implement and benefit from the presented methodological framework.

CHAPTER 8 CONCLUSION

8.1 Summary of the thesis

In conclusion, this research has shown that 76 driving cycles accurately represent the variety of driving patterns found throughout Montreal under various conditions. This was accomplished by using unsupervised machine learning techniques to categorize different road and weather conditions. An in-depth review of the literature, coupled with an analysis of extensive driving data, has deepened the author's understanding of driving cycle parameters and how they vary with road and weather types. As a result, effective methods have been developed to evaluate the loss of information and the level of detail in the driving cycles created, achieving the primary goal of this thesis.

The study also highlights the necessity of recognizing shifts in driving cycle components, considering the interplay between road types and weather conditions. The findings from this research are valuable for future studies and can be applied in the fields of traffic engineering and urban transport planning.

8.2 Future research

Looking forward, the thesis sets the stage for some relevant future research directions.

8.2.1 Climate change and driving cycle parameters

Given the reality of climate change, understanding its impact on driving cycle parameters becomes paramount. Driving cycle parameters characterize vehicle operation through variables such as speed, load, acceleration, deceleration, and idling. These parameters play a pivotal role in shaping fuel consumption and emissions.

Direct climatic influences on driving patterns are varied. Temperature extremes, both at the high and low ends, can have pronounced effects on vehicle performance. Hot conditions might spur increased use of air conditioning, which can elevate fuel consumption, while cold temperatures can decrease battery efficiency in electric vehicles and raise the viscosity of lubricants. This leads to greater energy expenditure. Moreover, precipitation in the form of rain, snow, or ice can modify driving behavior, often prompting drivers to reduce their speed. Even strong winds can challenge vehicle stability, especially for those vehicles with a high profile.

Infrastructure also comes under the purview of climate change. As weather patterns shift, roads may experience accelerated degradation. This could lead to more frequent roadworks, which, in turn, impacts the flow of traffic and the typical driving cycle. Furthermore, if certain climatic events, such as heatwaves, become commonplace or more intense, drivers might adapt their behavior over time. A potential scenario is one where heatwaves deter daytime travel, resulting in altered traffic patterns.

For policymakers, predictive modeling in this domain would be invaluable. The initial step in constructing such a model would be to amass data concerning both climatic and driving cycle parameters across a temporal spectrum. Modern models often employ machine learning algorithms to discern patterns and correlations within vast datasets. However, the validation of this model, using independent datasets, remains crucial to ensure its applicability and accuracy.

Yet, this research area is not without its challenges. Reliable data on driving cycles might be sparse, and there's the inherent complexity of merging disciplines like climate science with transportation engineering. Nevertheless, given the pressing nature of climate change and its potential ramifications on every aspect of human life, this research is both timely and essential.

8.2.2 Considering the geographical specificity of the study

Every city is unique in its infrastructure, topography, culture, and, importantly, its climatic conditions. When a study is geographically specific, the findings and outcomes often reflect the peculiarities of that particular environment. As such, there's always a question of how transferable these findings are to other cities or regions. The inherent characteristics of one city can influence many facets of daily life, from urban planning to transportation patterns, and this is especially true when considering the effects of climate.

For instance, a study conducted in a city with a predominantly temperate climate might not account for the unique challenges posed by cities with more extreme conditions, be it tropical humidity, arctic cold, or desert heat. Vehicles in a humid coastal city might have different maintenance challenges and wear-and-tear patterns compared to those in a dry desert environment. Moreover, driver behavior could vary significantly based on the climate. A city that frequently experiences heavy rainfall might have more cautious driving patterns than one which rarely sees a drop of rain.

Expanding the research to encompass cities with diverse climatic conditions can provide a more well-rounded understanding of the subject at hand. This multi-city approach not only adds depth to the findings but also tests the robustness of any models or hypotheses developed in the original study. It allows researchers to identify universal patterns and insights while also highlighting city-specific nuances.

Such an extension, while increasing the comprehensiveness of the research, also elevates its global relevance. Stakeholders from various cities can glean insights that are pertinent to their specific conditions, rather than trying to retrofit findings from a geographically distinct study. Whether it's urban planners, policymakers, or transportation authorities, the value of having data that resonates with their specific environment cannot be understated.

Furthermore, by examining cities across diverse climates, there's an opportunity to identify best practices and innovative solutions that emerge from the unique challenges each climate presents. A solution developed in a tropical city might be innovative and effective for a similar challenge faced by another city across the globe.

In conclusion, while geographically specific studies provide in-depth insights into a particular environment, the true potential of research is realized when it's broadened to encompass a spectrum of climatic conditions. This not only ensures that the findings are more holistic but also amplifies their global relevance, making the research a more powerful tool for change and improvement across diverse landscapes.

8.2.3 Adding new features and running the framework

In essence, while the core dataset used in this research offers foundational insights, the inclusion of supplementary features can greatly refine the output, providing a more comprehensive and accurate representation of the real world. For driving cycle research, both weather visibility and traffic information are paramount. These two features are not included in the current thesis. By accounting for these factors, researchers can ensure their findings are not only theoretically sound but also practically relevant, aiding in the development of informed solutions for contemporary transportation challenges.

8.2.4 Exploring the impact of COVID-19 on driving behavior

One significant area for future research is the examination of the impact of the COVID-19 pandemic on driving behavior. The dataset, which spans 2019, 2020, and 2022 — representing the years before, during, and after the pandemic — provides a valuable opportunity to analyze changes in driving patterns and behaviors throughout this period. The pandemic led to unprecedented shifts in traffic dynamics, influenced by lockdowns, social distancing measures, and changes in work patterns, such as the increase in remote working.

References

Abas, Mohd Azman, Srithar Rajoo, and Shaiful Fadzil Zainal Abidin. 2018. "Development of Malaysian Urban Drive Cycle Using Vehicle and Engine Parameters." *Transportation Research Part D: Transport and Environment* 63: 388–403.

Abdi Kordani, Ali, Omid Rahmani, Amir Saman Abdollahzadeh Nasiri, and Sid Mohammad Boroomandrad. 2018. "Effect of Adverse Weather Conditions on Vehicle Braking Distance of Highways." *Civil Engineering Journal* 4(1): 46.

Abele, Liva ;, Møller, and Mette. 2011. "General Rights The Relationship between Road Design and Driving Behavior." In *3rd International Conference on Road Safety and Simulation,*.

Abele, Liva, and Mette Møller. 2011. "The Relationship between Road Design and Driving Behavior." In *3rd International Conference on Road Safety and Simulation, ,* 17.

Achour, H., and A. G. Olabi. 2016. "Driving Cycle Developments and Their Impacts on Energy Consumption of Transportation." *Journal of Cleaner Production* 112: 1778–88.

Aditjandra, Paulus Teguh. 2013. "The Impact of Urban Development Patterns on Travel Behaviour: Lessons Learned from a British Metropolitan Region Using Macro-Analysis and Micro-Analysis in Addressing the Sustainability Agenda." *Research in Transportation Business and Management* 7: 69–80.

Agarwal, Ashish. 2004. "A Comparison of Weekend and Weekday Travel Behavior A Comparison of Weekend and Weekday Travel Behavior Characteristics in Urban Areas Characteristics in Urban Areas." University of South Florida. https://digitalcommons.usf.edu/etd.

Ahmed Al-Samari. 2017. "Real-World Driving Cycle: Case Study of Baqubah, Iraq." *Diyala Journal of Engineering Sciences* 10(2): 39–47. http://en.enginmag.uodiyala.edu.iq/pages?id=101.

Ahmed, Mohamed et al. 2018. *Driver Performance and Behavior in Adverse Weather Conditions: An Investigation Using the SHRP2 Naturalistic Driving Study Data-Phase 2*. https://rosap.ntl.bts.gov/view/dot/35664%0Ahttps://trid.trb.org/view/1540961.

Akin, Darcin, Virginia P. Sisiopiku, and Alexander Skabardonis. 2011. "Impacts of Weather on Traffic Flow Characteristics of Urban Freeways in Istanbul." In *Procedia - Social and*

Behavioral Sciences, Elsevier Ltd, 89–99.

Al-Samari, Ahmed. 2017. "Study of Emissions and Fuel Economy for Parallel Hybrid versus Conventional Vehicles on Real World and Standard Driving Cycles." *Alexandria Engineering Journal* 56(4): 721–26.

Ali, Elhashemi M., Mohamed M. Ahmed, and Guangchuan Yang. 2020. "Normal and Risky Driving Patterns Identification in Clear and Rainy Weather on Freeway Segments Using Vehicle Kinematics Trajectories and Time Series Cluster Analysis." *IATSS Research*.

Amirjamshidi, Glareh. 2015. "Assessment of Commercial Vehicle Emissions and Vehicle Routing of Fleets Using Simulated Driving Cycles." University of Toronto.

Amirjamshidi, Glareh, and Mattheh J. Roorda. 2013. "Development of Simulated Driving Cycles_ Case Study of the Toronto Waterfront Area _ Enhanced Reader." In *Transportation Research Board*, , 16.

An, Feng, Matthew Barth, George Scora, and Marc Ross. 1998. "Modeling Enleanment Emissions for Light-Duty Vehicles." *Transportation Research Record* (1641): 48–57.

Andersen, Camilla Sloth, and Kristian Hegner Reinau. 2016. "The Relationship Between Road Characteristics And Speed Collected From Floating Car Data." *Journal of Traffic and Transportation Engineering* 4(6): 10.

De Andrade, Guilherme Medeiros Soares, Fernando Wesley Cavalcanti De Araújo, Maurício Pereira Magalhães De Novaes Santos, and Fabio Santana Magnani. 2020. "Standardized Comparison of 40 Local Driving Cycles: Energy and Kinematics." *Energies* 13(20).

Andrey, Jean, Derrick Hambly, Brian Mills, and Sadia Afrin. 2013. "Insights into Driver Adaptation to Inclement Weather in Canada." *Journal of Transport Geography* 28: 192–203.

Anida, I. N. et al. 2017. "Characterisation and Development of Driving Cycle for Work Route in Kuala Terengganu." *International Journal of Automotive and Mechanical Engineering* 14(3): 4508–17.

Antonson, Hans, Selina Mårdh, Mats Wiklund, and Göran Blomqvist. 2009. "Effect of Surrounding Landscape on Driving Behaviour: A Driving Simulator Study." *Journal of Environmental Psychology* 29(4): 493–502.

Arbelaitz, Olatz et al. 2013. “An Extensive Comparative Study of Cluster Validity Indices.” *Pattern Recognition* 46(1): 243–56.

Aronsson, Karin F. M. 2006. *Speed Characteristics of Urban Streets Based on Driver Behaviour Studies and Simulation*. Division of transport och logistics, Royal Institute of Technology.

Arun, N. H., Srinath Mahesh, Gitakrishnan Ramadurai, and S. M. Shiva Nagendra. 2017. “Development of Driving Cycles for Passenger Cars and Motorcycles in Chennai, India.” *Sustainable Cities and Society* 32: 508–12.

Ashtari, Ali, Eric Bibeau, and Soheil Shahidinejad. 2014. “Using Large Driving Record Samples and a Stochastic Approach for Real-World Driving Cycle Construction: Winnipeg Driving Cycle.” *Transportation Science* 48(2): 170–83.

Azam, Md Shafiul, Ashish Bhaskar, and Md Mazharul Haque. 2022. “Driving Behaviour Modelling in the Context of Heterogeneous Traffic and Poor Lane Discipline Conditions: The State of the Art and Beyond.” *Transportmetrica A: Transport Science* 18(3): 367–434.

Babić, Dario, Mario Fiolić, Darko Babić, and Timothy Gates. 2020. “Road Markings and Their Impact on Driver Behaviour and Road Safety: A Systematic Review of Current Findings.” *Journal of Advanced Transportation* 2020.

Bakhshi, Vahid, Kayvan Aghabayk, Nasser Parishad, and Nirajan Shiwakoti. 2022. “Evaluating Rainy Weather Effects on Driving Behaviour Dimensions of Driving Behaviour Questionnaire.” *Journal of Advanced Transportation* 2022.

Barlow, T J, S Latham, I S Mccrae, and P G Boulter. 2009. *A Reference Book of Driving Cycles for Use in the Measurement of Road Vehicle Emissions*.

Ben-Bassat, Tamar, and David Shinar. 2011. “Effect of Shoulder Width, Guardrail and Roadway Geometry on Driver Perception and Behavior.” *Accident Analysis and Prevention* 43(6): 2142–52.

Bijleveld, Frits, and Tony Churchill. 2009. *The Influence of Weather Conditions on Road Safety*. Leidschendam.

Bodisco, Timothy, and Ali Zare. 2019. “Practicalities and Driving Dynamics of a Real Driving Emissions (RDE) Euro 6 Regulation Homologation Test.” *Energies* 12(12).

Broughton, Kathy L.M., Fred Switzer, and Don Scott. 2007. "Car Following Decisions under Three Visibility Conditions and Two Speeds Tested with a Driving Simulator." *Accident Analysis and Prevention* 39(1): 106–16.

Brundell-Freij, Karin, and Eva Ericsson. 2005. "Influence of Street Characteristics, Driver Category and Car Performance on Urban Driving Patterns." *Transportation Research Part D: Transport and Environment* 10(3): 213–29.

Canada, Government. 2021. *Provincial and Territorial Energy Profiles – Quebec*. Quebec. https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-quebec.html.

Carlson, Annelie, Ulf Hammarström, and Olle Eriksson. 2013. *Models and Methods for the Estimation of Fuel Consumption Due to Infrastructure Parameters*.

Chauhan, Boski P, Gaurang J Joshi, and Purnima Parida. 2020. "Development of Candidate Driving Cycles for an Urban Arterial Corridor of Vadodara City." *International Journal of Transport Economics, Engineering and Law* 4(75): 16.

Chen, Chen et al. 2019. "Influence of Adverse Weather on Drivers' Perceived Risk during Car Following Based on Driving Simulations." *Journal of Modern Transportation* 27(4): 282–92.

Chen, K. S. et al. 2003. "Motorcycle Emissions and Fuel Consumption in Urban and Rural Driving Conditions." *Science of the Total Environment* 312(1–3): 113–22.

Chen, Tao, Miao Zhang, and Lang Wei. 2014. "Driver Behavior on Combination of Vertical and Horizontal Curves of Mountainous Freeways." *Mathematical Problems in Engineering* 2014.

Chen, Xinqiang et al. 2017. "Evaluating the Impacts of Grades on Vehicular Speeds on Interstate Highways." *PLoS ONE* 12(9).

Chen, Xinqiang et al. 2020. "Sensing Data Supported Traffic Flow Prediction via Denoising Schemes and ANN: A Comparison." *IEEE Sensors Journal* 20(23): 14317–28.

Chiou, Yu Chiun, and Tai Chieh Chen. 2010. "Direct and Indirect Factors Affecting Emissions of Cars and Motorcycles in Taiwan." *Transportmetrica* 6(3): 215–33.

Chowdhury, Nazi Faisal A. 2015. "Ambient Temperature Effects on Driving." *Procedia Manufacturing* 3: 3123–27.

Christian, Thomas, and Thomas C Jensen. 2014. “General Rights Weather and Road Capacity Weather and Road Capacity.” *Danish Journal of Transportation Research* (1603–9696): 19. http://www.trafikdage.dk/artikelarkivwww.trafikdage.dk/artikelarkiv.

Chung, Edward, Hiroshi Warita, Osamu Ohtani, and Masao Kuwahara. 2006. “Does Weather Affect Highway Capacity?” In *Proceedings of the Symposium on Highway Capacity*, Tokyo, 9. https://www.researchgate.net/publication/287175635.

Cui, Mengying, and David Levinson. 2018. “Accessibility and the Ring of Unreliability.” *Transportmetrica A: Transport Science* 14(1–2): 4–21.

Cui, Yuepeng et al. 2022. “A Novel Optimization-Based Method to Develop Representative Driving Cycle in Various Driving Conditions.” *Energy* 247: 17.

Czégé, L, A Vámosi, and I Kocsis. 2020. “Review on Construction Procedures of Driving Cycles.” *International Journal of Engineering and Management Sciences (IJEMS)* 5(2).

Damsere-Derry, James, Francis K Afukaar, Peter Donkor, and Charles Mock. 2008. “Assessment of Vehicle Speeds on Different Categories of Roadways in Ghana.” *international journal of injury control and safety promotion* 15(2): 83–91. https://www.tandfonline.com/action/journalInformation?journalCode=nics20.

Degraeuwe, Bart, and Martin Weiss. 2017. “Does the New European Driving Cycle (NEDC) Really Fail to Capture the NOX Emissions of Diesel Cars in Europe?” *Environmental Pollution* 222: 234–41.

Dinh, Do Duy, and Hisashi Kubota. 2013. “Profile-Speed Data-Based Models to Estimate Operating Speeds for Urban Residential Streets with a 30km/h Speed Limit.” *IATSS Research* 36(2): 115–22.

Duran, Adam, and Matthew Earleywine. 2012. “GPS Data Filtration Method for Drive Cycle Analysis Applications.” In *SAE Technical Papers*, SAE International, 9.

Durkee, Shaun Michael. 2010. “The Effect Of Simulation Attributes On Driver Perception And Behavior.” Montana State University.

E.G.Tzirakis, and F.E. Zannikos. 2013. “On-Road Real Driving and Road Gradient Data Processing Methodologies to Form Driving Cycle Complete Dynamometer Test.” In

Proceedings of the 13th International Conference on Environmental Scince and Technology Athen, , 6.

Eckert, Jony Javorski et al. 2018. "Gear Shifting Multi-Objective Optimization to Improve Vehicle Performance, Fuel Consumption, and Engine Emissions." *Mechanics Based Design of Structures and Machines* 46(2): 238–53.

Elsagheer Mohamed, Samir A., Khaled A. Alshalfan, Mohammed A. Al-Hagery, and Mohamed Tahar Ben Othman. 2022. "Safe Driving Distance and Speed for Collision Avoidance in Connected Vehicles." *Sensors* 22(18).

EPA. 2014. *Evaporative Emissions from On-Road Vehicles in MOVES2014 (EPA-420-R-14-014, September 2014)*.

Ericsson, Eva. 2000. "Variability in Urban Driving Patterns." *Transportation Research Part D* 5: 18. www.elsevier.com/locate/trd.

Ericsson, Eva. 2001a. "Independent Driving Pattern Factors and Their In¯uence on Fuel-Use and Exhaust Emission Factors." *Transportation Research Part D* 6: 325–45. www.elsevier.com/locate/trd.

Ericsson, Eva. 2001b. "Independent Driving Pattern Factors and Their Influence on Fuel-Use and Exhaust Emission Factors." *Transportation Research Part D: Transport and Environment* 6(5): 325–45.

Ericsson, Eva, Hanna Larsson, and Karin Brundell-Freij. 2006. "Optimizing Route Choice for Lowest Fuel Consumption - Potential Effects of a New Driver Support Tool." *Transportation Research Part C: Emerging Technologies* 14(6): 369–83.

Esser, Arved, and Stephan Rinderknecht. 2020. "Process for the Validation of Using Synthetic Driving Cycles Based on Naturalistic Driving Data Sets." In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020*,.

Esteves-Booth, A et al. 2001. *The Measurement of Vehicular Driving Cycle within the City of Edinburgh*. www.elsevier.com/locate/trd.

European Commission Joint Research Center. 2013. *Regulated Emissions of a Euro 5 Passenger Car Measured over Different Driving Cycles*.

Faria, Marta V., Patricia C. Baptista, and Tiago L. Farias. 2017. “Main.” *transportation research procedia* 27: 953–60.

Faria, Marta V., Patrícia C. Baptista, Tiago L. Farias, and João M.S. Pereira. 2020. “Assessing the Impacts of Driving Environment on Driving Behavior Patterns.” *Transportation* 47(3): 1311–37.

Ferreira, Hugo, Carlos Manuel Rodrigues, and Carlos Pinho. 2019. “Impact of Road Geometry on Vehicle Energy Consumption and CO2 Emissions: An Energy-Efficiency Rating Methodology.” *Energies* 13(1).

FHWA. 2013. *Highway Functional Classification Concepts, Criteria and Procedures*.

Fotouhi, A., and M. Montazeri-Gh. 2012. “Tehran Driving Cycle Development Using the K-Means Clustering Method.” *Scientia Iranica* 20(2): 286–93.

Galgamuwa, Uditha, Loshaka Perera, and Saman Bandara. 2015. “Developing a General Methodology for Driving Cycle Construction: Comparison of Various Established Driving Cycles in the World to Propose a General Approach.” *Journal of Transportation Technologies* 05(04): 191–203.

Galgamuwa, Uditha, Loshaka Perera, and Saman Bandara. 2016. “A Representative Driving Cycle for the Southern Expressway Compared to Existing Driving Cycles.” *Transportation in Developing Economies* 2(2).

Gebisa, Amanuel, Girma Gebresenbet, Rajendiran Gopal, and Ramesh Babu Nallamothu. 2021. “Driving Cycles for Estimating Vehicle Emission Levels and Energy Consumption.” *Future Transportation* 1(3): 615–38.

Ghaemi, Mohammad Sajjad, Bruno Agard, Martin Trépanier, and Vahid Partovi Nia. 2017. “A Visual Segmentation Method for Temporal Smart Card Data.” *Transportmetrica A: Transport Science* 13(5): 381–404.

Giakoumis, Evangelos G. 2016. Driving and Engine Cycles *Driving and Engine Cycles*. Springer International Publishing.

Gonzalez, Luis C. et al. 2017. “Learning Roadway Surface Disruption Patterns Using the Bag of Words Representation.” *IEEE Transactions on Intelligent Transportation Systems* 18(11): . 2916.28

Goralzik, Anne, and Mark Vollrath. 2017. “The Effects of Road, Driver, and Passenger Presence on Drivers’ Choice of Speed: A Driving Simulator Study.” In *Transportation Research Procedia*, Elsevier B.V., 2061–75.

Gorges, Christian, Kemal Öztürk, and Robert Liebich. 2018. “Road Classification for Two-Wheeled Vehicles.” *Vehicle System Dynamics* 56(8): 1289–1314.

Greene, David L. et al. 2017. “What Is the Evidence Concerning the Gap between On-Road and Environmental Protection Agency Fuel Economy Ratings?” *Transport Policy* 53: 146–60.

Günther, R., T. Wenzel, M. Wegner, and R. Rettig. 2017. “Big Data Driven Dynamic Driving Cycle Development for Busses in Urban Public Transportation.” *Transportation Research Part D: Transport and Environment* 51: 276–89.

Hämäläinen, Joonas, Susanne Jauhiainen, and Tommi Kärkkäinen. 2017. “Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering.” *Algorithms* 10(3).

Han, Baorui et al. 2020. “Classification of Urban Street Networks Based on Tree-like Network Features.” *Sustainability (Switzerland)* 12(2).

Harris, Chioke B., and Michael E. Webber. 2012. “A Temporal Assessment of Vehicle Use Patterns and Their Impact on the Provision of Vehicle-to-Grid Services.” *Environmental Research Letters* 7(3): 11.

HCM. 2000. *Highway Capacity Manual*. Transportation Research Board, National Research Council.

He, Li, Bruno Agard, and Martin Trépanier. 2020. “A Classification of Public Transit Users with Smart Card Data Based on Time Series Distance Metrics and a Hierarchical Clustering Method.” *Transportmetrica A: Transport Science* 16(1): 56–75.

He, Yongjiang. 2020. *Research on the Construction Method of Vehicle Driving Cycle Based on Mean Shift Clustering*.

Heath, Joel P., and Peter Borowski. 2013. “Quantifying Proportional Variability.” *PLoS ONE* 8(12).

Hirschmann, Karin, Michael Zallinger, Martin Fellendorf, and Stefan Hausberger. 2010. “A New

Method to Calculate Emissions with Simulated Traffic Conditions." In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC,* , 33–38.

Hjelkrem, Odd André, and Eirin Olaussen Ryeng. 2017. "Driver Behaviour Data Linked with Vehicle, Weather, Road Surface, and Daylight Data." *Data in Brief* 10: 511–14.

Ho, Sze Hwee, Yiik Diew Wong, and Victor Wei Chung Chang. 2014. "Developing Singapore Driving Cycle for Passenger Cars to Estimate Fuel Consumption and Vehicular Emissions." *Atmospheric Environment* 97: 353–62.

Hoogendoorn, R. G., G. Tamminga, S. P. Hoogendoorn, and W. Daamen. 2010. "Longitudinal Driving Behavior under Adverse Weather Conditions: Adaptation Effects, Model Performance and Freeway Capacity in Case of Fog." In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC,* , 450–55.

Van Der Horst, Richard, and Selma De Ridder. 2007. "Influence of Roadside Infrastructure on Driving Behavior: Driving Simulator Study." *Transportation Research Record* (2018): 36–44.

Huertas, José I., Michael Giraldo, Luis F. Quirama, and Jenny Díaz. 2018. "Driving Cycles Based on Fuel Consumption." *Energies* 11(11).

Huertas, José Ignacio, Luis Felipe Quirama, Michael Giraldo, and Jenny Díaz. 2019. "Comparison of Three Methods for Constructing Real Driving Cycles." *Energies* 12(4).

Hung, W. T. et al. 2007. "Development of a Practical Driving Cycle Construction Methodology: A Case Study in Hong Kong." *Transportation Research Part D: Transport and Environment* 12(2): 115–28.

Hung, Wing Tat et al. 2005. "Comparison of Driving Characteristics in Cities of Pearl River Delta, China." *Atmospheric Environment* 39(4): 615–25.

IEA. 2019. *Tracking Transport 2019*. Paris. https://www.iea.org/reports/tracking-transport-2019.

IPCC. 2022. 2022 2022: Transport. In IPCC *The Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Climate Change.

Jägerbrand, Annika K., and Jonas Sjöbergh. 2016. "Effects of Weather Conditions, Light Conditions, and Road Lighting on Vehicle Speed." *SpringerPlus* 5(1).

Jiang, Ying et al. 2020. “Effects of Weather on Highway Traffic Capacity in China: Characteristics and Causes of Roadblocks Due to Fog Events.” *Pure and Applied Geophysics* 177(10): 5027–40.

Jun, Gwook, Randall Guensler, and Jennifer H. Ogle. 2006. “Smoothing Methods to Minimize Impact of Global Positioning System Random Error on Travel Distance, Speed, and Acceleration Profile Estimates.” *Transportation Research Record* (1972): 141–50.

Kamble, Sanghpriya H., Tom V. Mathew, and G. K. Sharma. 2009. “Development of Real-World Driving Cycle: Case Study of Pune, India.” *Transportation Research Part D: Transport and Environment* 14(2): 132–40.

Kaymaz, Habib, Hayriye Korkmaz, and Hasan Erdal. 2019. “Development of a Driving Cycle for Istanbul Bus Rapid Transit Based on Real-World Data Using Stratified Sampling Method.” *Transportation Research Part D: Transport and Environment* 75: 123–35.

Kazemi Miyangaskari, Mina et al. 2023. “Designing a Closed-Loop Supply Chain Multi-Objective Optimization Model with the Aim of Supplier Selection and Order Allocation (Case Study: Iran’s Retail Stores for Protein Products).” *Industrial Management Studies* 21(69). https://doi.org/10.22054/jims.2023.70278.2815.

Khanfar, Nour O. et al. 2023. “Driving Behavior Classification at Signalized Intersections Using Vehicle Kinematics: Application of Unsupervised Machine Learning.” *International Journal of Injury Control and Safety Promotion* 30(1): 34–44.

Khattak, Asad J., and Behram Wali. 2017. “Analysis of Volatility in Driving Regimes Extracted from Basic Safety Messages Transmitted between Connected Vehicles.” *Transportation Research Part C: Emerging Technologies* 84: 48–73.

Kilpeläinen, Markku, and Heikki Summala. 2007. “Effects of Weather and Weather Forecasts on Driver Behaviour.” *Transportation Research Part F: Traffic Psychology and Behaviour* 10(4): 288–99.

Kim, Rae Seon. 2011. “Standardized Regression Coefficients as Indices of Effect Sizes in Meta-Analysis.” The Florida State University.

Kliment Ohridski" -Bitola, St. 2009. “Role of Functional Classification of Highways in Road Traffic Safety.” *Transportation problem* 4(3): 1.

https://www.researchgate.net/publication/38104647.

Knez, Matjaz, Tariq Muneer, Borut Jereb, and Kevin Cullinane. 2014. "The Estimation of a Driving Cycle for Celje and a Comparison to Other European Cities." *Sustainable Cities and Society* 11: 56–60.

Ladubec, Chris et al. 2021. *Assessment of Energy Efficiency Ratios of Electric and Fuel Cell Light Duty Vehicles in Canada: Final Report*.

Lee, Chih Feng, and Per Öberg. 2015. "Classification of Road Type and Driving Style Using OBD Data." In *SAE Technical Papers,* SAE International.

Lee, Tae-Kyung, and Zoran S. Filipi. 2011. "Synthesis and Validation of Representative Real-World Driving Cycles for Plug-In Hybrid." *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*.

Lewandowski, Daniel, and Roger M Cooke. 2007. "Sample-Based Estimation of Correlation Ratio with Polynomial Approximation." *ACM Transactions on Modeling and Computer Simulation* v: 1–16.

Lin, Cheng, Li Zhao, Xingqun Cheng, and Wenwei Wang. 2015. "A DCT-Based Driving Cycle Generation Method and Its Application for Electric Vehicles." *Mathematical Problems in Engineering* 2015: 13.

Lin, Jane, Yi Chang Chiu, Suriya Vallamsundar, and Song Bai. 2011. "Integration of MOVES and Dynamic Traffic Assignment Models for Fine-Grained Transportation and Air Quality Analyses." In *2011 IEEE Forum on Integrated and Sustainable Transportation Systems, FISTS 2011*, , 176–81.

Lin, Jie, and Debbie A Niemeier. 2002. 36 Atmospheric Environment *An Exploratory Analysis Comparing a Stochastic Driving Cycle to California's Regulatory Cycle*.

Liu, Chengxi, Yusak O. Susilo, and Anders Karlström. 2016. "Estimating Changes in Transport CO2 Emissions Due to Changes in Weather and Climate in Sweden." *Transportation Research Part D: Transport and Environment* 49: 172–87.

Liu, Haobing, Michael O. Rodgers, Fang "Cherry" Liu, and Randall Guensler. 2020. "Bayesian Approach in Estimating the Road Grade Impact on Vehicle Speed and Acceleration on

Freeways." *Transportmetrica A: Transport Science* 16(3): 602–25.

Liu, Jing, Jing Cai, Shanshan Lin, and Jianyou Zhao. 2020. "Analysis of Factors Affecting a Driver's Driving Speed Selection in Low Illumination." *Journal of Advanced Transportation* 2020.

Liu, Jun, Asad Khattak, Lee Han, and Quan Yuan. 2020. "How Much Information Is Lost When Sampling Driving Behavior Data? Indicators to Quantify the Extent of Information Loss." *Journal of Intelligent and Connected Vehicles* 3(1): 17–29.

Liu, Xiaodong et al. 2020. "Study on Driving Cycle Synthesis Method for City Buses Considering Random Passenger Load." *Journal of Advanced Transportation* 2020.

Liu, Yanchi et al. 2010. "Understanding of Internal Clustering Validation Measures." In *Proceedings - IEEE International Conference on Data Mining, ICDM*, , 911–16.

Ma, Ruoyun et al. 2019. "Real-World Driving Cycles and Energy Consumption Informed by Large-Sized Vehicle Trajectory Data." *Journal of Cleaner Production* 223: 564–74.

Mafi, Shahab et al. 2022. "Developing Local Driving Cycle for Accurate Vehicular CO2 Monitoring: A Case Study of Tehran." *Journal of Cleaner Production* 336.

Mahayadin, A. R. et al. 2017. "Efficient Methodology of Route Selection for Driving Cycle Development." In *Journal of Physics: Conference Series*, Institute of Physics Publishing.

Mayakuntla, Sai Kiran, and Ashish Verma. 2018. "A Novel Methodology for Construction of Driving Cycles for Indian Cities." *Transportation Research Part D: Transport and Environment* 65: 725–35.

Mi, Jiayu, Jiuchun Jiang, Shaoyuan Wei, and Xinyu Jia Xinyu Jia. 2019. "Synthesis of Representative Driving Cycles for Energy Storage Based Urban Rail Vehicles." *DEStech Transactions on Environment, Energy and Earth Sciences* (iceee).

Min, Jing. 2015. *Quantifying the Effects of Winter Weather and Road Maintenance on Emissions and Fuel Consumptions*.

Morris, Drew. 2015. "The Cold Driver: Driving Performance under Thermal Stress." Clemson. https://tigerprints.clemson.edu/all_theses (January 8, 2021).

Muhammad, Jahanzaib. 2022. "Development of Driving Cycle for Passenger Cars and Estimation

of Vehicular Exhaust Emission Factors." *Journal Of Environmental Pollution And Control* 5(1): 1–11.

Nama, Suresh et al. 2016. "Vehicle Speed Characteristics and Alignment Design Consistency for Mountainous Roads." *Transportation in Developing Economies* 2(2).

Nesamani, K.S., and K.P. Subramanian. 2011. "Development of a Driving Cycle for Intra-City Buses in Chennai, India." *Atmospheric Environment* 33(15): 2323–35.

Nguyen, Yen-Lien T, Ngoc Dung Bui, Trung-Dung Nghiem, and Anh-Tuan Le. 2020. "GPS Data Processing For Driving Cycle Development In Hanoi, Vietnam." *Journal of Engineering Science and Technology* 15(2): 1429–40.

Nguyen, Yen Lien T., Trung Dung Nghiem, Anh Tuan Le, and Ngoc Dung Bui. 2019. "Development of the Typical Driving Cycle for Buses in Hanoi, Vietnam." *Journal of the Air and Waste Management Association* 69(4): 423–37.

Niemeier, Deb, and Douglas Eisinger. 2008. *Driving Cycles: A New Cycle-Building Method That Better Represents Real-World Emissions*. https://www.researchgate.net/publication/265495453.

Nouri, Pegah. 2015. "Enhancing the Gasoline Vehicles' CO2 Emissions Estimation in Montreal." Polytechnique Montreal.

Nouri, Pegah, and Catherine Morency. 2015. *Untangling the Impacts of Various Factors on Emission Levels of Light Duty Gasoline Vehicles*.

Nouri, Pegah, and Catherine Morency. 2015. 2017. "Evaluating Microtrip Definitions for Developing Driving Cycles." *Transportation Research Record* 2627: 86–92.

Nutramon, Tamsanya, and Chungpaibulpatana Supachart. 2009. "Influence of Driving Cycles on Exhaust Emissions and Fuel Consumption of Gasoline Passenger Car in Bangkok." *Journal of Environmental Sciences* 21(5): 604–11.

Nyberg, Peter. 2015. "Evaluation, Generation, and Transformation of Driving Cycles." Linköping University institue and technology.

Pilgerstorfer, Authors. 2011. "Small Scale Naturalistic Driving Pilot Studies • To Test and Refine the Practical and Technical." In *DaCoTA Conference 22-23*, , 18.

Pouresmaeili, Mohammad Amin, Iman Aghayan, and Seyed Ali Taghizadeh. 2018. "Development of Mashhad Driving Cycle for Passenger Car to Model Vehicle Exhaust Emissions Calibrated Using On-Board Measurements." *Sustainable Cities and Society* 36: 12–20.

Quirama, Luis F., Michael Giraldo, José I. Huertas, and Miguel Jaller. 2020. "Driving Cycles That Reproduce Driving Patterns, Energy Consumptions and Tailpipe Emissions." *Transportation Research Part D: Transport and Environment* 82.

Ragione, Livia Della, and Meccariello Giovanni. 2016. "The Influence of Road Gradient in an Integrated Approach of Real Driving Cycles and Emissions Factors Model." In *Transportation Research Procedia*, Elsevier B.V., 3179–88.

Ranacher, Peter et al. 2016. "Why GPS Makes Distances Bigger than They Are." *International Journal of Geographical Information Science* 30(2): 316–33.

Ratner, Bruce. 2009. "The Correlation Coefficient: Its Values Range between 1/1, or Do They." *Journal of Targeting, Measurement and Analysis for Marketing* 17(2): 139–42.

Rechkemmer, Sabrina K. et al. 2021. "Utilization of Smartphone Data for Driving Cycle Synthesis Based on Electric Two-Wheelers in Shanghai." *IEEE Transactions on Intelligent Transportation Systems* 22(2): 876–86.

Ribeiro, Silvana Mara. 2021. *Imputation by Decomposition and by Time Series Pattern: Novel Imputation Methods for Missing Data in Time Series*.

Rios-Torres, Jackeline, Jun Liu, and Asad Khattak. 2019. "Fuel Consumption for Various Driving Styles in Conventional and Hybrid Electric Vehicles: Integrating Driving Cycle Predictions with Fuel Consumption Optimization*." *International Journal of Sustainable Transportation* 13(2): 123–37.

Ritner, Mark, Kurt K. Westerlund, C. David Cooper, and Michael Claggett. 2013. "Accounting for Acceleration and Deceleration Emissions in Intersection Dispersion Modeling Using MOVES and CAL3QHC." *Journal of the Air and Waste Management Association* 63(6): 724–36.

Roso, Ruckert et al. 2015. "Evaluation Of A Real-World Driving Cycle And Its Impacts On Fuel Consumption And Emissions."

Roy, Frédérique, and Catherine Morency. 2020. "Comparing Driving Cycle Development Methods

Based on Markov Chains." *Transportation Research Record: Journal of the Transportation Research Board*: 036119812096882.

Sagberg, F. 2003. Institute of Transport Economics *Influencing Driver Behaviour and Safety by Road System Improvements*. https://trid.trb.org/view/697621.

Saleh, Wafaa, Ravindra Kumar, and Añil Sharma. 2010. "Driving Cycle for Motorcycles in Modern Cities: Case Studies of Edinburgh and Delhi." *World Journal of Science, Technology and Sustainable Development* 7(3): 263–74.

Salihu, Flamur, Yusuf Kağan Demir, and Hatice Göçmen Demir. 2023. "Effect of Road Slope on Driving Cycle Parameters of Urban Roads." *Transportation Research Part D: Transport and Environment* 118: 103676.

Samuel, S., L. Austin, and D. Morrey. 2002. "Automotive Test Drive Cycles for Emission Measurement and Real-World Emission Levels - A Review." *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 216(7): 555–64.

Sankar, Subramaniam Saravana, Yiqun Xia, Julaluk Carmai, and Saiprasit Koetniyom. 2020. "Optimal Eco-Driving Cycles for Conventional Vehicles Using a Genetic Algorithm." *Energies* 13(17).

Schober, Patrick, and Lothar A. Schwarte. 2018. "Correlation Coefficients: Appropriate Use and Interpretation." *Anesthesia and Analgesia* 126(5): 1763–68.

Seers, P., G. Nachin, and M. Glaus. 2015. "Development of Two Driving Cycles for Utility Vehicles." *Transportation Research Part D: Transport and Environment* 41: 377–85.

Shi, Shuming et al. 2016. "Research on Markov Property Analysis of Driving Cycles and Its Application." *Transportation Research Part D: Transport and Environment* 47: 171–81.

Silvano, Ary P., and Karl L. Bang. 2016. "Impact of Speed Limits and Road Characteristics on Free-Flow Speed in Urban Areas." *Journal of Transportation Engineering* 142(2).

Silvas, Emilia et al. 2016. "Synthesis of Realistic Driving Cycles with High Accuracy and Computational Speed, Including Slope Information." *IEEE Transactions on Vehicular Technology* 65(6): 4118–28.

Sithananthan, Masilamani, and Ravindra Kumar. 2020. "A Framework for Development of Real-

World Motorcycle Driving Cycle in India." *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*: 095440702097753. http://journals.sagepub.com/doi/10.1177/0954407020977533.

Skanda R. Kalgal et al. 2017. "Segmented Studies on Urban Driving Cycle and Traffic Patterns." In *2017 International Conference on Microelectronic Devices, Circuits and Systems (ICMDCS)*, , 6.

Stevanovic-Briatico, Vesna. 2013. *City of Toronto Road Classification System*. Toronto.

Supiyono, Ludfi Djakfar, and Achmad Wicaksono. 2022. "Drivers Behavior at Signalized Intersections with Countdown Timer Reduces Air Pollution by Gender." In *IOP Conference Series: Earth and Environmental Science*, Institute of Physics.

Svenson, Gunnar, and Dag Fjeld. 2016. "The Impact of Road Geometry and Surface Roughness on Fuel Consumption of Logging Trucks." *Scandinavian Journal of Forest Research* 31(5): 526–36.

Tayarani, Mohammad, Razieh Nadafianshahamabadi, Amir Poorfakhraei, and Gregory Rowangould. 2018. "Evaluating the Cumulative Impacts of a Long Range Regional Transportation Plan: Particulate Matter Exposure, Greenhouse Gas Emissions, and Transportation System Performance." *Transportation Research Part D: Transport and Environment* 63: 261–75.

Tharvin, R. et al. 2018. "Development of Driving Cycle for Passenger Car under Real World Driving Conditions in Kuala Lumpur, Malaysia." In *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing.

Tong, H. Y., and W. T. Hung. 2010. "A Framework for Developing Driving Cycles with On-Road Driving Data." *Transport Reviews* 30(5): 589–615.

Tong, Hing Yan, and Ka Wai Ng. 2021. "Development of Bus Driving Cycles Using a Cost Effective Data Collection Approach." *Sustainable Cities and Society* 69.

Tzirakis, Evangelos G, and Fanourios E Zannikos. 2013. *On-Road Real Driving and Road Gradient Data Processing Methodologies to Form Driving Cycle Complete Dynamometer*. https://www.researchgate.net/publication/314260289.

United Nations. 2014. *Addendum 15: Global Technical Regulation No. 15 Worldwide Harmonized Light Vehicles Test Procedure Global Technical Regulation on Worldwide Harmonized Light Vehicles Test Procedure Contents*.

Valera, J. J. et al. 2014. "Driving Cycle and Road Grade On-Board Predictions for the Optimal Energy Management in EV-PHEVs." In *2013 World Electric Vehicle Symposium and Exhibition, EVS 2014*,.

Varella, Roberto A. et al. 2019. "Assessing the Influence of Boundary Conditions, Driving Behavior and Data Analysis Methods on Real Driving CO2 and NOx Emissions." *Science of the Total Environment* 658: 879–94.

Viallard, Alexis, Martin Trépanier, and Catherine Morency. 2019. "Assessing the Evolution of Transit User Behavior from Smart Card Data." *Transportation Research Record* 2673(4): 184–94.

Vitkienė, Jūratė, Virgaudas Puodžiukas, and Daiva Žilionienė. 2017. "New Approach to the Lithuanian Road Classification Based on Worldwide Experience." In *10th International Conference on Environmental Engineering, ICEE 2017*, Vilnius Gediminas Technical University Publishing House "Technika."

Wang, Qidong et al. 2008. "Characterization of Vehicle Driving Patterns and Development of Driving Cycles in Chinese Cities." *Transportation Research Part D: Transport and Environment* 13(5): 289–97.

Wang, Zhenpo et al. 2019. "Driving Cycle Construction for Electric Vehicles Based on Markov Chain and Monte Carlo Method: A Case Study in Beijing." In *Energy Procedia*, Elsevier Ltd, 2494–99.

Wang, Zhenpo, Yang Li, Hao Luo, and Xue Xue. 2016. *An Improved Method of Vehicle Driving Cycle Construction: A Case Study of Beijing*. Beijing.

Wolfermann, Axel, Wael K.M. Alhajyaseen, and Hideki Nakamura. 2011. "Modeling Speed Profiles of Turning Vehicles at Signalized Intersections." *2004_Veolocity_Planning* (September 2016): 1–17.

Wood, Alice Louise. 2015. "A Climatology of North American Air Masses and Their Extremes in Montreal, Quebec." McGill University.

Wu, Yue et al. 2023. “Spatial–Temporal Data-Driven Full Driving Cycle Prediction for Optimal Energy Management of Battery/Supercapacitor Electric Vehicles.” *Energy Conversion and Management* 277.

Xu, Shuo et al. 2016. “Reviews on Determining the Number of Clusters.” *Applied Mathematics and Information Sciences* 10(4): 1493–1512.

Yan, Xuedong, Xiaomeng Li, Yang Liu, and Jia Zhao. 2014. “Effects of Foggy Conditions on Drivers’ Speed Control Behaviors at Different Risk Levels.” *Safety Science* 68: 275–87.

Yang, Jie et al. 2019. “Modeling of the Relationship between Speed Limit and Characteristic Speed of Expressway Traffic Flow.” *Sustainability (Switzerland)* 11(17).

Yang, Yang, Tiezhu Li, Tao Zhang, and Qian Yu. 2020. “Time Dimension Analysis: Comparison of Nanjing Local Driving Cycles in 2009 and 2017.” *Sustainable Cities and Society* 53.

Yang, Ying et al. 2018. “Markov Chain-Based Approach of the Driving Cycle Development for Electric Vehicle Application.” In *Energy Procedia*, Elsevier Ltd, 502–7.

Yang, Zifei, Rachel Muncrief, and Anup Bandivadekar. 2017. *Global Baseline Assessment of Compliance and Enforcement Programs for Vehicle Emissions and Energy Efficiency*. www.theicct.org.

Yarahmadi, Asad, Catherine Morency, and Martin Trepanier. 2023a. “Identifying Optimal Number of Driving Cycles in Different Driving Conditions Using Machine Learning.” In *World Conference on Transport Research - WCTR 2023*, www.sciencedirect.comwww.elsevier.com/locate/procedia2352-1465.

Yarahmadi, Asad, Catherine Morency, and Martin Trepanier. 2023b. “New Data-Driven Approach to Generate Typologies of Road Segments.” *Transportmetrica A: Transport Science*: 1–24.

Yeo, Jiho, Jooyoung Lee, and Kitae Jang. 2021. “The Effects of Rainfall on Driving Behaviors Based on Driving Volatility.” *International Journal of Sustainable Transportation* 15(6): 435–43.

Yu, Lei, Xiao Zhang, Fengxiang Qiao, and Yi Qi. 2010. “Genetic Algorithm-Based Approach to Develop Driving Schedules to Evaluate Greenhouse Gas Emissions from Light-Duty Vehicles.” *Transportation Research Record* (2191): 166–73.

Yue, Bingjian et al. 2015. “Study on the Design Method of Driving Cycle with Road Grade Based on Markov Chain Model.” In *2015 IEEE Vehicle Power and Propulsion Conference, VPPC 2015 - Proceedings*, Institute of Electrical and Electronics Engineers Inc.

Yue, Lei, and Hui Wang. 2019. “An Optimization Design Method of Combination of Steep Slope and Sharp Curve Sections for Mountain Highways.” *Mathematical Problems in Engineering* 2019.

Yuhui, Peng, Zhuang Yuan, and Yang Huibao. 2018. “Development of a Representative Driving Cycle for Urban Buses Based on the K-Means Cluster Method.” *Cluster Computing* 22: 6871–80.

Zähringer, Maximilian, Svenja Kalt, and Markus Lienkamp. 2020. “Compressed Driving Cycles Using Markov Chains for Vehicle Powertrain Design.” *World Electric Vehicle Journal* 11(3).

Zhai, Yikai et al. 2018. “A Context-Aware Evaluation Method of Driving Behavior.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 462–74.

Zhang, Jianbo et al. 2018. “Analysis of Rainfall Effects on Road Travel Speed in Beijing, China.” *IET Intelligent Transport Systems* 12(2): 93–102.

Zhang, Lihang et al. 2021. “Road Type-Based Driving Cycle Development and Application to Estimate Vehicle Emissions for Passenger Cars in Guangzhou.” *Atmospheric Pollution Research* 12(8).

Zhang, Man, Shuming Shi, Nan Lin, and Bingjian Yue. 2019. “High-Efficiency Driving Cycle Generation Using a Markov Chain Evolution Algorithm.” *IEEE Transactions on Vehicular Technology* 68(2): 1288–1301.

Zhang, Xiao, Dui Jia Zhao, and Jun Min Shen. 2011. “A Synthesis of Methodologies and Practices for Developing Driving Cycles.” In *Energy Procedia*, Elsevier BV, 1868–73.

Zhao, Xuan et al. 2018. “Development of a Representative EV Urban Driving Cycle Based on a K-Means and SVM Hybrid Clustering Algorithm.” *Journal of Advanced Transportation* 2018.

Zhao, Yunjie, Adel Sadek, and Daniel Fuglewicz. 2012. “Modeling the Impact of Inclement

Weather on Freeway Traffic Speed at Macroscopic and Microscopic Levels." *Transportation Research Record* (2272): 173–80.

Zhong, Ming, John Douglas Hunt, and Xuewen Lu. 2008. "Studying Differences of Household Weekday and Weekend Activities A Duration Perspective." *Transportation Research Record* (2054): 28–36.

Zito, R, and F Primerano. 2005. *Drive Cycle Development Methodology and Results*. Adelaide.

Appendix

1.Technical description of driving cycle generation

In constructing a driving cycle, it is crucial to have detailed driving data. As outlined in Section 6.3.1, we utilized approximately 10,567,000 taxi trips collected over the course of two years, from 2019 to 2020. This data encompasses different times of the day, including both rush and non-rush hours, ensuring a comprehensive representation of driving patterns. The key characteristics of this data are detailed in Table 6.2. The raw data were initially recorded at an average interval of every five seconds.

Further, as mentioned in Section 5.3.4, we applied a range of preprocessing techniques to the raw data to make it suitable for analysis. A significant step in this preprocessing was the modification of the timestamp frequency, changing it from every five seconds to every second. This increase in data granularity allows for a more detailed and accurate analysis of driving behaviors and patterns. To do that, first we used the *resample* function in Pandas library. Then we utilized KalmanFilter function in Pykalman library for predicting missing longitude and latitude values. The literature indicates that the Kalman filter is an effective method for filling missing values in time series data (Ribeiro 2021). A common method for validating the results of a Kalman filter is through visualization. To do that, QGIS is employed to conduct a comparative analysis. This involved loading both the dataset before and after the imputation, which was resampled to a one-second interval, into the QGIS platform. By overlaying these datasets within QGIS, the precision of the imputations can be visually evaluated.

The initial parameters of Kalman filter are as follows:

```
F = ([[1, 0, 0, 1, 0, 0],
      [0, 1, 0, 0, 1, 0],
      [0, 0, 1, 0, 0, 1],
      [0, 0, 0, 1, 0, 0],
      [0, 0, 0, 0, 1, 0],
      [0, 0, 0, 0, 0, 1]])
```

```
H = ([[1, 0, 0, 0, 0, 0],
      [0, 1, 0, 0, 0, 0],
       [0, 0, 1, 0, 0, 0]])
R = ([1e-4, 1e-4, 100])**2
Q = ([[  3.17720723e-09,  -1.56389148e-09,  -2.41793770e-07,
            2.29258935e-09,  -3.17260647e-09,  -2.89201471e-07],
         [  1.56687815e-09,   3.16555076e-09,   1.19734906e-07,
            3.17314157e-09,   2.27469595e-09,  -2.11189940e-08],
         [ -5.13624053e-08,   2.60171362e-07,   4.62632068e-01,
            1.00082746e-07,   2.81568920e-07,   6.99461902e-05],
         [  2.98805710e-09,  -8.62315114e-10,  -1.90678253e-07,
            5.58468140e-09,  -5.46272629e-09,  -5.75557899e-07],
         [  8.66285671e-10,   2.97046913e-09,   1.54584155e-07,
            5.46269262e-09,   5.55161528e-09,   5.67122163e-08],
         [ -9.24540217e-08,   2.09822077e-07,   7.65126136e-05,
            4.58344911e-08,   5.74790902e-07,   3.89895992e-04]])
```

To produce a driving cycle, driving data should be generated per second. It is because driving cycle parameters are calculated at a one-second interval.

To generate driving cycle, we used a python code developed by Roy and Morency (2020). The code has various classes including microtrip generation, clustering, driving cycle construction. To construct a driving cycle using this code each step should be initialized and performed manually.

The first step is preparing raw data, explained in Table 5-3, through steps explained in section 5.3.4. Table A.1 shows the result of prepared data.

Table A.1 prepared driving data

DateTime	Speed	Distance	gps_Lat	gps_Long
2020-01-02 19:47:55	0	0	45.50199608	-73.66405841
2020-01-02 19:47:56	12.19704351	3.388067642	45.50197022	-73.66401416
2020-01-02 19:47:57	18.05155883	8.40238954	45.50188089	-73.66392333

The second step is microtrip generation. As explained in section 5.3.5, the length of microtrip we selected was 250 m. This length is selected because (Nouri 2015; Roy and Morency 2020) showed it generates more representative result. In this study, the driving mode characteristics including idling, cruising, creeping, acceleration, and deceleration are calculated according to Table 6-3. Figure A.1 and Figure A.2 indicate the microtrips speed profile in different road types, weather conditions, and weekdays and microtrips characteristics respectively.

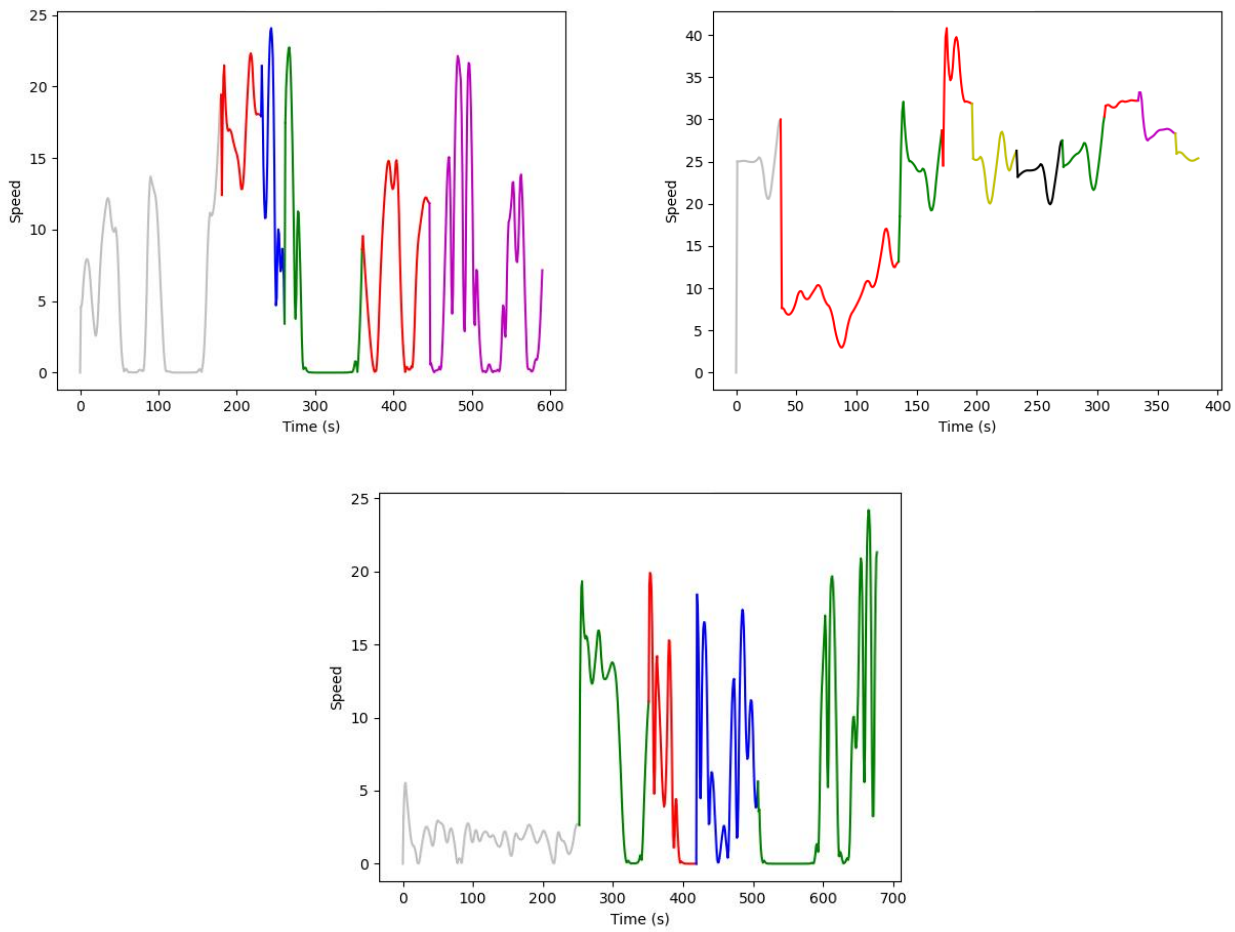


Figure A.1 Speed profile with microtrip segmentation in different road types, weather conditions and types of day.

Seg	DateTime	gps_Lat	gps_Long	V	V_m	V_r	V_std	Acc	Dcc	Acc_std	Acc_2	Idle_p	Acc_p	Cru_p	Cre_p	Dcc_p
0	2020-08-07 19:44:59	45.4954	-73.7063	8.4	16.2	8.4	6.2	0.1	-0.2	0.2	0.0	0.3	0.3	0.0	0.6	0.1
1	2020-08-07 19:46:46	45.5014	-73.6988	4.2	15.7	4.2	5.7	0.1	-0.1	0.2	0.0	0.6	0.1	0.0	0.7	0.2
2	2020-08-07 19:49:34	45.6363	-73.5949	0.1	1.2	0.1	0.1	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0

Figure A. 2 Microtrip segmentation characteristics

Next, generated microtrips are clustered using k-means algorithm. To do this, first the PCA is performed on the features and components accumulative contribution rate is 80% are kept. Then, k-means should be initialized by cluster numbers. For different datasets, the best number of clusters changes from 7 to 10 clusters, but Roy and Morency (2020) suggested to select a fixed number of clusters to be able to isolate the impact of microtrip definitions and clustering methods on the driving cycle performance. Therefore, 8 clusters are considered for different conditions. Figure A.3 displays the clustering results in different conditions.

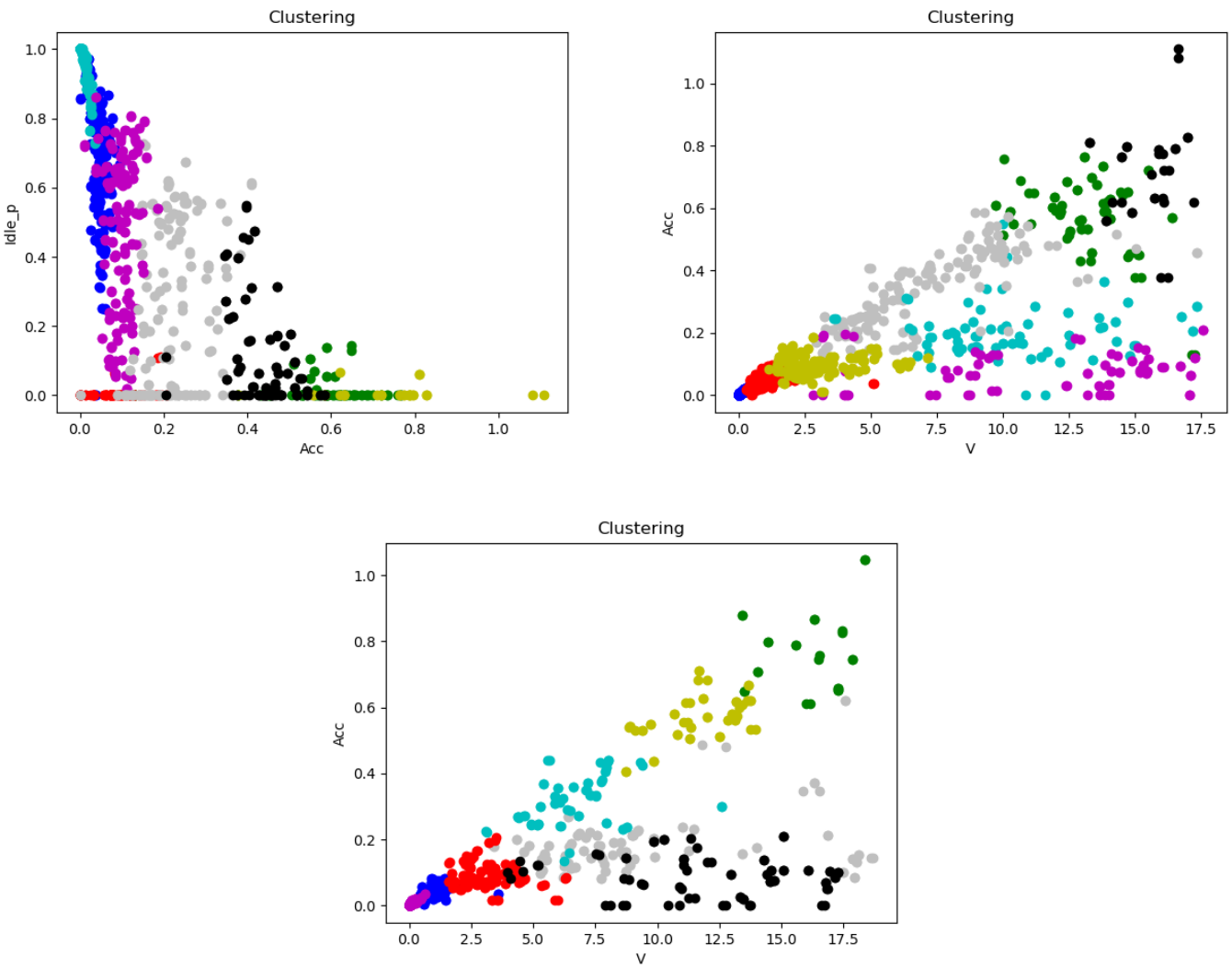


Figure A. 3 Clustering results in different conditions

Then, to construct driving cycle the code uses a Markov-chains approach. This process starts chaining microtrips by initializing with the starting microtrip, and the subsequent microtrip is chosen using a one-step Markov model. This method produces a transition matrix that represents the likelihood of transitioning from one cluster of microtrips to another within the dataset. The probabilities are derived from the sequences of microtrips throughout the database. The selection of each new microtrip is statistically determined by the calculated probabilities. Once a microtrip is chosen, the next one is identified based on these transition probabilities. To perform the Markov model, three parameters must be initialized: delta speed, length of the driving cycle, and the number of iterations. The delta speed parameter determines the maximum acceptable speed difference between the end of one cluster and the beginning of the next one. This parameter ensures continuity

in the driving cycle, with the recommended value by the developer being 2 km/h. The next parameter, the number of iterations, establishes how many candidate driving cycles should be constructed to identify the most representative one. The importance of iterations is tied to their impact on the reliability and reproducibility of the driving cycles produced. The selected number of iterations is designed to ensure that the resulting cycles accurately reflect real-world driving behavior. While the default value is 800, it has been increased to 1200 due to the number of trips in the dataset. The final parameter, the length of the driving cycle, depends on the average length of trips. In this study, the trip lengths ranged from 18 to 25 minutes, observed between December 2020 and September 2019, respectively. Ultimately, the most representative driving cycle is selected according to the methodology described in Nouri (2015).

2.Calculating density of signalized intersections

To calculate the density of intersections, we adopted a zoning approach. In this method, we imported a layer of signalized intersections into QGIS and overlaid a grid on it. The primary challenge in this approach was determining the optimal grid cell size. We tested three different cell sizes: 250 m, 500 m, and 750 m. To select the optimal size, we considered two criteria: the homogeneity of the distribution of points within each cell and the number of road segments located in more than one cell. With the smallest cell size, a considerable number of road links were observed to span more than one cell (sometimes even more than two cells). Furthermore, in dense areas like downtown, the number of traffic lights was excessively high, resulting in many cells empty of any lights. The issue of road segments overlapping multiple cells was resolved when the cell size was increased to 750 m, although this led to an inhomogeneous distribution of points. Specifically, one part of the cell might contain several points, while another side might have segments with no signalized intersections. A 500 m cell size provided a more homogeneous point distribution than the 750 m size, and the number of segments overlapping multiple cells decreased significantly. To address the problem of segments straddling two cells in QGIS, we split the segments at the cell boundaries where necessary. After determining optimal size, we generated a density layer and then density of each cell assigned to road segment. Figure A. 4 illustrates the density of intersection in Montreal.

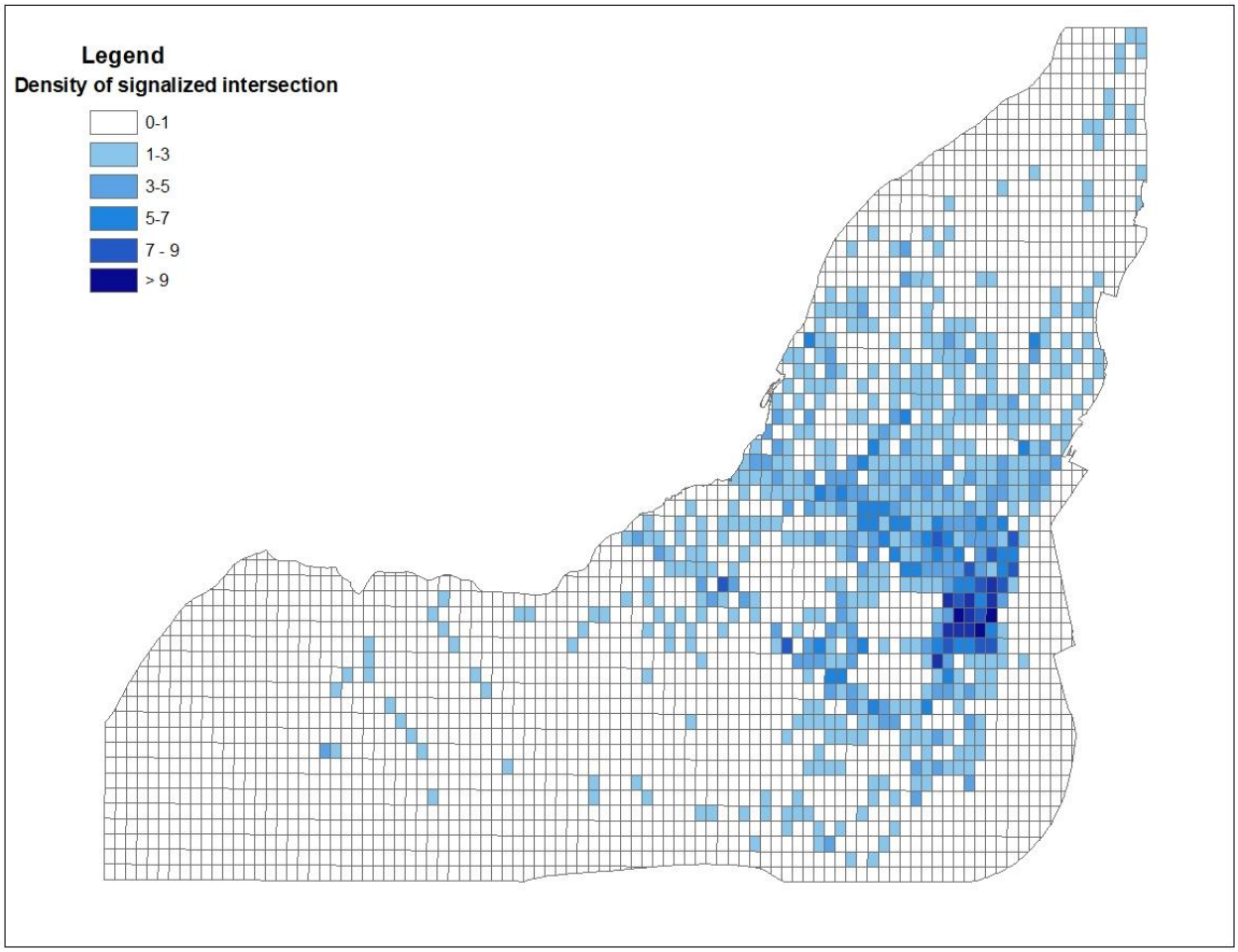


Figure A. 4 Density of signalized intersection in Montreal Road network

Additional results

1.MMRM for base model and weather types

DCP/ Weather		Base Model	
		Coef.	P-Value
V	T	-0.87	0.00
	H	0.02	0.42
	P	-0.01	0.85
	W	-0.01	0.04
Vr	T	-0.85	0.00
	H	0.03	0.34
	P	-0.01	0.80
	W	-0.02	0.05
Vm	T	-0.39	0.00
	H	0.06	0.17
	P	-0.05	0.32
	W	-0.04	0.40
V-std	T	-0.07	0.03
	H	0.02	0.48
	P	-0.02	0.24
	w	-0.01	0.44
Acc	T	0.12	0.04
	H	0.08	0.13
	P	-0.01	0.83
	W	-0.09	0.07
Dcc	T	-0.34	0.00
	H	0.00	0.96
	P	-0.12	0.02
	W	0.00	0.97
Acc2	T	0.25	0.00
	H	0.02	0.62
	P	0.08	0.11

	W	-0.04	0.39
Acc_ std	T	-0.04	0.36
	H	-0.02	0.68
	P	-0.12	0.03
	W	0.02	0.70
Idle_p	T	-0.31	0.00
	H	0.01	0.00
	P	-0.10	0.06
	W	0.01	0.91
Acc_p	T	0.77	0.00
	H	0.02	0.46
	P	0.06	0.10
	W	-0.05	0.11
Dcc_p	T	-0.17	0.00
	H	0.00	0.94
	P	-0.11	0.03
	W	0.05	0.28
Cru_p	T	-0.26	0.00
	H	-0.02	0.73
	P	-0.07	0.19
	W	0.01	0.82
Cre_p	T	-0.85	0.00
	H	0.00	0.00
	P	0.05	0.06
	w	-0.02	0.48

2.Mixed model results

Driving cycle parameters	The most important influencing factor according to highest coefficient		Coef.	P-Value
V	Fixed effect	RT3	-41	0.00
		WT7	4	0.04
		WE	8	0.03
	Combination of fixed effects	RT3WT8WE	24	0.03
Vr	Fixed effect	RT3	-38	0.00
		WT7	2	0.00
		WE	5	0.01
	Combination of fixed effects	RT3WT8WE	19	0.02
Vm	Fixed effect	RT5	-70	0.00
		WT6	-9	0.05
		WD	-6	0.02
	Combination of fixed effects	RT3WT6WD	-54	0.01
V-std	Fixed effect	RT4	-4	0.03
		WT8	-3	0.04
		WD	-5	0.00
	Combination of fixed effects	RT3WT2WD	-17	0.03
Acc	Fixed effect	RT5	0.1	0.00
		WT5	0.01	0.04
		WT7	0.01	0.03
		WE	0.02	0.00
	Combination of fixed effects	RT3WT7WE	0.5	0.01
Dcc	Fixed effect	RT3	0.2	0.02
		WT5	-0.2	0.02
		WD	-0.1	0.01
	Combination of fixed effects	RT3WT4WD	-0.7	0.00
		RT3WT7WE	-0.7	0.00
Acc2	Fixed effect	RT5	0.1	0.05
		WT5	0.1	0.00
		WD	-0.01	0.02

	Combination of fixed effects	RT3WT7WE	1.0	0.05
Acc_ std	Fixed effect	RT1	0.2	0.04
		WT5	0.2	0.00
		WD	-0.1	0.00
	Combination of fixed effects	RT2WT5WD	-0.7	0.04
Idle_p	Fixed effect	RT2	-0.75	0.00
		RT3	-0.75	0.00
		WT7	-0.57	0.00
		WD	-0.74	0.00
	Combination of fixed effects	RT3WT3WE	-1.37	0.02
Acc_p	Fixed effect	RT2	0.16	0.01
		WT7	0.02	0.04
		WE	0.07	0.03
	Combination of fixed effects	RT2WT7WE	0.46	0.01
Dcc_p	Fixed effect	RT2	0.17	0.04
		WT7	0.17	0.04
		WE	0.01	0.05
	Combination of fixed effects	RT3WT3WE	0.46	0.02
Cru_p	Fixed effect	RT3	0.6	0.01
		WT3	-0.01	0.05
		WD	-0.02	0.05
	Combination of fixed effects	RT3WT3WD	-0.4	0.04
Cre_p	Fixed effect	RT3	-0.86	0.00
		WT7	-0.37	0.05
		WE	-0.08	0.04
	Combination of fixed effects	RT3WT3WE	-0.46	0.04

3.Matrix of difference of post hoc analysis

	C1	**C2**	**C3**	**C4**	**C5**
C1	0	49.18	44.41	26.3	33.74
C2	-49.18	0	4.76	-22.88	-15.44
C3	-44.41	-4.76	0	-18.11	-10.67
C4	-26.3	22.88	18.11	0	17.44
C5	-33.74	15.44	10.67	-17.44	0

4.R-squares of driving cycle parameters in different weather conditions

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Clusterr 7	Cluster 8	Base
	R^2	R^2	R^2	R^2	R^2	R^2	R^2	R^2	R^2
V	0.43	0.58	0.55	0.44	0.46	0.42	0.45	0.44	0.38
Vr	0.44	0.58	0.56	0.42	0.48	0.4	0.42	0.45	0.39
Vm	0.6	0.75	0.69	0.52	0.67	0.4	0.42	0.57	0.4
V_std	0.53	0.75	0.66	0.32	0.66	0.4	0.41	0.46	0.41
Acc	0.49	0.68	0.65	0.32	0.54	0.24	0.28	0.38	0.2
Dcc	0.55	0.55	0.67	0.25	0.59	0.34	0.23	0.4	0.3
Acc2	0.15	0.32	0.50	0.2	0.13	0.15	0.15	0.18	0.15
Acc_std	0.45	0.37	0.60	0.33	0.57	0.2	0.29	0.39	0.23
Idle_p	0.41	0.29	0.60	0.15	0.47	0.23	0.27	0.15	0.14
Acc_p	0.61	0.72	0.65	0.38	0.57	0.35	0.45	0.43	0.4
Dcc_p	0.6	0.71	0.35	0.41	0.6	0.45	0.42	0.44	0.42
Cru_p	0.18	0.23	0.21	0.19	0.3	0.2	0.13	0.15	0.18
Cre_p	0.57	0.7	0.74	0.26	0.67	0.25	0.38	0.28	0.38