UNIVERSITÉ DE MONTRÉAL

MÉTHODE D'ANALYSE ET DE CLASSIFICATION DES SEGMENTS DU RÉSEAU ROUTIER SUPÉRIEUR DE LA RÉGION DE MONTRÉAL

ANOUAR KALBOUSSI

DÉPARTEMENT DE MATHÉMATIQUES ET GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION

DU DIPLÔME DE MAÎTRISE ÉS SCIENCES APPLIQUÉES

(GÉNIE INDUSTRIEL)

2010

© Anouar KALBOUSSI, 2010.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

MÉTHODE D'ANALYSE ET DE CLASSIFICATION DES SEGMENTS DU RÉSEAU ROUTIER SUPÉRIEUR DE LA RÉGION DE MONTRÉAL

Présenté par KALBOUSSI Anouar :

En vue de l'obtention du diplôme de : Maîtrise ès Sciences Appliquées

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

M. AGARD Bruno, Ph.D, président

M. SAUNIER Nicolas, ing.,Ph.D, membre

M. TRÉPANIER Martin, ing., Ph.D., membre et directeur de recherche

Mme. MORENCY Catherine, ing., Ph.D, membre et codirectrice de recherche

DÉDICACE

Je dédie ce travail à ma mère Wassila et mon père Brahim. Pour tous les efforts, les souffrances et sacrifices consentis pour nos personnes. Que Dieu leur accorde longue vie et santé afin qu'ils puissent jouir et bénéficier du fruit de toutes leurs privations.

À mes chers frères et sœurs, plus particulièrement à Amir, Fairouze, Mohammed et Yasmine qui n'ont cessé de m’encourager même dans les moments les plus difficiles. Qu'ils reçoivent ici toute mon affection.

À tous ceux qui m’aiment et me sont chers.

REMERCIEMENTS

Je tiens à exprimer mes vifs remerciements à tous ceux qui ont participé de près ou de loin à l'élaboration de ce travail. Je tiens à exprimer ma profonde gratitude et à remercier en premier lieu les professeurs Martin Trépanier et Catherine Morency pour l'aide et les compétences qu'ils m'ont apporté, pour leur disponibilité et leurs encouragements à finir ce travail.

Je remercie chaleureusement le professeur Bruno Agard qui m'a enseigné les méthodes et les techniques de Data Mining. Je n'oublie pas toutes les rencontres, organisées par le professeur Nicolas Saunier, qui ont nourri cette recherche.

Je tiens à remercier tous mes amis du laboratoire Génie des Transports et ceux du laboratoire Mathématiques et Génie industriel; plus précisément Amira, Khaoula et Louiselle pour leurs encouragements et leur soutien, sans lesquels je n'aurais pas été intégré.

Aussi, ses remerciements s'adressent :

- à l'École Polytechnique et aux départements de génie civil, géologique et des mines, et mathématiques et de génie industriel pour les efforts employés pour assurer la bonne formation;

- au Ministère des Transports du Québec pour son aide financière;

- à la Mission Universitaire de la Tunisie à Montréal;

- aux respectueux membres du jury pour l'honneur que vous faites en jugeant ce travail.

RÉSUMÉ

La congestion routière est un phénomène général, considérée comme une externalité extrêmement coûteuse (Prud'homme, 1999). Elle a des effets néfastes sur la prospérité économique d'un pays, le bien-être des individus et la stabilité de l'écosystème. La compréhension de ce phénomène est fondamentale afin que les autorités du transport routier soient en mesure d'apporter des solutions appropriées à ce problème.

Ce mémoire présente la méthode d'analyse et de classification des segments du réseau routier supérieur de la grande région montréalaise à partir de leurs caractéristiques physiques, puis étudie la similarité entre cette nouvelle classification et celle obtenue à partir des relevés de temps de parcours observés sur ces différents segments routiers durant la période de pointe du matin (AM).

Après avoir recensé, à l'issue d'une revue de littérature, les caractéristiques physiques des routes ayant un lien avec l'état de la circulation, ce document met en avant la méthodologie proposée afin d'atteindre l'objectif maître de ce mémoire. L'approche proposée comporte 5 étapes : faire le montage de la base de données relationnelle, dresser le portrait global du réseau routier échantillonné, synthétiser et agréger les données de facteurs physiques, grouper les tronçons selon leurs caractéristiques physiques, confronter les méthodes de groupement de tronçons et prédire l'appartenance d'un tronçon à un groupe TP à partir des facteurs physiques.

Plusieurs outils ont été utilisés pour créer la base de données des facteurs physiques, tels que Google Maps, Street View et ArcGis. Le choix de caractéristiques physiques est fondé sur de multiples études effectuées par divers auteurs sur la thématique de la congestion routière. Au moyen d'un logiciel de traitement de base de données, la table des facteurs physiques a été intégrée avec deux autres tables qui proviennent directement des travaux de Loustau *et al.(*2009), soit la table des groupes TP et la table d'état de la circulation. Ces travaux consistent à grouper les segments routiers de la grande région de Montréal selon les temps de parcours durant la période de pointe du matin (AM).

Une analyse descriptive de la base de données relationnelle qui vise à dresser le portait du réseau routier de la grande région de Montréal, a été menée. En effet, une qualification de l'état de la circulation du réseau routier montréalais a été, au premier lieu, effectué. En second lieu, une analyse descriptive des facteurs physiques des tronçons routiers échantillonnés a été réalisée.

Certaines caractéristiques se sont avérées dominantes : le type de voie « autoroute », la vitesse maximale autorisée qui varie entre 70 et 100km/h, le nombre d'intersections très faible et le nombre de voies qui varie entre 2 et 3.

Une analyse factorielle des correspondances multiples (AFCM) a été utilisée pour étudier simultanément les liens entre les différents facteurs physiques et pour agréger les données. À partir de cette méthode, 32 axes factoriels ont été obtenus, dont 12 composantes principales (représentant 63% de l'inertie totale) ont été retenues.

Ensuite, les segments routiers ont été regroupés en s'appuyant sur la matrice des composantes principales issues de l'AFCM. Pour ce faire, l'algorithme des k-moyennes a été utilisé. Douze groupes, nommés groupes FP, ont été obtenus. Puis, on a examiné le profil de chaque groupe FP, vérifiant par la suite le degré d'hétérogénéité inter et intragroupe.

Une fois que le groupement des tronçons selon les facteurs physiques a été achevé, une analyse bivariée entre les groupes TP et FP a pu être effectuée. Cette analyste vise à étudier la relation entre ces groupes au moyen de test du khi-deux (écart d'indépendance). Pour déterminer le lien entre l'état de circulation et les caractéristiques physiques des segments routiers, et également pour prédire l'appartenance des tronçons aux groupes TP, un arbre de décision a été construit. Les facteurs physiques ayant une influence significative sur l'affectation d'un segment routier à un groupe TP sont le type de voie, la vitesse autorisée, le nombre d'entrées et de sorties, le sens de la circulation et la présence de pont.

Ce travail peut servir à des études menées sur les incidents afin d'expliquer la dégradation du niveau de circulation sur certains segments routiers par leurs caractéristiques physiques.

ABSTRACT

Urban congestion is a general phenomenon, regarded as an extremely costly externality (Prud'homme, 1999). Indeed, congestion has adverse effects on the economy, society and the environment. The urban transportation authorities must understand the congestion phenomenon in order to provide adequate solutions to this problem.

This report presents the method of analysis and classification of road segments of the Greater Montreal Region highway system based on their physical characteristics. It then compares this classification to the traffic condition of road sections according to the frequency distribution of travel times observed on these various road sections during the morning rush hour (AM).

Following a definition of road congestion and of the indicators to measure it, this paper puts forward the proposed methodology to reach the main objective of this work. The approach is proposed in 5 different steps: to built the relational database, to make a global description of the sampled road system, to summarize and aggregate the data relating to physical factors, to cluster the different sections based on their physical characteristics, to compare the methods with which the sections were grouped together and to predict the TP group of section road.

Several tools have been used to create the physical factors database, such as Google Maps, Street View and ArcGIS. The choice of physical characteristics is based on a number of studies carried out by various authors on the subject of urban congestion. Using database management software, the physical factors table has been combined with two other tables taken directly from work done by Loustau *and al.* (2009), i.e. the TP groups table and the traffic conditions table. These studies consisted in grouping the road sections of the Greater Montreal Region according to travel times during the morning rush hour (AM).

A descriptive analysis of the relational database was carried out with the objective of drawing a global picture of the Greater Montreal Region road system. Thus a classification of traffic conditions on the Montreal road network was first made. Then, a descriptive analysis of the physical factors of the sampled road sections was made. Some characteristics turned out to prevail over others: the "highway" type of lanes, the maximum speed authorized, which varies between 70km/h and 100km/h, the very small number of intersections and the number of lanes, which varies between 2 and 3.

Given the high number of modalities of physical factors, the method of multiple correspondence analysis (MCA) was used to summarize and aggregate data relating to the physical factors of road sections. Using this method, 32 factorial axes were obtained, out of which 12 principal components (representing 63% of total inertia) were selected.

The road sections were then grouped on the basis of the matrix of principal components derived from the MCA. To achieve this, a segmentation algorithm used in data mining techniques was used: the K-mean method algorithm. Eight groups, called the FP groups, were obtained. We then examined the profile of each FP group and verified the degree of inter and intra-groups heterogeneity.

Once the grouping of road sections according to physical factors was completed, a bivariate analysis between FP and TP groups was carried out. This analysis aimed at determining the degree of independence between these two types of groups using the Chi-square test. We built a decision tree to determine the relationship between traffic conditions and the physical characteristics of the road sections, as well as to predict the TP groups of sections roads. The physical factors that have a significant influence on the allocation of a road section to a TP group are the type of road, the speed limit, the direction of traffic, number of inputs and outputs, and the presence of a bridge.

This work can be used for research projects on incidents in order to explain deterioration in the level of traffic on certain road sections.

TABLE DES MATIÈRES

DÉDICACE III

REMERCIEMENTS IV

RÉSUMÉ V

ABSTRACT VII

TABLE DES MATIÈRES IX

LISTE DES TABLEAUX XII

LISTE DES FIGURES XIII

CHAPITRE 1 INTRODUCTION 1

1.1 Problématique 1

1.2 Objectif 2

1.3 Structure du document 3

CHAPITRE 2 REVUE DE LITTÉRATURE 5

2.1 La congestion routière 5

2.1.1 Le concept de la congestion 5

2.1.2 Les effets néfastes de la congestion 7

2.1.3 Les typologies de la congestion 8

2.1.4 Les mesures et les indicateurs de la congestion routière 9

2.2 Les facteurs physiques liés à la congestion routière 11

2.3 Conclusion 13

CHAPITRE 3 MÉTHODOLOGIE ET MONTAGE DE LA BASE DE DONNÉES 15

3.1 Méthodologie 15

3.2 Montage de la base de données 20

3.2.1 Base de données : table des facteurs physiques 20

3.2.2 Base de données : table des groupes TP 25

3.2.3 Base de données : table d'état de la circulation 25

3.2.4 Synthèse 25

3.3 Conclusion 26

CHAPITRE 4 PORTRAIT DU RÉSEAU ROUTIER DE LA GRANDE RÉGION DE MONTRÉAL 27

4.1 L'état de la circulation sur les segments du réseau routier de la grande région de Montréal 27

4.2 Les caractéristiques physiques du réseau routier 29

4.2.1 Type de voie et nombre d'intersections 29

4.2.2 Sens de circulation 30

4.2.3 Types de barrières accotement et trottoir 31

4.2.4 Ponts et tunnels 33

4.2.5 Vitesse autorisée 34

4.2.6 Nombre de voies 36

4.3 Conclusion 38

CHAPITRE 5 GROUPEMENT DES SEGMENTS DU RÉSEAU ROUTIER DE LA GRANDE RÉGION DE MONTRÉAL SELON LEURS CARACTÉRISTIQUES PHYSIQUES40

5.1 Préparation des données 40

5.2 Analyse factorielle des correspondances multiples 40

5.3 Groupement des tronçons 43

5.3.1 Choix du nombre de groupes FP 43

5.3.2 Analyse de groupes FP 46

5.4 Conclusion 47

CHAPITRE 6 CONFRANTATION DES RESULTATS ET PREDICTION DE L'APPARTENANCE D'UN TRONÇON À UN GROUPE TP 51

6.1 Confrontation des résultats (groupes FP vs groupes TP)..... 51

6.2 Arbre de décision 54

6.3 Conclusion 60

CHAPITRE 7 CONCLUSIONS ET PERSPECTIVES 62

7.1 Contributions..... 62

7.2 Limitations 63

7.3 Perspectives..... 63

BIBLIOGRAPHIE..... 65

ANNEXE A : ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES 72

ANNEXE B : METHODE DE CLASSIFICATION PAR PARTITIONNEMENT (K-MOYENNES)..... 78

ANNEXE C: ARBRE DE DECISION C4.5 81

ANNEXE D : ÉVALUATION ET COMPARAISON DES MODÈLES..... 84

ANNEXE E : RECODAGE..... 86

LISTE DES TABLEAUX

Tableau 3.1: Description des champs de la table des facteurs physiques 21

Tableau 3.2: Type de voie 21

Tableau 3.3: Sens de déplacement sur le réseau de la grande région montréalaise 22

Tableau 3.4: Les codes de type de barrières 23

Tableau 3.5: Codes de présence d'accotement 23

Tableau 3.6: Codes de présence de trottoir 23

Tableau 3.7: Codes de présence de pont 24

Tableau 3.8: Codes de présence de tunnel 24

Tableau 3.9: Codes des vitesses autorisées 24

Tableau 3.10: Description des champs de la table d'état de la circulation 25

Tableau 5.1: Proportion des tronçons dans chaque groupe FP par rapport aux modalités de chaque facteur physique 47

Tableau 6.1: Tri croisé à l'aide de la variable « Groupe TP » 52

Tableau 6.2: Tableau de khi-deux pour le cas des groupes TP et des groupes FP 53

Tableau 6.3: Matrice de confusion des groupes TP 59

Tableau 6.4: Le taux de précision (prédiction des groupes TP) 60

Tableau A.1: Tableau disjonctif complet T (adapté de Foucart, 1984) 72

Tableau A.2: Tableau disjonctif complet 73

LISTE DES FIGURES

Figure 2.1: Diagramme fondamental : les relations entre la vitesse, la densité et le débit (FHWA, 2003) 6

Figure 2.2: Les pourcentages de la congestion récurrente et non récurrente (ECMT, 2007) 9

Figure 3.1: Méthodologies 16

Figure 3.3: Intersection avec signalisation entre Boulevard Décarie et Rue Jean Talon (A) et sortie de l'autoroute 40 (B) 22

Figure 3.4: Modèle relationnel de la base de données 26

Figure 4.1: Répartition des segments routiers par groupe TP, période AM (Loustau *et al.*, 2009) 28

Figure 4.2: Moyenne et variabilité des temps de parcours des groupes TP (Loustau *et al.*, 2009)28

Figure 4.3: Répartition des types de voies 29

Figure 4.4: Nombres moyens d'entrées, de sorties et d'intersections 30

Figure 4.5: Sens de déplacement par groupe TP 31

Figure 4.6: Accotement et trottoir à droite 32

Figure 4.7: Répartition des types de barrières à droite, par groupe TP 32

Figure 4.8: Présence de pont et/ou de tunnel par groupe TP 33

Figure 4.9: Présence de pont et de tunnel 34

Figure 4.10: La distribution des vitesses autorisées 34

Figure 4.11: Les vitesses autorisées par Groupe TP 35

Figure 4.12: Carte thématique de vitesses autorisées sur les segments routiers de la grande région de Montréal échantillonnés par le MTQ 36

Figure 4.13: Répartition de nombre de voies 37

Figure 4.14: Nombre de voies sur les segments du réseau routier de la région de Montréal échantillonnés par le MTQ 38

Figure 5.1: La valeur propre, le pourcentage d'inertie, pourcentage cumulé expliquée par les axes factoriels 41

Figure 5.2: Contribution des modalités 42

Figure 5.3: Dispersion intragroupe (w) et intergroupe (BSS) 44

Figure 5.4: valeur de pseudo-F et la différence de dispersion intragroupe (Diff W) 45

Figure 5.5: Répartition des tronçons par groupe FP 45

Figure 5.6: Valeur de dispersion intragroupe et indice de dispersion intragroupe 46

Figure 6.1: Évolution de taux de précision 55

Figure 6.2: L'arbre de décision de groupe FP 56

Figure 6.3: Répartition des tronçons par niveau de circulation 60

CHAPITRE 1 INTRODUCTION

Le transport joue un rôle déterminant dans le développement d'un pays, notamment dans le développement économique et le bien-être de la société. Toutefois, il entraine des externalités négatives ayant des effets néfastes que ce soit sur l'environnement ou sur l'être humain (Aïchour, 2006), telles que la congestion routière, les accidents de la route, la nuisance sonore, la pollution atmosphérique, etc.

La congestion routière est un phénomène général considéré comme une externalité extrêmement coûteuse (Prud'homme, 1999). Elle correspond à la gêne, directe ou indirecte, que les véhicules de la route s'imposent les uns aux autres lorsque l'utilisation du système de transport se rapproche de la capacité de ce système (Dargay et Goodwin, 1999).

Dans le but d'avoir un système de transport urbain durable et de réduire les effets néfastes de la congestion, plusieurs solutions peuvent être envisagées. Ces moyens d'action consistent à agir sur la demande et l'offre de transport, par exemple par le biais de politiques tarifaires, en assurant une meilleure gestion de la circulation, en faisant la promotion du transport public, etc (OCDE, 1999). Alors, la compréhension de la congestion est fondamentale afin que les autorités du transport soient en mesure d'apporter des solutions appropriées à ce problème. À l'instar de l'offre et de la demande de transport, la congestion varie dans le temps et dans l'espace (OCDE, 1999). En conséquence, les solutions ne peuvent pas être adoptées d'une manière uniforme sur la totalité du réseau routier de la région, d'où la nécessité d'élaborer des indicateurs pour cibler les zones les plus problématiques.

1.1 Problématique

La congestion routière ne cesse de croitre dans les grandes agglomérations et villes canadiennes. Au Québec, différents processus de collecte de données ont été menés afin de mesurer la congestion routière. Des données de temps de parcours (entre 1998 et 2004) sur différents segments routiers jugés critiques dans la région de Montréal ont été collectées par le Ministère des Transports du Québec (MTQ). À partir de ces données, des études ont été menées par les consultants de la firme MIRO (2006), puis par les ressources techniques du MTQ, afin de pouvoir caractériser et mesurer l'évolution de la congestion. Les différentes études n'ont pas fait

l'objet d'une analyse concluante parce que les relevés accumulés ne permettent pas de mesurer clairement l'évolution de la congestion (Loustau, 2009).

Dans cette perspective interviennent les travaux conduits par Loustau *et al. (*2009). L'étude menée par ces auteurs présente une analyse et une modélisation des relevés de temps de parcours sur le réseau routier montréalais au moyen des techniques statistiques avancées, afin d'en tirer son potentiel informationnel. Ces auteurs ont regroupé les segments routiers selon les distributions fréquentielles des relevés de temps de parcours durant la période de pointe du matin (Groupe TP). Dans les travaux effectués par Loustau *et al.* (2009), l'analyse de l'état de la circulation pour chaque groupe TP a été faite à partir des temps de parcours moyens sans tenir compte des caractéristiques physiques des différents segments routiers (type de voies, nombre d'intersections, nombre de voies, etc.).

1.2 Objectif

L'objectif principal de ce mémoire est de caractériser les segments du réseau routier de la grande région de Montréal à partir de leurs attributs physiques (nombre de voies, type de voie, vitesse autorisée maximale, etc.), de créer des regroupements de segments similaires puis d'étudier la similarité entre ces nouveaux regroupements et ceux bâtis sur les distributions fréquentielles des relevés de temps de parcours durant la période de pointe du matin obtenus par Loustau *et al.* (2009). Le but est de déterminer un lien possible entre les caractéristiques physiques des segments routiers et l'état de la circulation sur le réseau routier dans la grande région de Montréal.

Les étapes proposées dans l'atteinte de ces objectifs sont:

1- créer une base de données décrivant les caractéristiques physiques des segments routiers de la région de Montréal et tracer le portrait global du réseau routier échantillonné par le Ministère des Transports du Québec;

2- regrouper les segments routiers à partir de leurs caractéristiques physiques à l'aide de méthodes de *data mining* (création de groupes appelés « FP »);

3- étudier la similarité entre le nouveau regroupement obtenu et celui fait à partir des temps de parcours observés sur les segments routiers durant la période matinale (groupes appelés « TP »);

4- prédire l'appartenance d'un tronçon à un groupe TP selon leurs caractéristiques physiques.

1.3 Structure du document

Ce mémoire se divise en six chapitres. Le premier chapitre est une revue de littérature qui présente de multiples études effectuées par différents auteurs sur la thématique de la congestion routière. Le concept et les typologies de ce phénomène sont, en premier lieu, définis. Les indicateurs de la congestion routière sont par la suite décrits. Ces indicateurs traduisent implicitement la relation entre les facteurs physiques et la congestion routière. Finalement, une synthèse des caractéristiques physiques des segments routiers ayant un lien avec l'état de la circulation est effectuée.

Le second chapitre aborde, pour sa part, la méthodologie pour atteindre l'objectif de ce projet ainsi que les méthodes statistiques utilisées dans chaque étape (décrites en détails en annexe). Par conséquent, le processus du montage de la base de données relationnelle, qui regroupe les caractéristiques physiques des routes et des attributs caractérisant l'état de la circulation durant la période de pointe du matin (AM), est décrit, et les outils utilisés lors de la création de la base de données des facteurs physiques sont présentés. Les méthodes et les outils d'analyse, de groupement des tronçons et de prédiction sont aussi abordés.

Le troisième chapitre, quant à lui, dresse le portrait global du réseau routier de la grande région de Montréal échantillonné par le Ministère des Transports du Québec (MTQ). Ainsi, une caractérisation de l'état de la circulation et des caractères physiques des routes est effectuée.

Le quatrième chapitre se focalise sur le groupement des segments routiers selon leurs caractéristiques physiques. La méthode d'analyse factorielle des correspondances multiples (AFCM) est, en premier lieu, employée. Cette méthode étude simultanément les relations entre les différentes modalités, synthétise et agrège les données, puis les représente dans un plan factoriel. La contribution des modalités dans le processus de construction des plans factoriels est, par la suite, effectuée. Finalement, les segments routiers sont regroupés en s'appuyant sur la matrice des composantes principales issues de la méthode AFCM. Les groupes obtenus sont nommés groupe FP. Pour ce faire, l'algorithme de classification par partitionnement (k-moyennes) a été employé.

Le cinquième chapitre étudie le lien entre le nouveau groupement des tronçons et celui fait à partir des distributions fréquentielles des relevés de temps de parcours observés durant la période de pointe du matin. Pour ce faire, une analyse bivariée basée sur le test de khi-deux est effectuée. De plus, ce chapitre consiste à prédire l'appartenance d'un tronçon à un groupe TP à partir ses facteurs physiques. À partir de l'arbre de décision C4.5, des règles de décision sont analysées pour déterminer les facteurs physiques ayant une influence significative sur l'affectation des tronçons aux groupes TP. Une évaluation de la performance de cet arbre de décision est effectuée par la méthode de la validation croisée.

CHAPITRE 2 REVUE DE LITTÉRATURE

Ce chapitre consiste à recenser les caractéristiques physiques des routes ayant une relation avec l'état de la circulation. Toutefois, le nombre des travaux, présentant exclusivement cette relation, est limité. C'est la raison pour laquelle on a abordé des études effectuées par différents auteurs sur la thématique de la congestion routière qui ont trait, entres autres, à cette relation. Il convient préalablement de définir le concept et les typologies de la congestion. Ensuite, on se focalise sur les indicateurs et les mesures de la congestion routière. Ceci a pour objet de mettre en avant, même d'une manière implicite, les différents facteurs physiques employés par certains auteurs lors de la mesure de ce phénomène ou son évolution. Une synthèse de ces facteurs physiques qui sont liés à l'état de la circulation termine ce chapitre.

2.1 La congestion routière

2.1.1 Le concept de la congestion

Dans la littérature, on trouve plusieurs définitions de la congestion routière. On peut les regrouper selon deux approches : l'approche économique et l'approche des ingénieurs du trafic. Notre étude portera exclusivement sur la deuxième approche parce qu'elle se base sur les caractéristiques du trafic.

Dans son rapport « *Freeway Management and Operations* », la *Federal Highway Administration* (FHWA) (2003) a abordé le concept de la congestion et a présenté les événements qui se produisent dans la circulation en tant qu'une forme de la congestion. Dans le même rapport, la FHWA a présenté la relation fondamentale de la circulation entre les trois variables macroscopiques suivantes :

- La densité (D, *Density*): elle décrit la répartition des véhicules dans l'espace. Elle correspond au nombre de véhicules par unité de distance.
- Le débit (V, *Volume*) : il correspond à la répartition des véhicules dans le temps. Il est calculé comme le nombre de véhicules par unité de temps.
- La vitesse moyenne spatiale (S, *Speed*) : se calcule comme la moyenne des vitesses des véhicules entre x_1 et x_2, à l'instant t.

La relation fondamentale entre ces variables est la suivante : V=D*S. cette relation est présentée par le diagramme fondamental suivant.

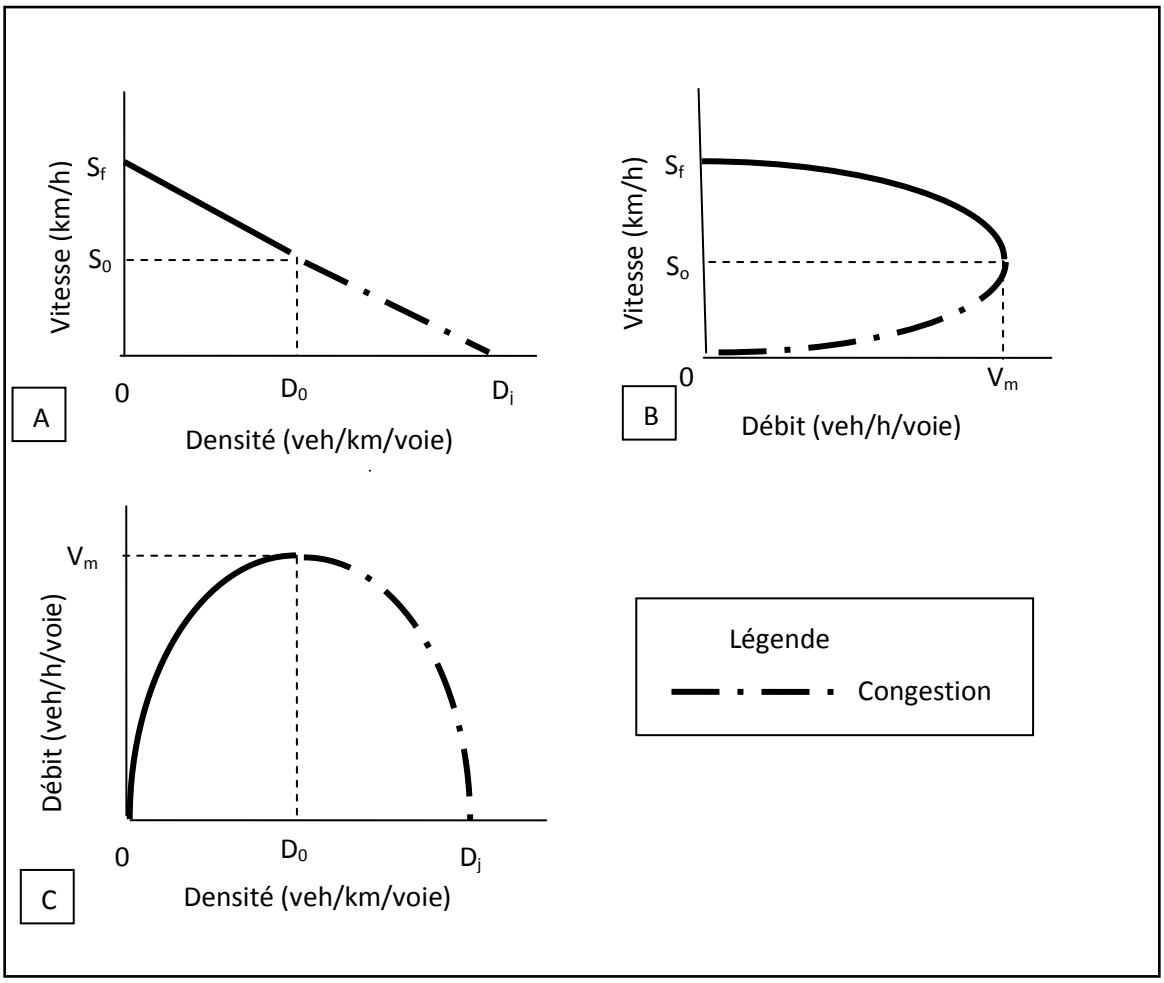

Figure 2.1: Diagramme fondamental : les relations entre la vitesse, la densité et le débit (FHWA, 2003)

Il faut noter tout d'abord que les relations entre les variables présentées par ce diagramme sont des idéalisations des données réelles et que la forme linéaire de la relation entre la vitesse et la densité revient à Greenshield (1935). À partir de la figure 2.1.B, on note que la vitesse est égale à la vitesse libre (S_f) dans la situation d'écoulement libre (lorsqu'il y a peu de circulation sur la route). Lorsque la densité atteint la densité critique (D_0), l'autoroute se rapproche de sa capacité maximale (V_m) et la vitesse d'écoulement du trafic se réduit à (S_0) (voir les figures 2.1.B et 2.1.C). Lorsque la densité atteint la densité de congestion D_j, le débit et la vitesse s'annulent (file

d'attente) (voir la figure 2.1.C). La circulation est considérée comme saturée lorsque les densités sont supérieures à la densité critique.

Le but de cette section n'est pas de recenser dans la littérature, toutes les définitions de la congestion. Cependant, on peut noter qu'il y a d'autres définitions, outres celles présentées dans le rapport de FHWA (2003), notamment celle expliquée en fonction du temps de parcours. Les prochaines sections de ce chapitre permettront, entres autres, d'exposer d'autres définitions de ce phénomène.

2.1.2 Les effets néfastes de la congestion

Le rapport « la congestion routière en Europe » de la Table Ronde 110, publié par l'OCDE à 1999, présente les conséquences de la congestion classées comme suit :

Conséquences écologiques

Lorsque le nombre de véhicules (densité) augmente sur la route, la vitesse diminue, et le temps de déplacement se prolonge. Ceci entraîne une émission supplémentaire de polluants ainsi que des nuisances sonores (Gourvil et Joubert, 2004). Les riverains de la route et les automobilistes sont les plus affectés par ces émissions. Dans ce contexte, plusieurs études ont été menées, notamment par Nesamani *et al. (*2005) qui ont montré que la qualité de l'air se dégrade lorsque la circulation n'est pas fluide.

Conséquences économiques

La congestion routière engendre un temps supplémentaire pour se déplacer du point d'origine au point de destination. Ce temps additionnel influence les délais de livraison des marchandises ainsi que le temps de déplacement. En outre, la congestion routière réduit le bassin de main-d'œuvre ainsi que l'accessibilité à des activités économiques (Robitaille et Nguyen, 2003).

À Montréal, une étude a été menée par le Ministère des Transports du Québec (2004). Le but de cette étude était d'évaluer les coûts engendrés par la congestion. D'après Gourvil et Joubert (2004), les coûts socio-économiques de la congestion en 1998 dans la région de Montréal furent estimés à 779 M$.

Conséquences sociales

Outre les conséquences écologiques et économiques, la congestion a un effet néfaste sur la société. D'après l'OCDE (1999), la réduction de la vitesse peut engendrer une diminution de contacts sociaux entre les personnes, notamment lorsque le temps de déplacement toléré est dépassé à cause de la congestion. Outre les conséquences écologiques, la pollution atmosphérique supplémentaire due à ce phénomène a des impacts négatifs sur la santé, par exemple l'effet de certains gaz sur la capacité respiratoire. De plus, la pollution acoustique a un effet sur le cadre de vie, à titre d'exemple : le bruit a un effet sur l'état physique des personnes (stress, qualité du sommeil, etc.), ainsi que sur la fréquence et la qualité des activités urbaines telles que l'activité personnelle, culturelle, etc (OCDE, 1999).

2.1.3 Les typologies de la congestion

Après avoir défini le concept de la congestion selon l'approche des ingénieurs du trafic, plusieurs études (FHWA, 2003 et 2005; ECMT, 2007) ont distingué entre deux formes de ce phénomène, soit la congestion récurrente et la congestion non récurrente.

La congestion récurrente

La FHWA (2003) a associé la congestion récurrente à un excès de la demande par rapport à la capacité de la route. En outre, il a lié ce phénomène aux déplacements pour le motif travail. En effet, la demande de transport croît en période de point du matin (AM) et du soir (PM), quand les gens se rendent au lieu de travail ou le quittent. La FHWA (2005) a ajouté dans le rapport « *Trafic Control and Systems* » que la congestion récurrente est un phénomène quotidien qui apparait sur les segments du réseau routier urbain, et peut être anticipée par les usagers de la route puisqu'elle se produit régulièrement dans la même localisation et dans la même période.

La congestion non récurrente

La congestion non récurrente est associée à la diminution de la capacité de la route, par contre la demande reste la même (FHWA, 2003) ou au moins supérieure à la capacité réduite. D'après la FHWA (2005), ce type de congestion résulte des événements aléatoires ou difficilement prévisibles qui varient d'un segment routier à un autre. Les événements principaux qui sont à l'origine de la congestion non récurrente sont les incidents dans la circulation (les accidents

graves, les véhicules en panne, etc.), les conditions météorologiques, les travaux sur les routes et les événements exceptionnels. Bien que les accidents soient imprévisibles, les segments routiers accidentogènes peuvent être identifiés au moyen de l'analyse statistique (ECMT, 2007).

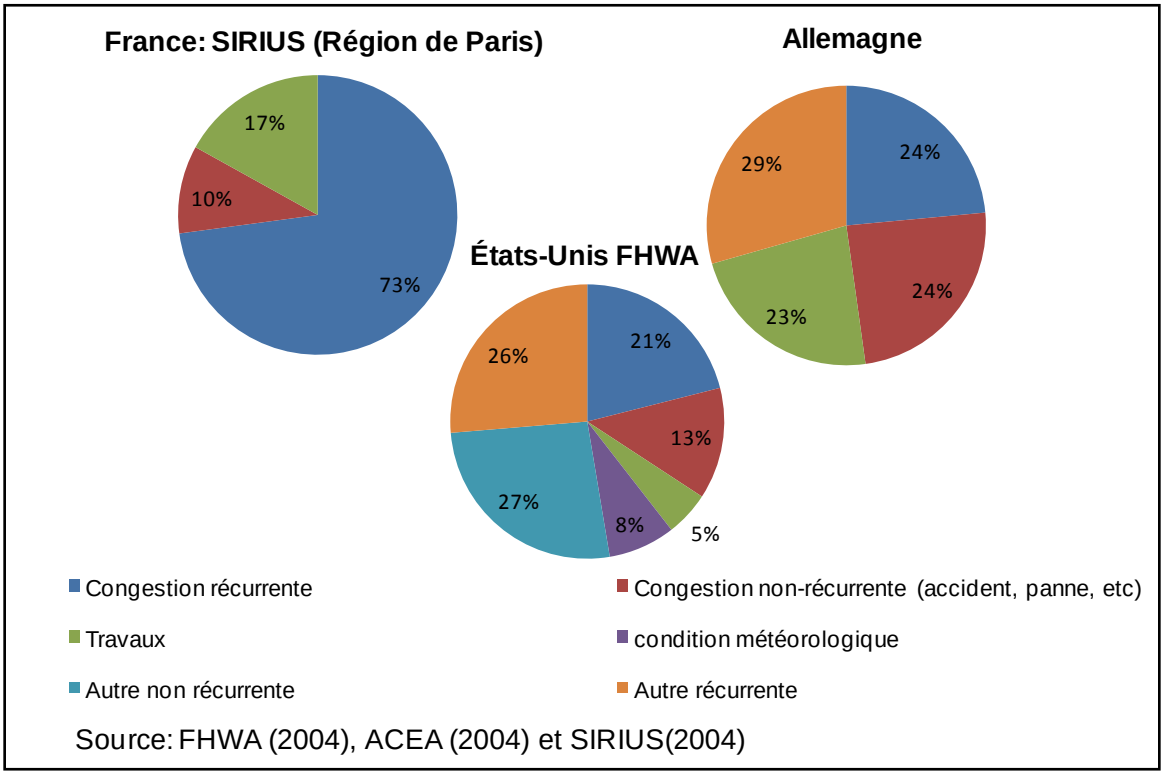

Figure 2.2: Les pourcentages de la congestion récurrente et non récurrente (ECMT, 2007)

La figure 2.2 illustre la part de chaque type de congestion dans les trois pays suivants : la France, l'Allemagne et les États-Unis. On constate à partir de cette figure que la part de la congestion récurrente, en France (région parisienne), est plus élevée par rapport aux autres types de congestion, soit de 83% de la congestion totale. Toutefois, en Allemagne, les pourcentages de la congestion récurrente, les incidents de véhicules (accident, panne, etc.) et les événements spéciaux sont assez équilibrés. Par contre, aux États-Unis, le pourcentage de la congestion récurrente est inférieur à celui de la congestion non récurrente.

2.1.4 Les mesures et les indicateurs de la congestion routière

Les indicateurs de la congestion ont un rôle important dans le processus décisionnel. Ils visent à simplifier et à mieux comprendre le phénomène de la congestion, et à détecter les défaillances du système de transport routier. Plusieurs recherches ont été entreprises sur ce sujet, notamment le

rapport 463 du NCHRP intitulé *« Economic implication of congestion »*, publié par le *Transportation Research Board* (TRB) en 2001. Ce rapport propose une classification exhaustive des mesures et des indicateurs de la congestion routière.

Les mesures basées sur le temps

D'après NCHRP (2001), les mesures basées sur le temps sont utilisées pour évaluer la congestion. Les avantages de ces mesures sont qu'elles peuvent se faire à tout moment et impliquent tous les modes de transport (NCHRP, 2001). En outre, elles trouvent leur essor dans les systèmes intelligents, notamment les systèmes d'informations en temps réel. Le rapport du NCHRP (2001) cite d'autres indicateurs qui découlent de ce type de mesures, tels que:

- le temps de parcours sur une route est la mesure la plus connue et sert de référence pour les usagers de la route afin d'évaluer la congestion.
- Le temps de parcours origine-destination est le temps estimé pour se déplacer d'une zone d'origine vers une zone de destination pour un réseau routier donné.

Les mesures basées sur les débits

Les mesures basées sur les débits sont plus attractives à cause de la grande disponibilité des données concernant le débit du trafic et le nombre de « véhicules-miles voyagés », noté VMT (*Vehicle Miles Traveled*) (NCHRP, 2001). Ce dernier est utilisé aussi dans des études menées sur la qualité de l'air. Le débit observé est souvent comparé à l'offre disponible et cette relation est exprimée en fonction du ratio débit-capacité. Tel qu'illustrée par la figure 2.1, le débit est utilisé avec la densité et la vitesse pour définir l'état de la circulation.

Les indices de la congestion

Ces indices décrivent l'état de la congestion avec un niveau d'agrégation élevé (NCHRP, 2001). En effet, ce sont des outils capables d'estimer la congestion globale sur un réseau routier, Parmi ces indices, on cite le RCI (*Roadway congestion Index)* qui est développé par Hanks et Lomax (1992) dans une étude qui porte sur la mobilité urbaine. Dans cette étude, ces derniers ont utilisé simultanément les données des « véhicules-mile voyagés » (VMT, *Vehicules-miles of Travel*) et la longueur de la voie (*lane-mile*) pour évaluer le niveau de la mobilité urbaine dans cinquante régions des États-Unis. Le RCI (*Roadway Congestion Index*) est défini comme suit :

$$RCI = \frac{\frac{(VMT_{journalier}^{aut})^2}{\text{L_Voie}_{\text{aut}}} + \frac{(VMT_{journalier}^{art})^2}{L\text{_Voie}_{\text{art}}}}{(a * VMT_{journalier}^{aut}) + (b * VMT_{journalier}^{art})}$$

Où :

$VMT_{journalier}^{aut}$ est le nombre de véhicules-mile voyagés sur une autoroute par jour

$VMT_{journalier}^{art}$ est le nombre de véhicules-mile voyagés sur une artère principale par jour

$\text{L_Voie}_{\text{aut}}$ est la longueur des voies de type autoroute en mile

$\text{L_Voie}_{\text{art}}$ est la longueur des voies de type artère en mile

a et b sont deux paramètres qui varient selon le type de voie

Les mesures de retard

Cet indicateur est utilisé pour décrire l'état de la congestion et pour illustrer la performance du réseau de transport. Le retard est la différence entre le temps de parcours observé et le temps de parcours en écoulement libre (Robitaille et Nguyen, 2003). Outre le temps de retard, le rapport 398 du NCHRP intitulé *« Quantifying Congestion »*, publié par Transportation Research Board (TRB) à 1997, met en avant d'autres mesures de retard et qui sont les suivantes :

- Taux de déplacement réel (*Travel rate*) : est le rapport entre le temps de déplacement et la longueur du segment parcouru (exprimé en minute par mile)
- Taux de retard (*Delay rate*) : représente le taux de déplacement perdu crée par la congestion et qui est égal à la différence entre le taux de déplacement réel et le taux de placement acceptable (exprimé en minute par mile). On note que le taux de déplacement acceptable est le rapport entre le temps de déplacement dans un écoulement libre et la longueur du segment parcouru.
- Retard total (*Total delay*) : représente le produit du nombre de véhicules sur le segment routier congestionné et le taux de retard (exprimé en véhicule minute).

2.2 Les facteurs physiques liés à la congestion routière

Dans la littérature, multiples travaux ont été effectués pour expliquer la relation entre la congestion et les caractéristiques des routes que ce soit d'une manière implicite ou explicite. La

congestion récurrente, telle que mentionnée dans la deuxième partie de ce chapitre, est le résultat d'un excès de la demande du transport par rapport à la capacité de la route qui reste la même (FHWA, 2003). La demande de transport dépend du motif de déplacement et les caractéristiques de la zone (zone commerciale, d'habitat, industrielle, etc.). Par ailleurs, la capacité de la route qui représente elle-même une caractéristique de la route, est calculée selon la formule issue du HCM (1997) :

$$C_f = C_i * \left(\frac{V}{C}\right) * f_d * f_m$$

Où :

C_i : la capacité horaire dans des conditions idéales

C_f : capacité horaire

V/C : le rapport volume/ capacité

f_d : facteur de réduction de la capacité pour déséquilibre directionnel

f_m : facteur de réduction pour voie et accotements étroits.

À partir de cette formule, on peut conclure que la capacité dépend de quatre facteurs physiques de la route, soit la largeur de l'accotement, le nombre de voies, le sens de circulation et la largeur de la voie. Le HCM a proposé en 2000 une nouvelle formule pour calculer la capacité en ajoutant d'autres facteurs, outre ceux présentés en 1997, tels que les facteurs de la composition de la circulation et du conflit de circulation, la présence de stationnement, etc.

De plus, certains indicateurs de la congestion sont calculés en fonction de facteurs physiques des segments routiers, notamment le RCI « Road Congestion Index » issu de Hanks et Lomax (1992). En effet, ces derniers ont exprimé cet indicateur en fonction de la longueur et le type de voie, en distinguant entre le débit du trafic sur une autoroute et celui sur une artère principale.

En outre, les caractéristiques des routes ont été mises en avant dans des études d'accidents qui représentent l'une des causes de la congestion non récurrente. Parmi ces études, on cite celle de Wang et *al. (*2009) qui porte sur les effets de la congestion sur les accidents routiers. Les facteurs physiques qui ont été évoqués sont la longueur de la route, le sens de déplacement, le nombre de voies et la géométrie de la route.

La vitesse autorisée représente aussi un facteur pour caractérisés la congestion. Dans certaines études, la congestion surviendrait lorsque la vitesse (observée) ne dépasse pas un certain seuil de la vitesse autorisée. Dans le rapport «Congestion and Accident Risque », publié par le *Département for Transport* (2003), un segment routier urbain est considéré congestionné si la vitesse moyenne est inférieure à 50% de la vitesse autorisée.

Sur le réseau routier urbain, les conducteurs se trouvent, dans certains cas, être obligés de réduire leur vitesse et par conséquent le temps de parcours augmente. En effet, ils doivent s'arrêter ou diminuer leur vitesse dans les intersections ou dans les zones à vitesse limitée, notamment à côté des écoles et des zones commerciales (Archer *et al.,* 2008).

À Montréal, le comité interrégional pour le transport des marchandises (1998) précise dans son rapport « la congestion routière et le transport des marchandises : diagnostic », que la congestion sur quelques sections de l'autoroute découle des caractéristiques géométriques de la route, à savoir :

- le nombre de voies
- la présence des échangeurs complexes qui favorisent l'utilisation de la voie gauche pour les mouvements entre les autoroutes, tels que l'échangeur Décarie
- l'absence d'accotements sur certains tronçons autoroutiers; la présence d'accotement est très utile lorsqu'un incident survient sur la route (accidents, panne, etc.)
- Le nombre élevé d'entrées et de sorties qui peut entrainer un nombre important de mouvements d'entrecroisement de véhicules sur les segments autoroutiers à débits élevés.

Ce même rapport souligne que la circulation augmente sur les ponts durant la période de pointe du matin et du soir, notamment les ponts autoroutiers.

2.3 Conclusion

Bien que le transport joue un rôle déterminant dans le bien-être des individus et dans le développement économique d'un pays, mais il représente un signe d'un « mal-développement », dont la congestion routière est un bel exemple. Ce phénomène a des effets néfastes sur la société, l'environnement et l'économie. La relation entre les caractéristiques de la route et la congestion a été évoquée implicitement dans plusieurs travaux effectués par différents auteurs. Ces travaux

portent principalement sur les mesures et l'évaluation de la congestion à partir des données telles que des données de temps de parcours, de débits, etc.

CHAPITRE 3 MÉTHODOLOGIE ET MONTAGE DE LA BASE DE DONNÉES

Afin de dresser le portait du réseau routier de la grande région de Montréal et d'étudier le lien entre les caractéristiques physiques des routes et l'état de la circulation, ce chapitre expose les différentes étapes à effectuer et les outils à employer. Il permet, entre autres, de détailler la première étape de cette étude, soit la création de la base de données des facteurs physiques et la présentation des outils ayant servi à la créer. Cette base de données sera, par la suite, croisée avec d'autres tables de données connexes afin d'étudier le lien entre les caractéristiques physiques et l'état de la circulation.

3.1 Méthodologie

La démarche scientifique proposée, telle qu'illustrée à la figure 3.1, comporte 5 étapes. Ces dernières représentent les objectifs spécifiques de ce mémoire et qui sont numérotés de 1 à 5. Il est à noter que les flèches représentent les données entrantes et sortantes pour chaque étape.

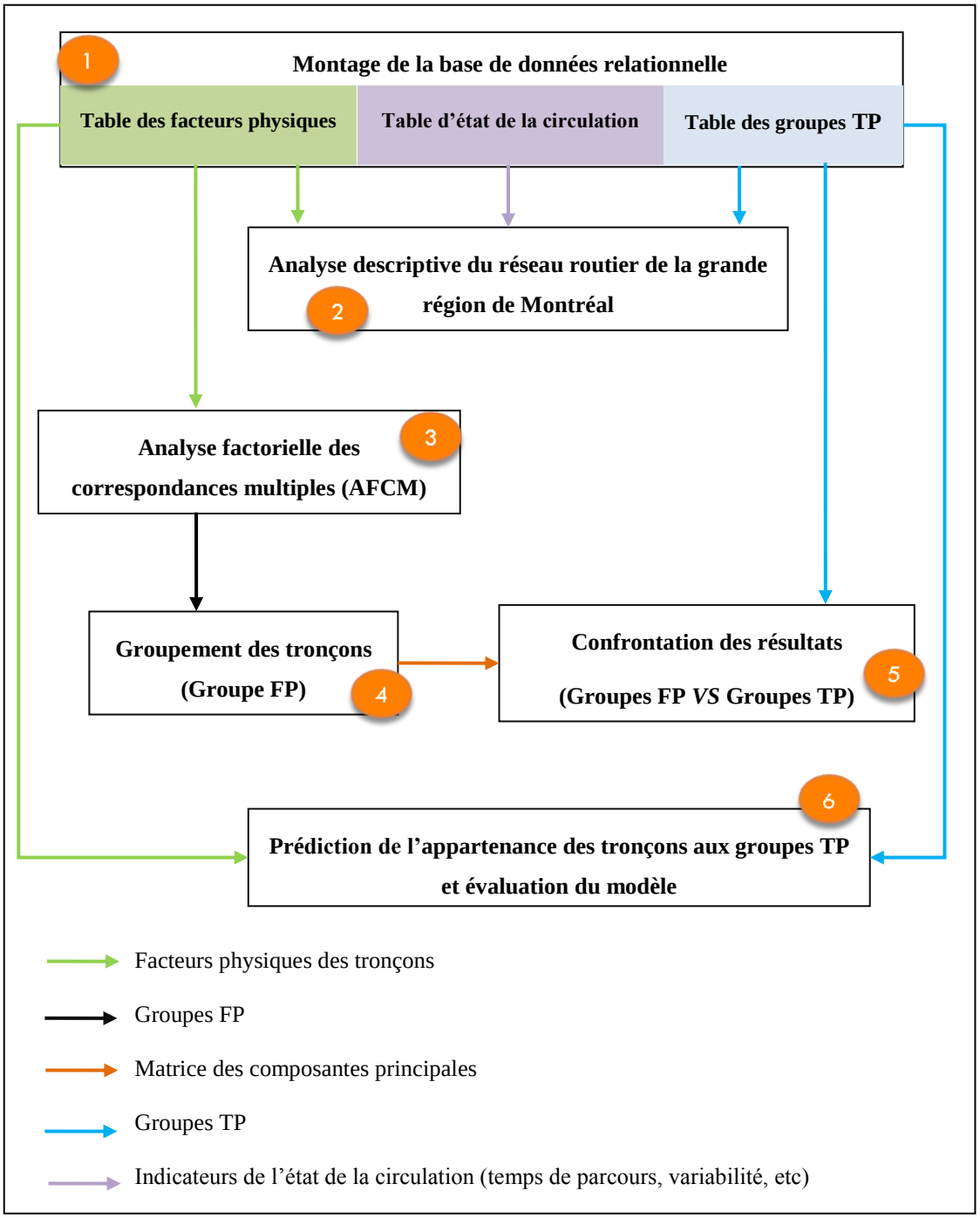

Figure 3.1: Méthodologies

Étape 1: Montage de la base de données des facteurs physiques

Afin d'étudier le lien entre l'état de la circulation et les caractéristiques physiques des segments du réseau routier de la grande région de Montréal, on a créé une base de données regroupant les caractéristiques physiques des segments routiers et les attributs caractérisant l'état de la

circulation. Cette base de données relationnelle se compose de trois tables de données provenant de sources diverses.

Les deux premières tables de données proviennent directement des travaux de Loustau *et al.(*2009), soit la table des groupes TP et la table d'état de la circulation. La première table de données est issue du regroupement des tronçons en s'appuyant sur les distributions fréquentielles des relevés de temps de parcours durant la période de pointe du matin (AM). Les groupes obtenus sont nommés groupes TP. La deuxième table de données caractérise l'état de la circulation dans chaque groupe TP par des indicateurs, soit le temps de parcours moyen et le coefficient de variabilité. La troisième table de données est appelée la table « Facteurs physiques ». Les attributs de cette dernière représentent les caractéristiques physiques des segments routiers de la grande région de Montréal échantillonnés par le MTQ.

La création de la base de données des facteurs physiques des segments routiers est le premier objectif spécifique de ce mémoire. Le but de cette étape est de mettre en exergue le montage de la base de données, de présenter les outils utilisés lors de sa création et de définir clairement les champs de la table des facteurs physiques. La suite de ce chapitre détaille cette étape.

Étape 2 : Analyse du comportement du réseau routier de la grande région de Montréal

Une fois la tâche du montage de la base de données relationnelle est achevée, une analyse descriptive permettant de dresser le portrait global du réseau routier montréalais sera effectuée. Il s'agit d'évaluer l'état de la circulation du réseau routier échantillonnée, de souligner de façon préliminaire ses caractéristiques physiques et de présenter un cadre d'analyse nécessaire aux études explicatives.

L'évaluation de l'état de la circulation sur les tronçons se repose sur la table des groupes TP et la table d'état de la circulation provenant directement des travaux de Loustau *et al.* (2009). Il s'agit de présenter les caractéristiques de chaque groupe TP, soit le temps de parcours moyen et la variabilité, pour dégager finalement l'état de la circulation sur les segments routiers pour la période du matin (AM).

L'analyse descriptive du réseau routier de la grande région du Montréal s'appuie sur la base de données relationnelle. Il s'agit de croiser les attributs de la table des facteurs physiques avec ceux

de la table des groupes TP. Cette étape permettra, ainsi, de présenter de façon préliminaire le lien entre les caractéristiques physiques des routes et l'état de la circulation.

Étapes 3 : Analyse factorielle des correspondances multiples

Cette étape est primordiale avant de regrouper les tronçons selon leurs facteurs physiques. Elle vise à étudier simultanément les relations entre les différents facteurs physiques caractérisant les segments routiers, et aussi à agréger les modalités des caractéristiques physiques liées entre elles à des variables indépendantes qui seront utilisées, dans la prochaine étape, pour regrouper les tronçons. Pour ce faire, la méthode d'analyse factorielle des correspondances multiples (AFCM) a été utilisée. Cette méthode a le même principe que l'analyse du tableau de contingence qui permet d'étudier les liens entre deux variables qualitatives au moyen du test de khi-deux. La méthode AFCM est utilisée dans cette étude vu qu'il y a plus de deux variables qualitatives (facteurs physiques) à étudier.

Le choix de nombre de dimensions de plan factoriel où les tronçons seront représentés est fonction de la valeur propre (Jambu, 1989). De plus, les coordonnées de chaque segment routier dans le nouveau plan multidimensionnel sont représentées dans une matrice appelée «matrice des composantes principales». Il faut noter que les composantes principales sont les modalités des facteurs physiques agrégées à des variables indépendantes. Enfin, les liens entre les caractéristiques physiques seront étudiés à partir de l'analyse de la contribution des modalités dans la construction des axes factorielles. La méthode d'AFCM est expliquée en détail dans l'annexe A.

Étape 4 : Groupement des tronçons selon les facteurs physiques

Cette étape consiste à grouper les segments routiers à partir de la matrice des composantes principales issue de l'AFCM. Pour ce faire, la méthode des k-moyennes développée par MacQueen (1967) a été utilisée. Les groupes obtenus sont nommés groupes FP. L'algorithme de la méthode k-moyennes et les critères de fonctionnement sont expliqués en détail dans l'annexe B.

Pour choisir le nombre de groupes, on a appliqué plusieurs fois l'algorithme de classification k-moyennes, en augmentant chaque fois le nombre de groupes. Puis, on a choisi le nombre de groupes en s'appuyant sur deux indicateurs proposés par Clanski et Harabasz (1974), et Guidici

(2003), soit pseudo-F et R-carré. Ces indicateurs sont calculés en fonction de la valeur de dispersion intragroupe et la valeur de dispersion intergroupe. Ces indicateurs sont définis dans l'annexe B.

Enfin, une analyse descriptive des groupes obtenus sera menée. Ceci permet de mettre avant les particularités des tronçons échantillonnés par le MTQ.

Étape 5 : Confrontation des deux méthodes de groupements (groupes TP vs groupes FP)

Cette étape consiste à confronter les résultats obtenus à partir des deux méthodes de classification, soit le groupement à partir de temps de parcours et le groupement selon les caractéristiques physiques des routes. Pour ce faire, on a analysé le profil des groupes TP par rapport aux groupes FP et vice-versa. Ensuite, une analyse bivariée des groupes TP et FP au moyen de la méthode khi-deux, a été réalisée. Enfin, on a étudié les interactions entres les différentes modalités de lesdits groupes à partir de la contribution aux Khi-deux.

Étape 6 : Prédiction de l'appartenance des tronçons aux groupes TP et validation des résultats

Dans la dernière étape, il s'agit de mieux appréhender l'état de la circulation et de prédire l'appartenance des segments routiers aux groupes TP à partir de leurs facteurs physiques. Il s'agit, entres autres, de dégager la circulation normale; c'est à dire celle exprimée par les caractéristiques physiques des tronçons. Pour ce faire, l'algorithme C4.5 développé par Quinlan (1993) sera utilisé. Ceci permet de construire un arbre de décision. Les règles issues de cet arbre seront analysées pour en ressortir les facteurs physiques ayant une influence significative sur l'affectation des segments routiers au groupe TP. On fait appel à cet algorithme vu qu'il tient en compte, entre autres, des variables discrets et des poids des observations, et qu'il se caractérise par l'élagage qui facilite la décision finale (Devéze et Fouquin, 2004).

Pour choisir la taille optimale de l'arbre de décision, on a appliqué plusieurs fois l'algorithme C4.5, en modifiant à chaque fois le critère d'arrêt de la segmentation des nœuds (nombre minimal de tronçons dans le nœud à segmenter). Le choix de la taille de l'arbre est fonction du taux de précision calculé par la méthode de validation croisée. Cette méthode consiste à mesurer la précision du modèle à classer correctement les tronçons. L'algorithme C4.5 et la méthode de validation croisée sont présentés respectivement dans l'annexe C et l'annexe D.

3.2 Montage de la base de données

Tel que mentionné dans la section précédente, le montage de la base de données est la première étape dans la méthodologie et qui représente également le premier objectif spécifique de cette étude. La base de données relationnelle est constituée par trois tables qui sont liées, soit la table des facteurs physiques, la table d'état de la circulation et la table des groupes TP.

3.2.1 Base de données : table des facteurs physiques

Afin de décrire le portrait du réseau routier de la grande région de Montréal, la création d'une base de données relative aux caractéristiques physiques des tronçons a été effectuée. Pour ce faire, plusieurs outils ont été utilisés, tels ArcGIS, Google maps, Google Street View, etc.

Au moyen du logiciel ArcGIS, il s'agit, tout d'abord, d'identifier la position du tronçon à caractériser sur la carte du réseau routier de la grande région de Montréal échantillonné par le MTQ. Ensuite, on a repéré la position de ce tronçon sur la carte fournie par Google maps. Certains facteurs nécessitent l'utilisation de Google Street View pour bien décrire les caractéristiques des tronçons, tels que la vitesse affichée, le type de barrière, etc. Finalement, les caractéristiques de chaque segment seront saisies dans un fichier Excel, nommé facteurs physiques.

Le tableau suivant représente la table des facteurs physiques telle qu'elle a été insérée dans la base de données relationnelle. Il faut noter que seulement 681 tronçons ont été caractérisés parmi 811 tronçons échantillonnés par le MTQ. En effet, les outils qu'on a utilisés tels que Google Maps, n'ont pas permis de caractériser 130 tronçons. Ces derniers sont exclus de cette étude. Les champs sont décrits dans la suite de cette partie.

Tableau 3.1 : Description des champs de la table des facteurs physiques

Nom du Champ	Exemple	Description	Type de champ
TRUNI	102	Identifiant du tronçon unique	continue
ID_FP	102	Identifiant du tronçon caractérisé	continue
Nbr_sorties	0	Nombre de sorties	continue
Nbr_entrées	0	Nombre d'entrées	continue
Nbr_intersections	2	Nombre d'intersections	continue
sens_HC/VC	HM	Sens de la circulation	catégorielle
Type_voie	BV/AV	Type de voie	catégorielle
Acc_D	Non	Présence d'accotement à droite	catégorielle
Tr_D=Oui	Oui	Présence de trottoir	catégorielle
Type_Barr_D	Absence de Barrière	Type de barrière à droite	catégorielle
Acc_G	Non	Présence d'accotement à gauche	catégorielle
Type_Barr_G	Absence de Barrière	Type de barrière à gauche	catégorielle
sur_pont	Non	Présence de ponts	catégorielle
Tunnel	Non	Présence de tunnels	catégorielle
Nbr_voies	3	Nombre de voies	continue
Vitesse_aut	70Km/h	Vitesse affichée en km/h	catégorielle

Type de voie

Les tronçons sont classés en trois types de voiries et qui sont représentés par le tableau suivant. À partir de Google Maps et de la carte du réseau routier échantillonné par le MTQ (fournie par ArcGIS), on a identifié le nom du segment routier. À partir de cela, on a déterminé le type de voie de chaque tronçon. À titre d'exemple, le tronçon ayant un nom « Autoroute 40 », sont classé dans la classe autoroute. De plus, on a agrégé les types de voies « boulevard » et « avenue », et aussi les voies de type « Route » et « Rue ».

Tableau 3.2 : Type de voie

Code	Type de Voie
Rue /Route	rue ou route
BV/AV	boulevard ou avenue
AU	autoroute

Sens de circulation

Les segments routiers sont classés sous deux catégories ayant les codes suivants.

Tableau 3.3 : Sens de déplacement sur le réseau de la grande région montréalaise

Code	Sens de déplacement
VM	Vers Montréal
HM	Hors Montréal

Nombre d'intersections, d'entrées et de sorties

Le nombre d'intersections concerne les tronçons de type « Boulevard/Avenue » et « Rue/Route » où le flux de trafic peut être interrompu par les feux de signalisation, les carrefours giratoires, les panneaux « Stop », etc. Les entrées et les sortis concernent seulement les autoroutes. La figure suivante représente un exemple d'une intersection et d'une sortie.

Figure 3.2: Intersection avec signalisation entre Boulevard Décarie et Rue Jean Talon (A) et sortie de l'autoroute 40 (B)

Types de barrières, accotements et trottoir

Pour caractériser les tronçons, on a déterminé le type de barrière de sécurité. Ce facteur intervient dans le cadre de la sécurité routière. En effet, la barrière permet de limiter les conséquences de la sortie des véhicules des voies de circulation. En utilisant Google Street view, on a pu remplir les colonnes « Type_Barr_D » et « Type_Barr _G » du fichier Excel « Facteurs physiques ». Le point faible de ces données est la présence, sur le même segment, de plus d'un type de barrière.

De ce fait, on a retenu celui le plus représentatif, soit la barrière la plus longue. Le tableau 3.4 représente les codes des types de barrières.

Tableau 3.4 : Les codes de type de barrières

Code	**Description**
Absence de barrière	il n'y a pas de barrière
Présence d'un mur	mur
Barrière métallique	barrière métallique
Barrière béton	barrière béton

Un autre facteur qui permet d'évaluer le niveau de sécurité sur une route est la présence d'accotements (Tableau 3.5). Cette bande, appelée aussi une bande d'arrêt d'urgence, peut être utilisé comme une voie dans lorsqu'un incident est survenu ou aussi dans le cas d'arrêt d'urgence d'un véhicule. Les codes de présence de trottoir sont présentés dans le tableau3.6.

Tableau 3.5 : Codes de présence d'accotement

Code	**Description**
Non	Sans accotement
Oui	Avec accotement

Les codes de trottoir sont représentés par le tableau suivant.

Tableau 3.6 : Codes de présence de trottoir

Code	**Description**
Oui	Présence de trottoir
Non	Absence de trottoir

Ponts et tunnels

On a ajouté deux autres facteurs physiques, soit la présence de ponts ou de tunnels. Ces facteurs nous permettent de savoir si le passage sur un pont ou dans un tunnel influence les

comportements des conducteurs et l'état de la circulation. On a récolté ces données via Google Street View, en alimentant le fichier des facteurs physiques.

Tableau 3.7 : Codes de présence de pont

Code	Description
Oui	Présence de pont
Non	Absence de pont

Tableau 3.8 : Codes de présence de tunnel

Code	Description
Oui	Présence de tunnel
Non	Absence de tunnel

Vitesse autorisée

Au moyen de Google maps, on a déterminé les vitesses autorisées sur les tronçons. Sur certains segments, il n'y pas des panneaux affichant les vitesses autorisées. Dans ce cas, le tronçon sera qualifié par la même vitesse autorisée de tronçon qui le précède. Dans certains cas, la vitesse affichée sur certains tronçons est représentée par une vitesse maximale de 100km/h et une vitesse minimale de 60 km/h. Le tableau suivant représente les codes des vitesses autorisées.

Tableau 3.9 : Codes des vitesses autorisées

Code	Description
35km/h	La vitesse autorisée est égale à 35 km/h
45km/h	La vitesse autorisée est égale à 45 km/h
50km/h	La vitesse autorisée est égale à 50 km/h
60km/h	La vitesse autorisée est égale à 60 km/h
65km/h	La vitesse autorisée est égale à 65 km/h
70km/h	La vitesse autorisée est égale à 70 km/h
80km/h	La vitesse autorisée est égale à 80 km/h
90km/h	La vitesse autorisée est égale à 90 km/h
100-60km/h	La vitesse autorisée est entre 100 km/h et 60 km/h

Nombre de voies

Le nombre de voies est l'un des facteurs physiques qui caractérisent les segments du réseau routier. La capacité de la route varie proportionnellement avec ce facteur. À l'aide de Google Street View, on a déterminé le nombre de voies qui, dans certains cas, varie sur le même tronçon. Cette variation peut se répercuter légèrement sur la fiabilité des données récoltées.

3.2.2 Base de données : table des groupes TP

La table des groupes TP provient directement des travaux de Loustau *et al.* (2009). Ces derniers ont regroupé les tronçons selon les distributions fréquentielles des relevés de temps de parcours durant la période de pointe du matin (AM). Pour ce faire, un algorithme issu du data ming, a été utilisé. La table des groupes TP se compose des deux champs, soit le numéro du tronçon unique et le numéro du groupe TP. Cette table permet d'identifier à quel groupe TP appartient un tronçon.

3.2.3 Base de données : table d'état de la circulation

La table d'état de la circulation apporte des informations relatives à chaque groupe TP. Elle aussi provient des travaux de Loustau *et al.* (2009). Le tableau suivant illustre les caractéristiques des groupes TP en décrivant chacun des champs.

Tableau 3.10 : Description des champs de la table d'état de la circulation

Champs	Exemple	Description
Groupe TP	TP2	Le nom du groupe TP
Temps_parcours_moy	39,2	Le temps de parcours moyen (secondes) durant la période de pointe du matin (AM)
Temps_parcours _Var	63,8%	Le coefficient de variabilité de temps de parcours
Seg(%)	32,6%	Le pourcentage des segments appartenant au groupe TP

3.2.4 Synthèse

Pour des fins d'analyse, les tables de données mentionnées plus haut doivent être utilisées conjointement. Ainsi, ces tables qui sont présentées chacune dans un fichier Excel ont été

intégrées dans un logiciel de traitement de base de données. La figure suivante illustre les liens entre ces différentes tables.

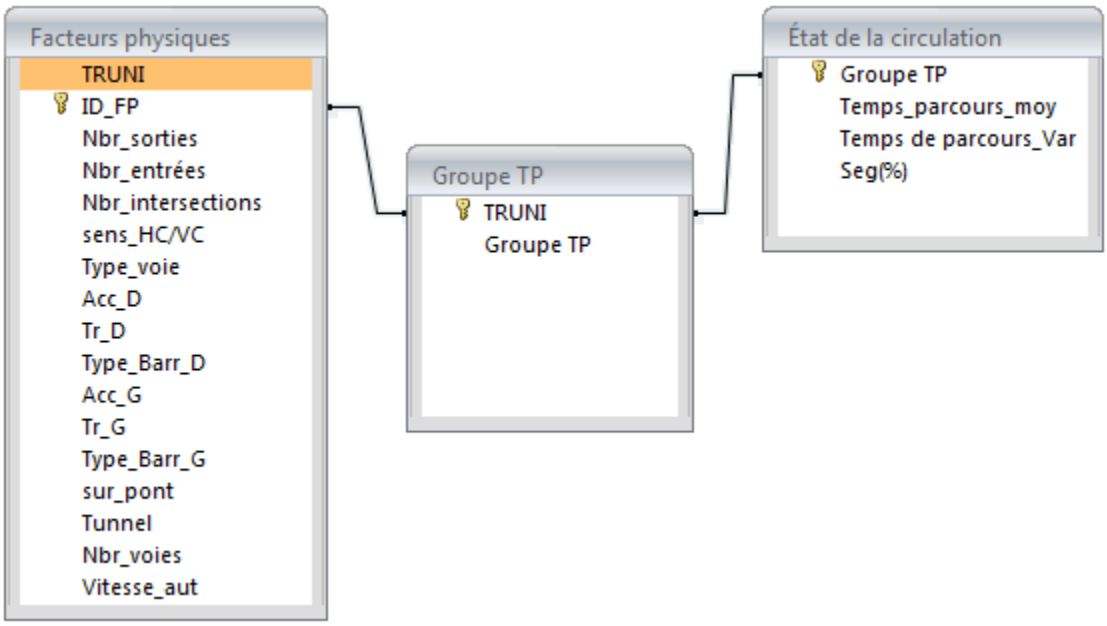

Figure 3.3: Modèle relationnel de la base de données

3.3 Conclusion

La partie méthodologie présente la démarche scientifique à suivre afin d'atteindre l'objectif de ce mémoire. De plus, elle a exposé les méthodes et les outils d'analyse empruntés. Ensuite, ce chapitre a détaillé la première étape de la méthodologie proposée, soit le montage de la base de données relationnelle. En effet, il s'agit de définir les champs de différentes tables de données ainsi que les outils ayant servi à les créer. Finalement, une base de données relationnelle a été obtenue en intégrant les tables de données dans un logiciel de traitement de base de données.

CHAPITRE 4 PORTRAIT DU RÉSEAU ROUTIER DE LA GRANDE RÉGION DE MONTRÉAL

Ce chapitre consiste à dresser le portrait global du réseau routier de la grande région de Montréal échantillonné par le MTQ. Ainsi, une évaluation de l'état de la circulation est, au premier lieu, effectuée. Pour ce faire, il s'agit d'exploiter conjointement les données de table des groupes TP et celles de table d'état de la circulation qui proviennent directement des travaux de Loustau *et al.*(2009). Ensuite, une caractérisation physique du réseau routier est menée. Cette caractérisation effectuée souligne, d'une manière préliminaire, le lien entre les caractéristiques des segments du réseau routier et l'état de la circulation. Il s'agit donc d'exploiter la base de données relationnelle en croisant les caractéristiques physiques des routes avec les groupes TP.

4.1 L'état de la circulation sur les segments du réseau routier de la grande région de Montréal

En se référant aux travaux de Loustau *et al.* (2009), l'état de la circulation des segments durant la période de pointe du matin (AM) est représenté par huit groupes, nommés groupes TP. Dans cette étude, 811 segments routiers ont été échantillonnés. La figure 4.1 illustre la répartition de tronçons dans chaque groupe TP. À partir de cette figure, on observe que les tronçons ne sont pas répartis d'une manière égale. Notamment, le groupe TP2 regroupe un tiers du nombre total de segments, contre 2% pour le groupe TP8.

La figure 4.2 représente la moyenne et le coefficient de variation de temps de parcours de chaque groupe TP. À partir de la courbe de temps de parcours, on observe que les groupes TP 4 et 8 se caractérisent par un temps de parcours moyen très élevé par rapport aux autres groupes TP. De plus, les temps de parcours moyens de groupes TP 2 et 3 sont les moins élevés, soit moins de 50 secondes. Ceci signifie que l'état de la circulation sur les tronçons de ces groupes est plus fluide. À partir de la courbe de variabilité, on constate que les groupes TP 2 et 4 se caractérisent par une variabilité de temps de parcours plus élevée par rapport aux autres groupes TP. On peut noter que le groupe TP 2 est fluide, mais avec une variabilité remarquable. La dispersion de temps de parcours des groupes TP6 et 8 est plus faible que celle des groupes TP2 et 4.

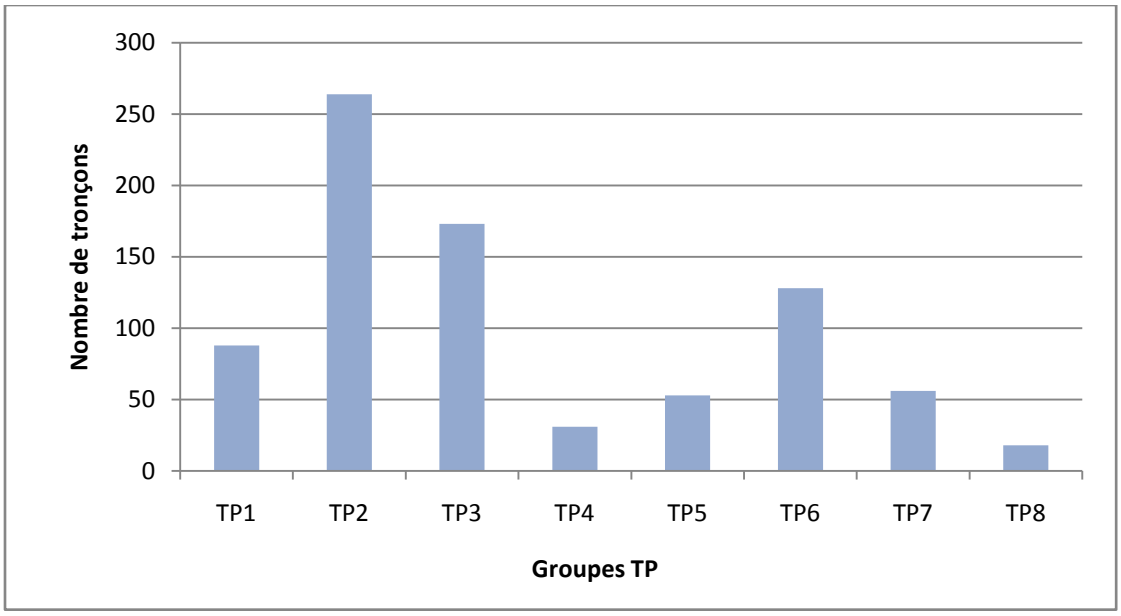

Figure 4.1: Répartition des segments routiers par groupe TP, période AM (Loustau *et al.*, 2009)

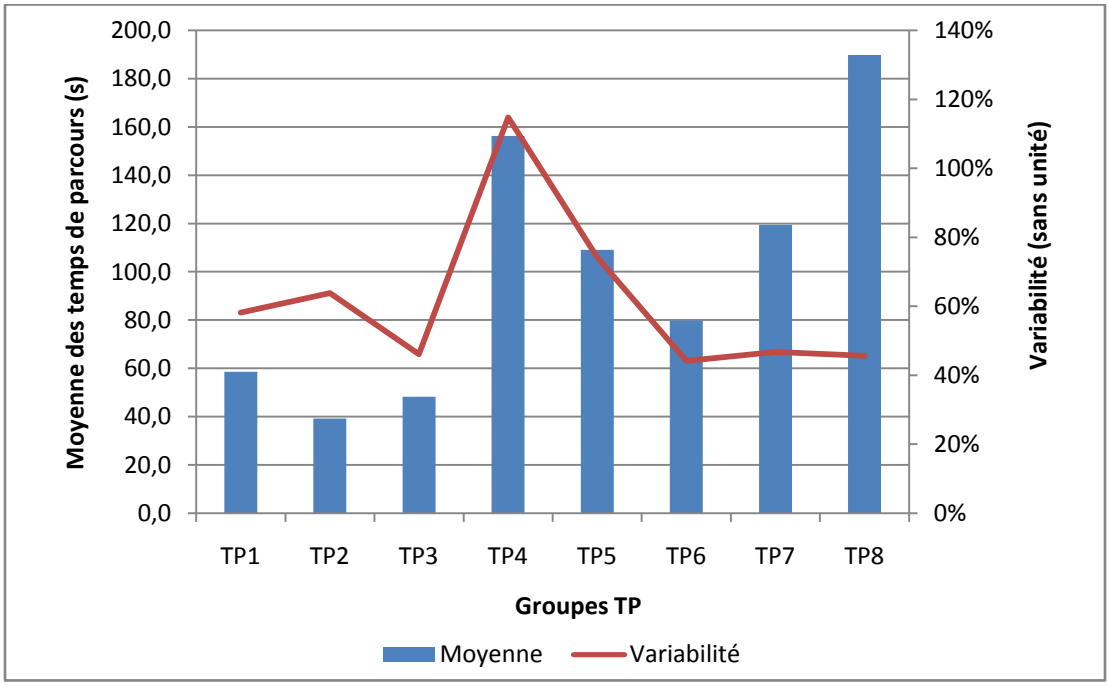

Figure 4.2: Moyenne et variabilité des temps de parcours des groupes TP (Loustau *et al.*, 2009)

4.2 Les caractéristiques physiques du réseau routier

4.2.1 Type de voie et nombre d'intersections

La majorité des tronçons sont de type «Autoroute», soit plus de 68% de nombre total de segments, contre près de 25 % pour les tronçons de type « Boulevard/Avenue », et de 7 % de type «Rue/Route» (figure 4.3).

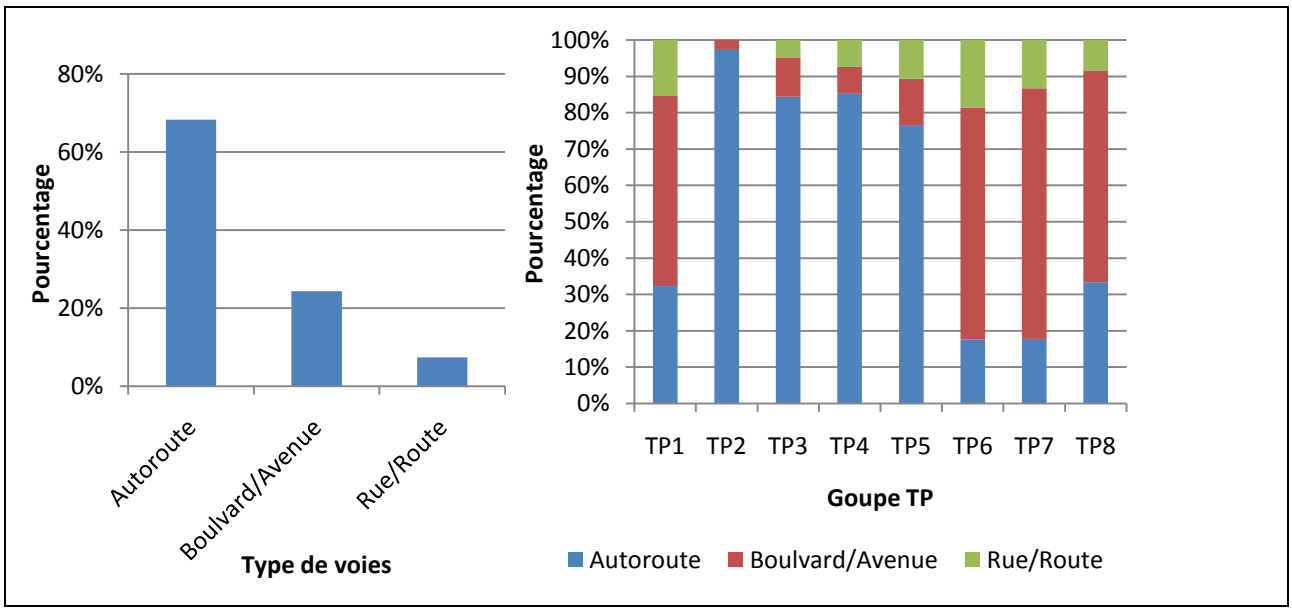

Figure 4.3: Répartition des types de voies

En vertu de cette figure, on observe aussi que la moitié des groupes se caractérisent par la domination de type de voie « Autoroute » et le reste par le type « Boulevard/Avenue ». Notamment, 97 % des tronçons du groupe TP2 sont des autoroutes et 69 % des tronçons du groupe TP7 sont des boulevards et des avenues. On s'aperçoit que l'état de la circulation est plus fluide sur les tronçons de type « autoroute », hormis les tronçons de groupe TP4.

La figure 4.4 représente le nombre moyen d'intersections, d'entrées et de sorties. À partir de celle-ci, on constate que l'état de la circulation d'un groupe ayant un nombre d'intersections le plus élevé est le moins fluide, tels que les groupes TP6, TP7 et TP8. De plus, on remarque que les nombres moyens d'entrées et de sorties des groupes TP2, TP3, TP4 et TP5 sont plus élevés par rapport aux nombres moyens d'intersections. Ceci peut être expliqué par le fait que la majorité des tronçons de ces groupes sont des autoroutes. En outre, le groupe TP4 se distingue des groupes

TP2 et TP3 au niveau d'état de la circulation; il se caractérise par un faible nombre d'intersections, mais le temps de parcours moyen sur les tronçons de ce groupe est plus élevé.

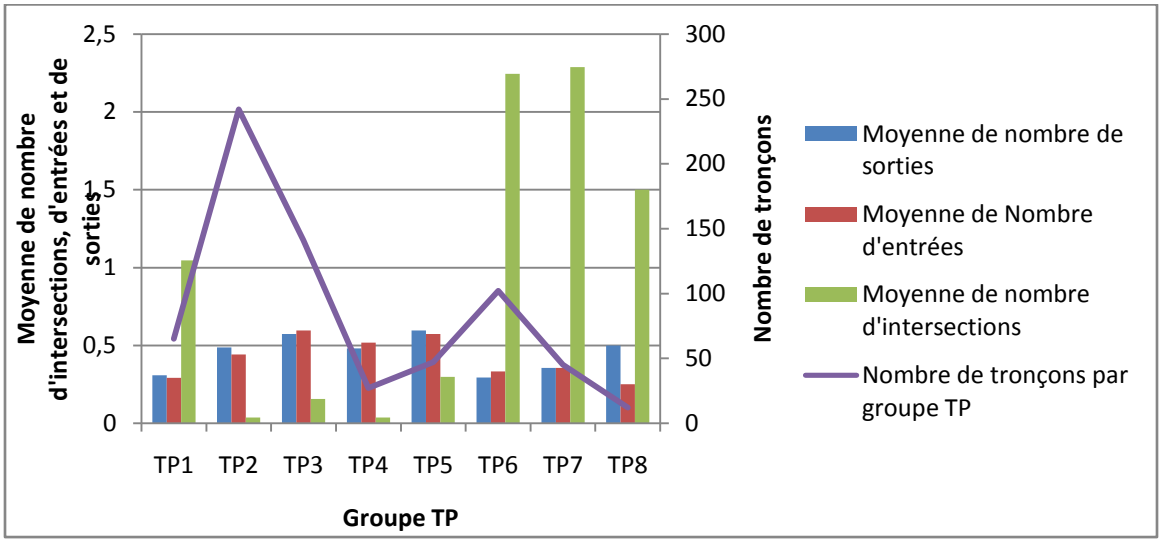

Figure 4.4: Nombres moyens d'entrées, de sorties et d'intersections

4.2.2 Sens de circulation

La figure 4.5 présente la répartition des tronçons par groupe TP selon le sens de circulation. En vertu de cette figure, on constate que la majorité des tronçons qui appartiennent aux groupes TP2 et TP3, caractérisés par la fluidité de la circulation, sont orientés hors le centre ville de Montréal. Les groupes TP4 et TP8 qui se caractérisent par une circulation non fluide incluent plus des tronçons orientés vers le centre ville, soit plus de 80% des tronçons. Ces résultats nous semblent logiques; les segments routiers qui amènent au centre ville sont plus congestionnés (durant la période de pointe du matin AM) que les tronçons de type « hors centre ville». On constate aussi qu'il n'y a pas un lien entre l'état de la circulation et le sens de déplacement pour les groupes TP1 et TP6, les tronçons de ces groupes sont les plus hétérogènes.

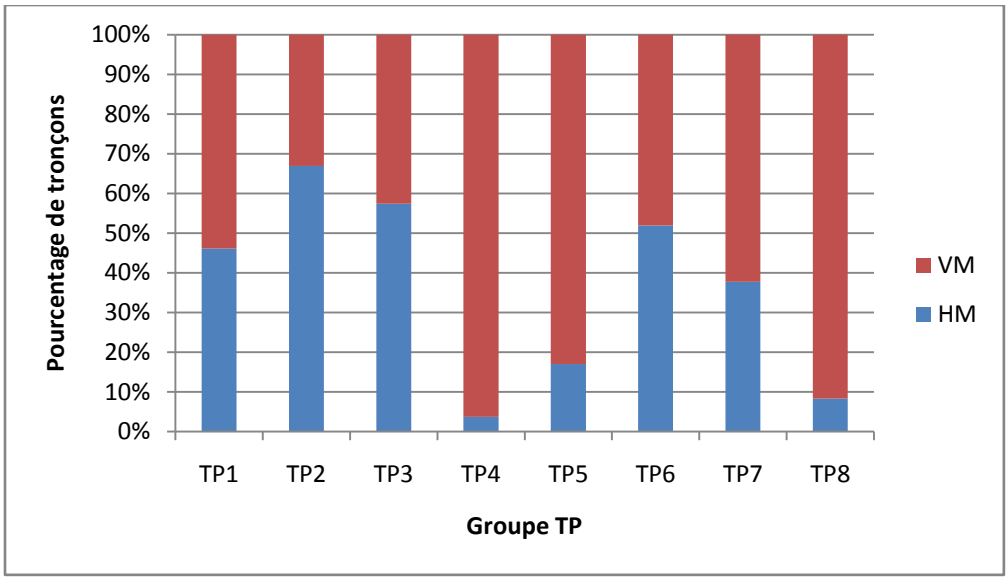

Figure 4.5: Sens de déplacement par groupe TP

4.2.3 Types de barrières, accotement et trottoir

La figure 4.6 illustre la présence ou l'absence d'accotement et de trottoir. On constate que plus de 90% des tronçons du groupe TP4 ont un accotement et que 66% des tronçons de groupe TP8 se caractérisent par l'absence d'accotement. L'absence d'accotement sur les segments de groupe TP8 peut réduire la capacité de la route lors de l'arrêt d'urgence des voitures en panne. Le groupe TP1 est le groupe le plus hétérogène; il regroupe autant des tronçons avec accotement que sans accotement. On observe aussi à partir de cette figure que les groupes TP6, TP7 et TP8 se caractérisent par la présence de trottoir par rapport aux autres groupes. Ceci nous semble logique parce que la majorité des segments de ces groupes sont des boulevards ou des avenues. On s'aperçoit que les tronçons ayant un état de la circulation fluide sont ceux qui se caractérisent par la présence d'accotement et l'absence de trottoir, hormis les tronçons des groupes TP1 et TP4.

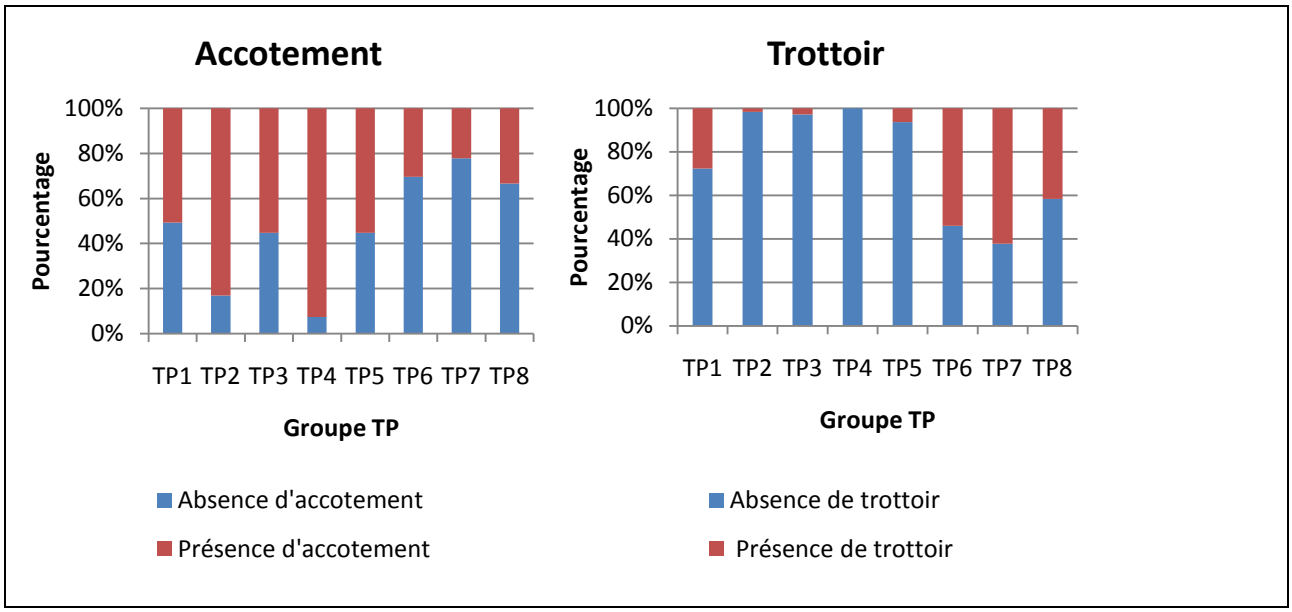

Figure 4.6 : Accotement et trottoir à droite

La figure suivante représente la répartition des types de barrières à droite de la chaussée par groupe TP. La majorité des tronçons des groupes TP6 et TP7 se caractérisent par l'absence de barrière à droite, ceci est logique parce que les tronçons de ces groupes sont de type « Boulevard/Avenue » ou « Rue/Route ». En vertu de cette figure, il semble que l'état de circulation n'est pas lié au type de barrière. En effet, les groupes ont presque les mêmes caractéristiques, hormis les groups TP6 et TP7.

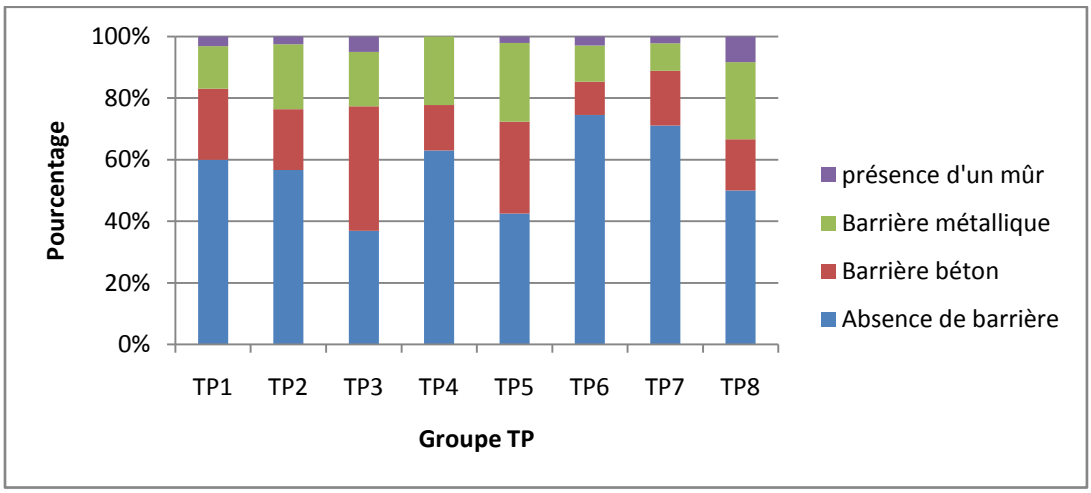

Figure 4.7 : Répartition des types de barrières à droite, par groupe TP

4.2.4 Ponts et tunnels

La figure 4.8 illustre la présence de pont et de tunnel par groupe TP. À partir de cette figure, il semble que la présence de pont a une influence négative sur l'état de la circulation. En effet, le temps de parcours moyen sur les tronçons de groupe TP8 qui se caractérise par la présence de pont, est le plus élevé. De même, si on fait une comparaison entre les groupes de type « autoroute », hormis le groupe TP4, on constate que la présence de pont engendre une augmentation de temps de parcours. À partir de cette figure, on observe aussi qu'il y a un faible pourcentage de tronçons par groupe TP ayant de tunnel. Ceci s'explique par le faible nombre de tunnels par rapport au nombre total de tronçons. Tel qu'illustrée par la figure 4.9, trois tunnels ont été caractérisés sur le réseau routier échantillonné par le MTQ. On observe aussi de cette figure que la plupart des ponts sont des ponts autoroutiers, notamment sur l'autoroute 40.

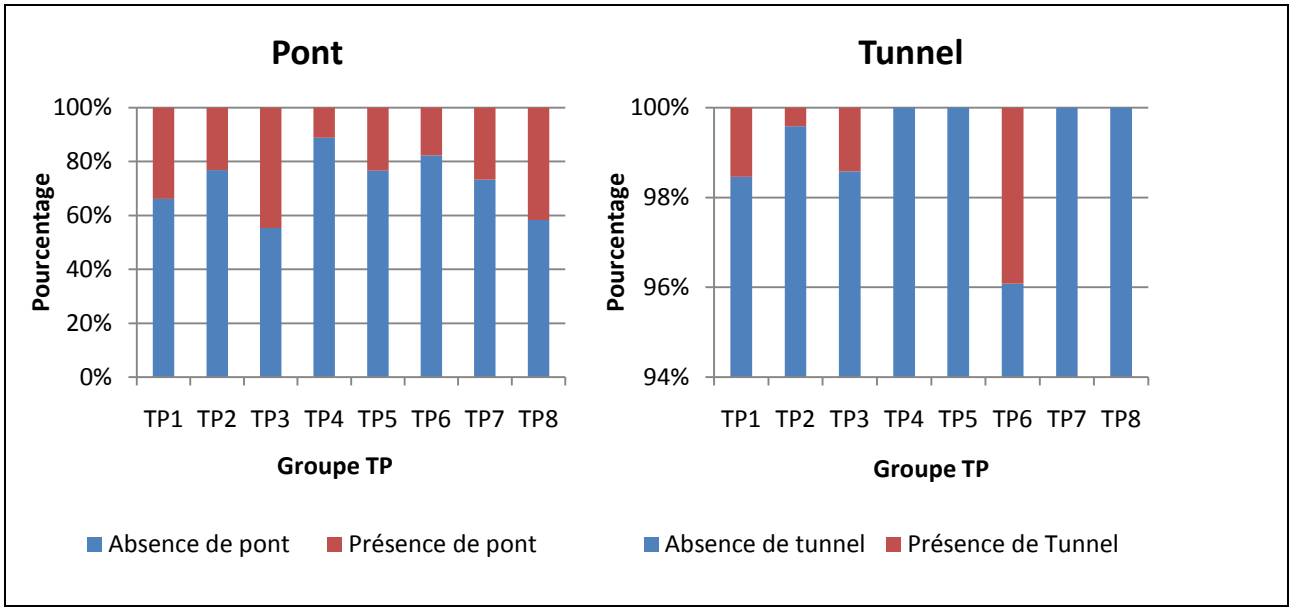

Figure 4.8 : Présence de pont et/ou de tunnel par groupe TP

Figure 4.9: Présence de pont et de tunnel

4.2.5 Vitesse autorisée

La figure 4.10 illustre la réparation des vitesses autorisées. D'après cette figure, on constate que sur 43% des segments routiers échantillonnés par le MTQ, la vitesse autorisée est de 100 km/h et que 30 % des tronçons ont une vitesse de 70 km/h. Ceci correspond au nombre élevé de tronçons de type autoroute.

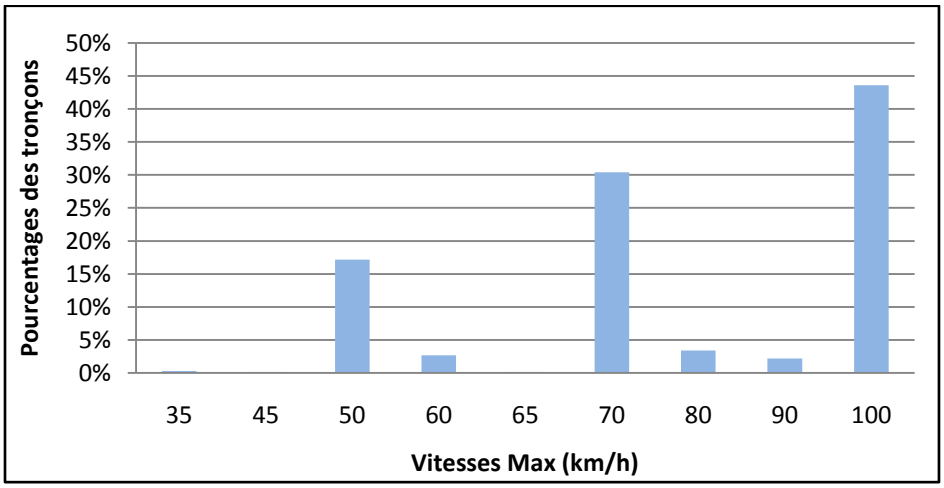

Figure 4.10 : La distribution des vitesses autorisées

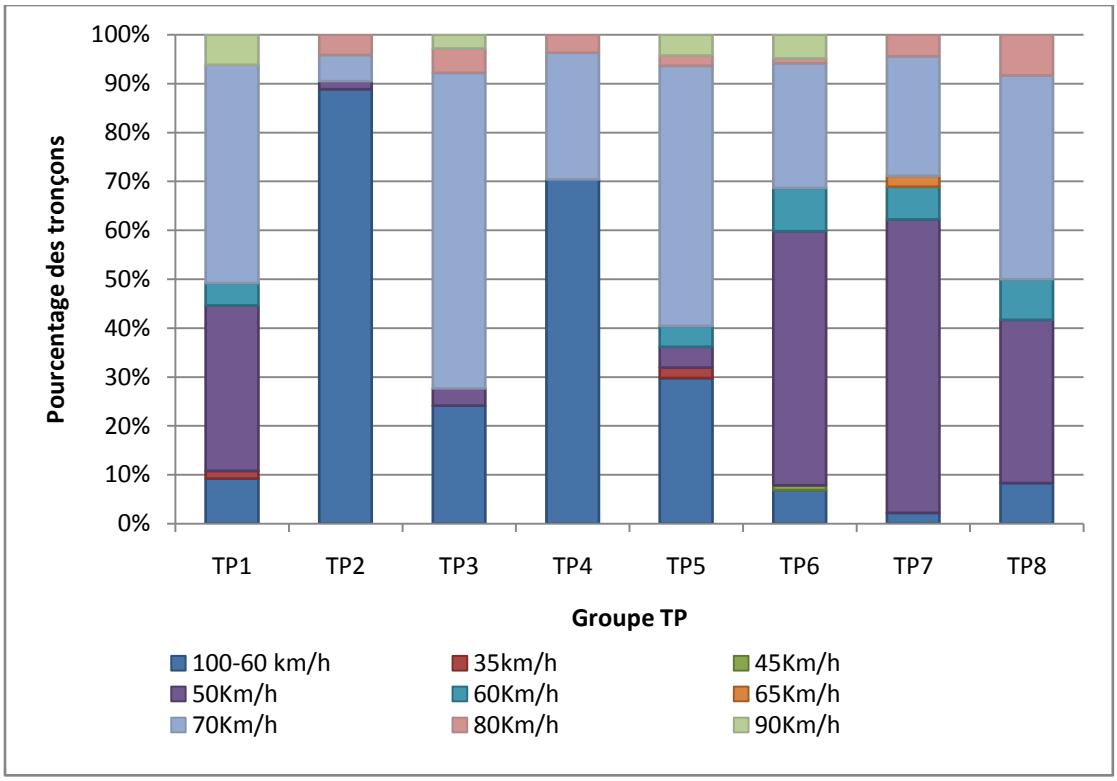

Figure 4.11: Les vitesses autorisées par Groupe TP

La figure 4.11 illustre le pourcentage de vitesse autorisée par groupe TP. À partir de cette figure, on a tiré les constats suivants :

- La vitesse autorisée pour plus de 70 % des tronçons de groupes TP2 et TP4, est limitée entre 100 et 60 km/h. Ceci correspond au fait que ces groupes incluent plus de tronçons de type « autoroute ».
- 65% des tronçons de groupe TP3 se caractérisent par une vitesse autorisée de 70Km/h.
- La vitesse autorisée pour plus de 52% des tronçons de groupes TP6 et TP7, est de 50 Km/h. Ce résultat est logique, car ces groupes TP se caractérisent par des tronçons de type « Boulevard/Avenue » et « Rue/ Route ».

À partir de ces constats, il est clair que le temps de parcours moyen est lié à la vitesse autorisée. En effet, les groupes caractérisés par une vitesse autorisée la plus élevée sont les groupes les plus fluides, hormis le groupe TP4.

La figure 4.12 illustre sous forme d'une carte thématique la vitesse affichée sur chaque segment du réseau routier de la grande région de Montréal. À partir de cette carte, on observe qu'il y a une domination de la couleur rouge et bleue, soit respectivement la vitesse 100km/h et la vitesse 70km/h. On peut noter également que plus on s'approche du centre-ville de Montréal, plus la vitesse autorisée diminue.

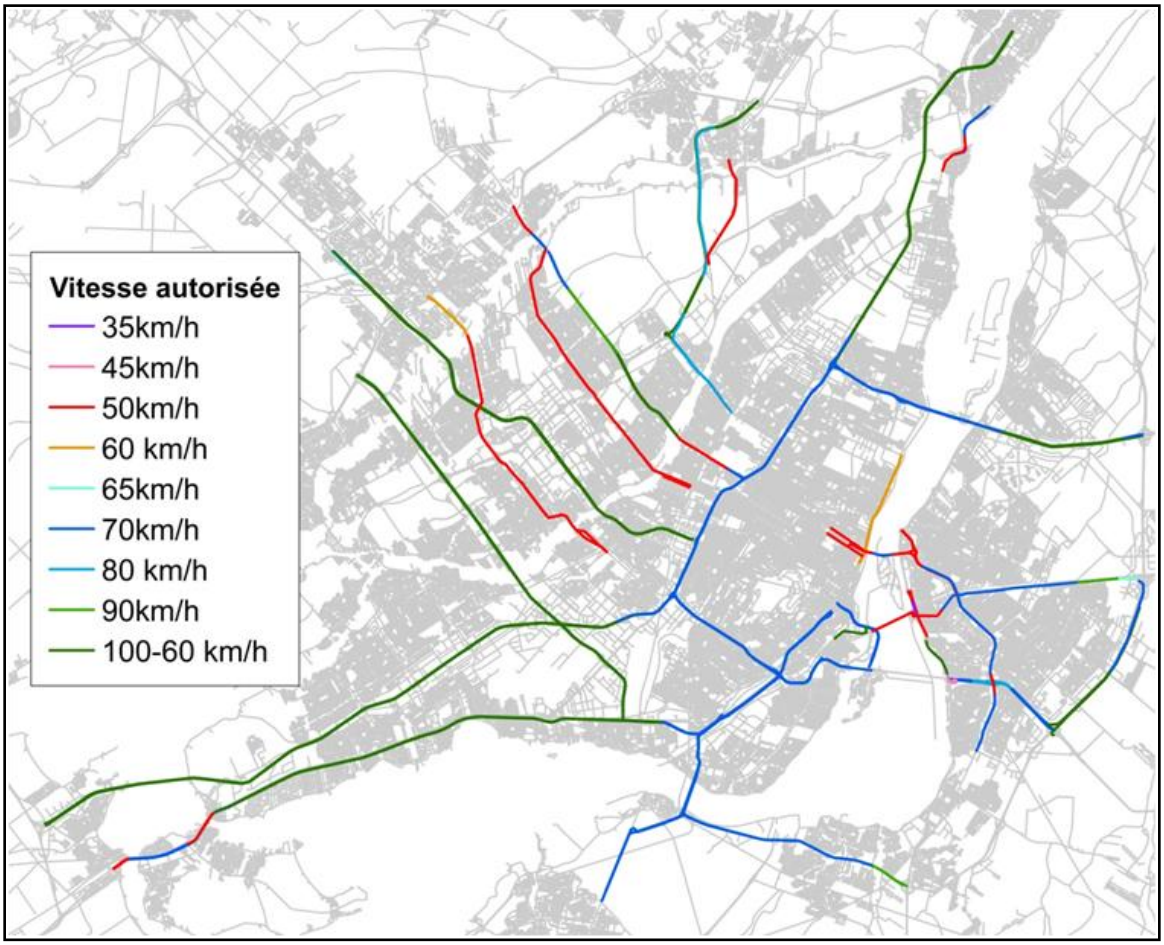

Figure 4.12 : Carte thématique de vitesses autorisées sur les segments routiers de la grande région de Montréal échantillonnés par le MTQ

4.2.6 Nombre de voies

La figure 4.13 représente la répartition des segments routiers par nombre de voies ainsi que la réparation de nombre de voies par groupe TP. En vertu de cette figure, on peut conclure que le réseau routier échantillonné par le MTQ se caractérise par des routes de 2 et 3 voies. De même pour chaque groupe TP, les pourcentages des tronçons de deux ou trois voies sont très

importants. Les groupes TP4 et TP8 sont les seuls groupes TP qui se caractérisent par l'absence des tronçons de type 4 voies. De plus, le pourcentage des tronçons à deux voies est plus élevé dans les groupes TP6, TP7 et TP8 que les autres groupes TP. Ceci peut justifier le non fluidité de la circulation sur les tronçons de ces groupes.

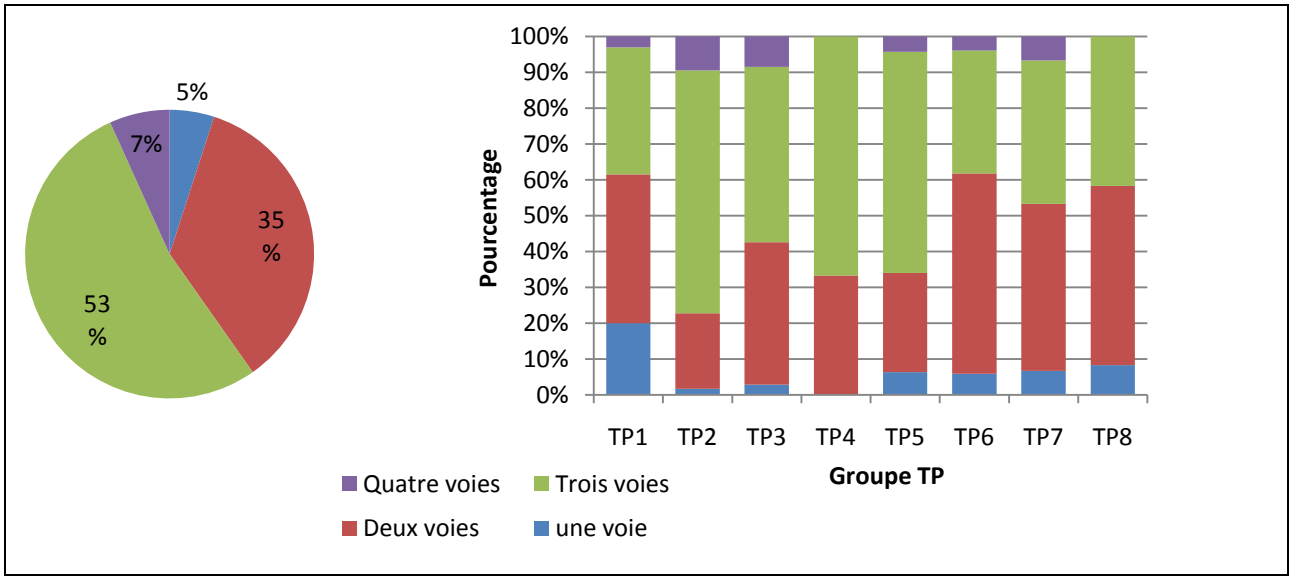

Figure 4.13 : Répartition de nombre de voies

La carte ci-dessous illustre le nombre de voies de chaque segment du réseau routier supérieur de la région de Montréal échantillonné par le MTQ. Sur certaines routes, le nombre de voies varie dans chaque sens, notamment l'autoroute 40 et l'autoroute 20.

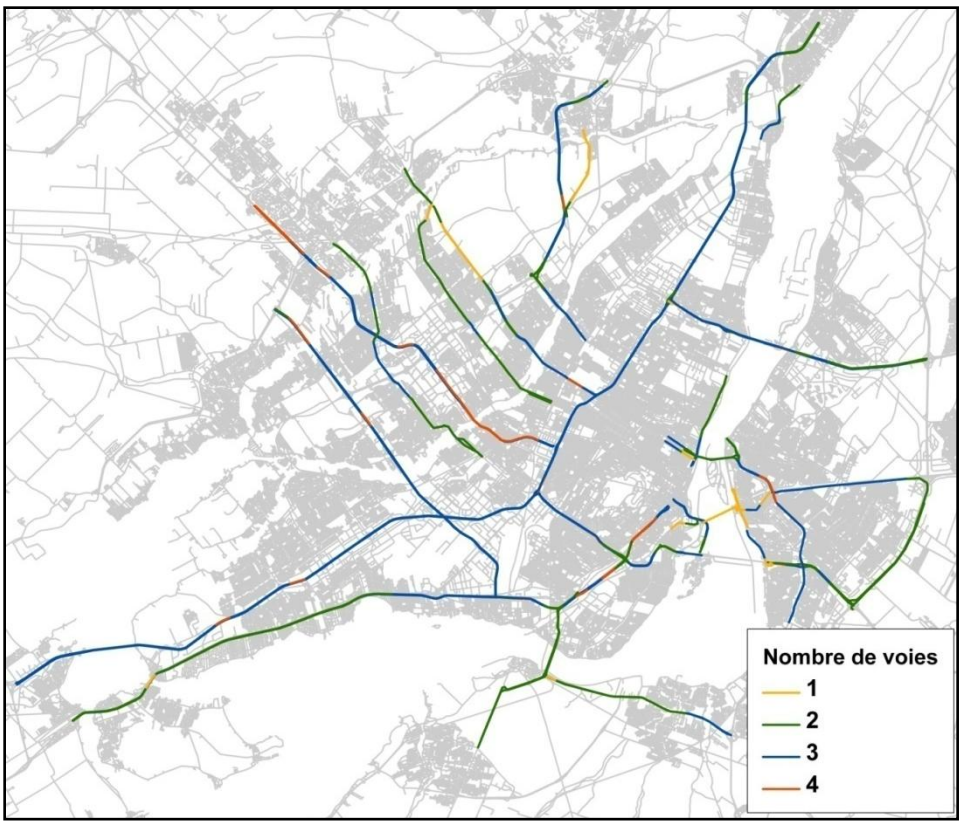

Figure 4.14 : Nombre de voies sur les segments du réseau routier de la région de Montréal échantillonnés par le MTQ

4.3 Conclusion

L'analyse de la base de données relationnelle a permis d'une part, de tracer le portrait global du réseau routier de la grande région de Montréal échantillonné par le MTQ, et d'autre part, d'étudier le lien entre l'état de la circulation et les facteurs physiques des segments routiers.

La première partie de ce chapitre a trait l'état de la circulation sur le réseau routier échantillonné par le MTQ. À partir des relevés de temps de parcours, Lousteau *et al.* (2009) ont regroupé les tronçons en huit groupes dont un seul groupe contient près de la moitié des tronçons échantillonnés, soit le groupe TP2. Ce groupe se caractérise par l'état de la circulation le plus fluide.

La deuxième partie, quant à elle, a permis d'exposer les caractéristiques physiques des segments routiers échantillonnés par le MTQ. Ensuite, une analyse descriptive de la base de données relationnelle qui croise les caractéristiques physiques avec les attributs liés au temps de parcours

a été effectuée. Cette analyse a mis en évidence l'existence d'une certaine analogie entre l'état de la circulation et les facteurs physiques, notamment le type de voie et la vitesse autorisée. En effet, ces facteurs physiques ont permis de distinguer entre les groupes fluides et les groupes non fluides et de mieux appréhender l'état de la circulation de chaque groupe TP. On constate aussi, à partir de cette analyse, que certains groupes ont les mêmes caractéristiques physiques, mais leurs états de la circulation sont différents, notamment les groupes TP2 et TP4. Ici, on peut noter, par analogie avec les caractéristiques physiques de groupe TP2, que le temps de parcours moyen élevé de groupe TP4 ne s'explique pas par les caractéristiques physiques, mais il est possible de l'appréhender par la présence de la congestion,

CHAPITRE 5 GROUPEMENT DES SEGMENTS DU RÉSEAU ROUTIER DE LA GRANDE RÉGION DE MONTRÉAL SELON LEURS CARACTÉRISTIQUES PHYSIQUES

Ce chapitre se divise en trois parties. D'abord, il s'agit de synthétiser et d'agréger les facteurs physiques des tronçons en appliquant la méthode de l'analyse factorielle des correspondances multiples (AFCM). Par la suite, la matrice de composantes principales issues de l'AFCM sera utilisée pour regrouper les tronçons similaires. Les groupes obtenus sont nommés groupes FP. Finalement, une analyse descriptive de ces groupes est effectuée afin de décrire leur comportement

5.1 Préparation des données

Après l'analyse descriptive du réseau routier échantillonné par le MTQ, il devient intéressant de mettre l'accent sur la structure des données qui seront utilisées que ce soit pour la segmentation ou pour la prédiction des groupes. On a recodé manuellement les variables continues pour qu'elles soient des variables catégorielles, soit le nombre de voies, le nombre d'intersections, le nombre d'entrées et le nombre de sorties. Pour les facteurs physiques «vitesse autorisée» et «nombre d'intersections», on a agrégé manuellement les modalités à faible effectif par rapport au nombre total de tronçons. Concernant la vitesse autorisée, la modalité «65km/h» est agrégée avec la modalité «70km/h», parce qu'un seul tronçon, parmi 681tronçons, a une vitesse autorisée de «65km/h». De plus, les modalités «35km/h» et «45km/h» ont été agrégées avec la modalité «50km/h» et qui sont recodées «moins de 50km/h». On a aussi agrégé les nombres d'intersections qui sont supérieurs à quatre et on les a recodés par «NI>=5». L'annexe E présente les nouveaux codes des différentes modalités.

5.2 Analyse factorielle des correspondances multiples

La méthode de l'analyse factorielle des correspondances multiples (AFCM) vise à étudier simultanément les relations entre les différentes modalités et permet d'agréger les données. Cette méthode est expliquée en annexe A. L'AFCM a permis d'identifier 32 axes résumant les informations fournies par la base de données des facteurs physiques. La figure 5.1 illustre les

valeurs propres des axes factoriels triées dans un ordre décroissant ainsi que le pourcentage cumulé d'inertie. Le nombre d'axes factoriels à retenir pour procéder au groupement des tronçons est égal au numéro du premier axe factoriel ayant une valeur propre supérieure à la moyenne des valeurs propres qui est égale à 0,067 (Jubiter, 1989; Saporta, 2006). Selon ce principe, les douze premiers axes sont retenus et qui représentent 63% de d'inertie total.

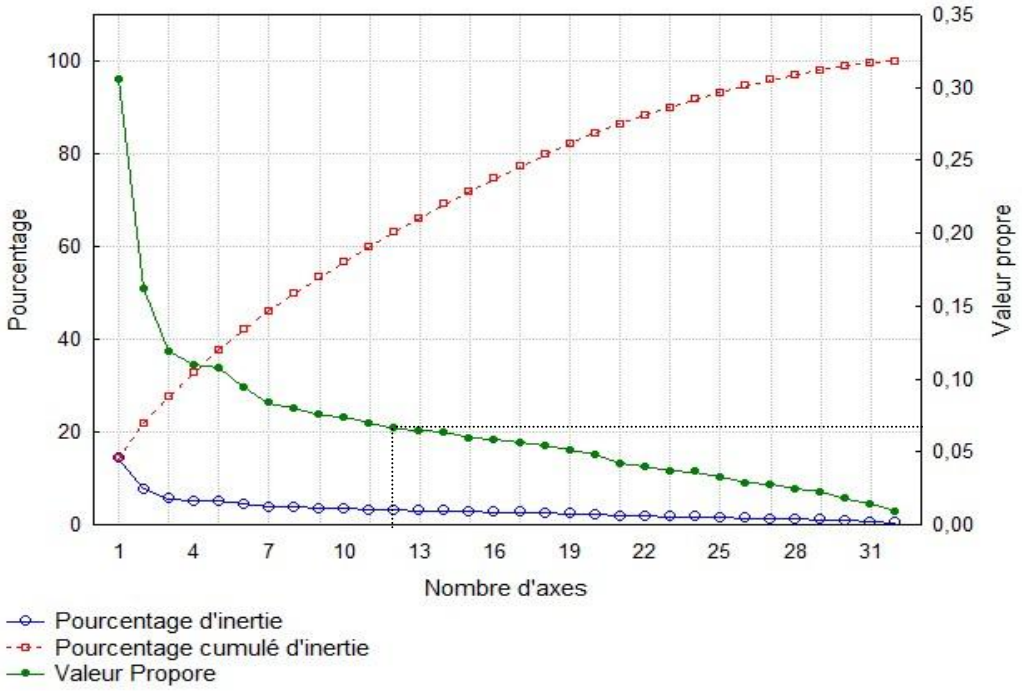

Figure 5.1: La valeur propre, le pourcentage d'inertie, pourcentage cumulé expliquée par les axes factoriels

Une fois que le nombre d'axes factoriels est déterminé, on s'intéresse à étudier simultanément les relations entre les différentes modalités des facteurs physiques. Ceci est présenté par la figure suivante. Les cellules colorées correspondent aux modalités qui ont une forte contribution à la construction des axes factoriels, c'est-à-dire, celle ayant une valeur de contribution supérieure à un certain seuil (voir l'annexe A). Les cellules colorées en rouge et en vert représentent respectivement le signe positif et négatif de la coordonnée de la modalité j sur l'axe factoriel r. Ceci permet de savoir s'il s'agit d'une association positive ou négative; si deux modalités ont le même signe, alors il s'agit d'un lien positif, sinon, le lien entre ces deux modalités en question est négatif. Les cellules non colorées correspondent à une faible contribution à la construction des axes factoriels.

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8	Axe 9	Axe 10	Axe 11	Axe 12
Nbr_sorties:NS0												
Nbr_sorties:NS1												
Nbr_sorties:NS2												
Nbr_entrées:NE0												
Nbr_entrées:NE1												
Nbr_entrées:NE2												
Nbr_intersections:NI0												
Nbr_intersections:NI1												
Nbr_intersections:NI2												
Nbr_intersections:NI3												
Nbr_intersections:NI4												
Nbr_intersections:NI>=5												
sens_HC/VC:VM												
sens_HC/VC:HM												
Type_voie:AU												
Type_voie:BV/AV												
Type_voie:Rue/Route												
Acc_D:Acc_D_Oui												
Acc_D:Acc_D_Non												
Tr_D:Tr_D_Non												
Tr_D:Tr_D_Oui												
Type_Barr_D:Absence de barrière_D												
Type_Barr_D:Barrière béton_D												
Type_Barr_D:Barrière métallique_D												
Type_Barr_D:Mûr_D												
Acc_G:Acc_G_Oui												
Acc_G:Acc_G_Non												
Tr_G:Tr_G_Non												
Tr_G:Tr_G_Oui												
Type_Barr_G:Absence de barrière_G												
Type_Barr_G:Barrière béton_G												
Type_Barr_G:Barrière métallique_G												
Type_Barr_G:Mûr_G												
sur_pont:Sur_p=Oui												
sur_pont:Sur_p=Non												
Tunnel:Tun=Non												
Tunnel:Tun=Oui												
Nbr_voies:NV1												
Nbr_voies:NV2												
Nbr_voies:NV3												
Nbr_voies:NV4												
Vitesse_aut:100-60km/h												
Vitesse_aut:90Km/h												
Vitesse_aut:80Km/h												
Vitesse_aut:70Km/h												
Vitesse_aut:60km/h												
Vitesse_aut:moins de 50km/h												

Figure 5.2: Contribution des modalités

Si on examine le premier axe factoriel, on s'aperçoit qu'il y a une association positive entre la présence d'intersection, les types de voies «boulevard/avenue» et «Rue/avenue», les vitesses

autorisées «60km/h» et «moins de 50 km/h». Ces modalités ont un lien négatif avec les modalités «autoroute» et «100-60km/h».

5.3 Groupement des tronçons

Après la simplification des données par la méthode d'analyse factorielle des correspondances multiples (AFCM), on va procéder au groupement des tronçons similaires au moyen de l'algorithme de segmentation présenté dans l'annexe B, soit la méthode de k-moyennes. Cet algorithme permet de grouper les observations selon les composantes principales issues de la méthode AFCM, soit les coordonnées des tronçons sur les axes factoriels.

Afin que la différence de l'amplitude des composantes principales n'influence pas sur le regroupement, les données ont été normalisées par le test Z avant d'appliquer l'algorithme des k-moyennes. En effet, les composantes principales ayant l'amplitude la plus grande peuvent avoir une forte influence sur les résultats. Supposons que x_i et x_i' sont respectivement la valeur originale de l'ensemble des données de la composante principale i, noté X_i , et la valeur normalisée, alors:

$$x_i' = \frac{x_i - \text{moyenne}(X_i)}{\text{écart type } (X_i)}$$

D'après Tuffèry (2010), le nombre maximal d'itérations préconisé pour l'algorithme des k-moyennes est supérieur ou égal à 10. Dans cette étude, on a fixé à 50 le nombre maximal d'itérations. Vu que le choix des centres initiaux est aléatoire, on a appliqué l'algorithme des k-moyennes 100 fois (nombre d'essais est choisi aléatoirement). L'essai à choisir est celle qui a la plus grande valeur de R-carré. L'algorithme des k-moyennes et la valeur de R-carré sont expliqués dans l'annexe B.

5.3.1 Choix du nombre de groupes FP

Tel que mentionné dans le partie méthodologie et aussi dans l'annexe B, la méthode des k-moyennes nécessite le choix du nombre de groupes. Pour ce faire, Cet algorithme a été appliqué plusieurs fois sur la base de données des composantes principales normalisées, en augmentant successivement le nombre de groupes, noté k. Pour choisir le nombre de groupes (k), on s'est appuyé sur deux critères, soit la valeur de R-carré et la valeur pseudo-F. Ces critères sont définis dans l'annexe B.

La figure suivante illustre, pour chaque essai, la valeur de dispersion intragroupe (w) et intergroupe (BSS) ainsi que la valeur de R-carré. Il faut note que, plus la valeur de R-carré s'approche de 1, plus la qualité de la classification est meilleure (Giudici, 2003). Dans notre étude, on ne cherche pas à avoir R-carré égale à 1, sinon chaque tronçon représentera un groupe. D'après ce critère, le nombre de groupes suggéré est supérieur ou égal à 10 (R=0,51), c'est à partir de ce valeur que la dispersion intergroupe (BSS) est supérieure à la dispersion intragroupe (W).

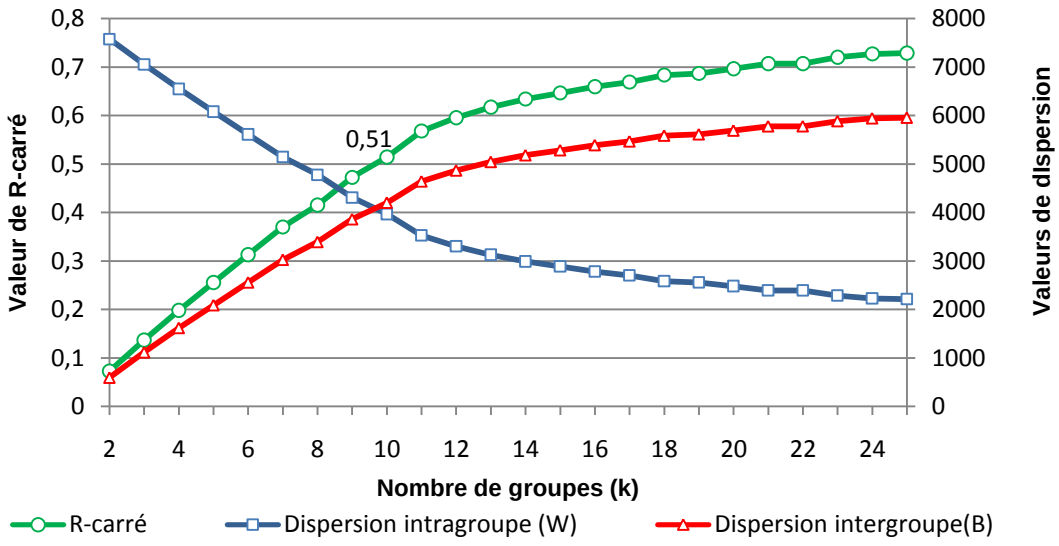

Figure 5.3: Dispersion intragroupe (w) et intergroupe (BSS)

La figure 5.4 illustre la valeur de pseudo-F proposée par Clinski et Harabasz (1974) comme un critère de choix du nombre de groupes. En s'appuyant sur cette figure, le nombre de groupes suggéré est égal à 13, ceci correspond à la valeur maximale de pseudo-F.

À partir de ces deux critères de mesure de la qualité de regroupement, le nombre de groupes suggéré est entre 10 et 13. Si on observe la courbe de perte de dispersion intragroupe (Diff W) en passant de k à k+1 groupes, on constate que le passage de 11 à 12 groupes entraine une perte sensiblement plus forte de la dispersion intragroupe (W) que le passage de 12 à 13 groupes. Cependant, le passage de 10 à 11 groupes occasionne une augmentation de la dispersion intragroupe. Ainsi, le nombre de groupes FP à retenir dans cette étude, est égal à 12.

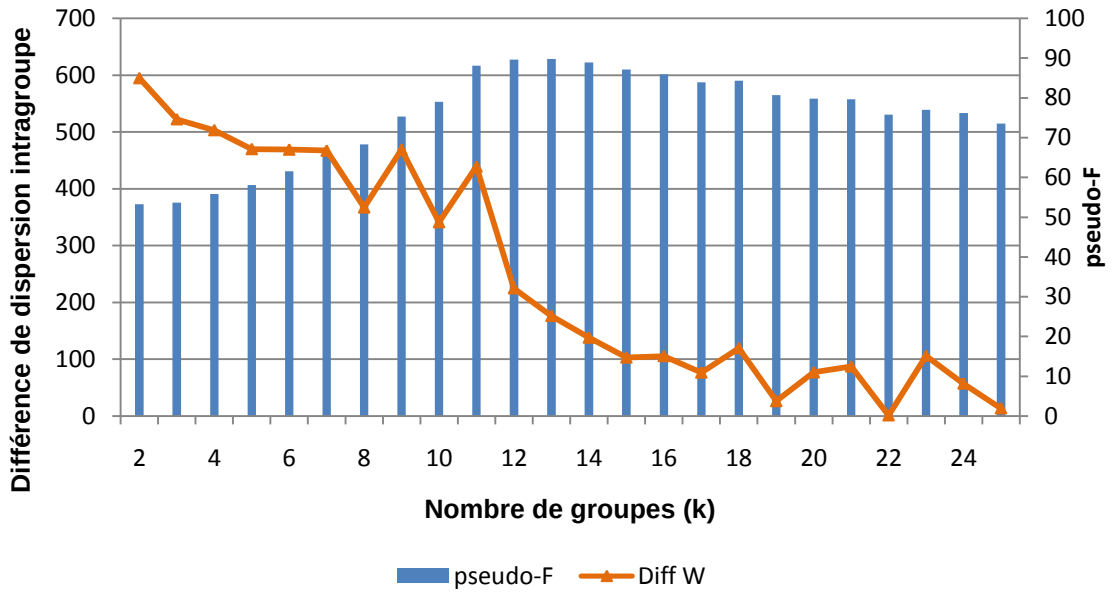

Figure 5.4: valeur de pseudo-F et la différence de dispersion intragroupe (Diff W)

Après le choix du nombre de groupes, on va analyser les groupes FP obtenus. La figure 5.5 présente la répartition des tronçons par groupe FP. On s'aperçoit que les tronçons ne sont pas repartis d'une manière égalitaire sur les douze groupes. Le nombre de tronçons dans le groupe FP11 est plus élevé que celui des autres groupes, soit plus de 30% de nombre total de tronçons. Le groupe FP12, quant à lui, regroupe 0,2% de nombre total de tronçons. La moitié des groupes ont presque le même nombre de tronçons, soit les groupes FP 4, 5, 7, 8, 9 et 10.

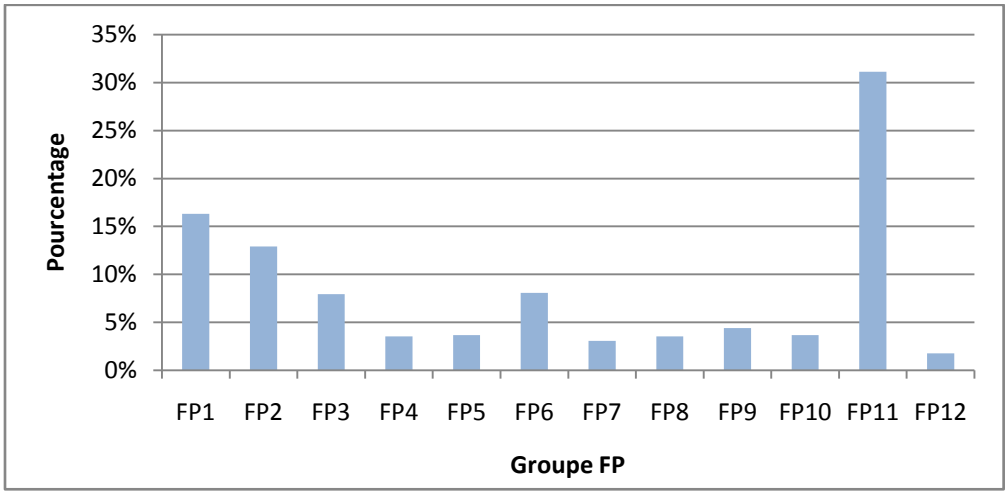

Figure 5.5: Répartition des tronçons par groupe FP

La figure suivante illustre les valeurs de dispersion intragroupe. Bien que le groupe FP6 et F11 n'aient pas le même nombre de tronçons, mais leurs valeurs dispersion intragroupe sont très proches. De plus, le degré de similarité entre les tronçons de groupe FP7 est plus élevé que celui de groupe 1. La courbe de l'indice de dispersion intragroupe (IW) donne des indications un peu contradictoires avec ceux donnés par l'histogramme de dispersion intragroupe (W), car elle présente la valeur W normalisée par le nombre de tronçons dans chaque groupe FP. À partir de la courbe IW, on constate que les groupes FP1et 11 se caractérisent par un degré de similarité le plus élevé par rapport aux autres groupes; leurs valeurs IW sont les plus faibles.

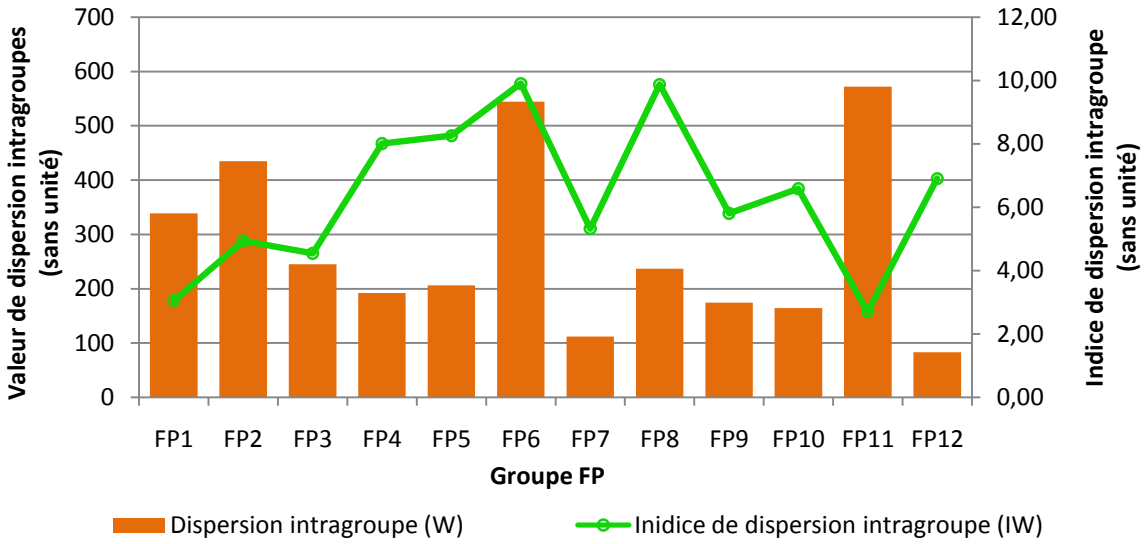

Figure 5.6: Valeur de dispersion intragroupe et indice de dispersion intragroupe

5.3.2 Analyse de groupes FP

Une fois que le nombre de groupes FP a été choisi et les segments ont été regroupés à partir de la base de données des composantes principales issues de la méthode de l'analyse des correspondances multiples, on procède, dans un but exploratoire, à analyser les caractéristiques physiques des ces groupes. En d'autres termes, il s'agit d'étudier la répartition des tronçons de chaque groupe FP par rapport aux modalités de chaque facteur physique.

Le tableau suivant présente les caractéristiques physiques des tronçons dans chaque groupe FP.

Tableau 5.1: Proportion des tronçons dans chaque groupe FP par rapport aux modalités de chaque facteur physique

Attributs	Modalités	FP1	FP2	FP3	FP4	FP5	FP6	FP7	FP8	FP9	FP10	FP11	FP12
Nbr_sorties	NS0	52%	59%	59%	83%	72%	55%	33%	83%	97%	68%	57%	50%
	NS1	38%	32%	39%	17%	20%	27%	62%	4%	3%	32%	42%	33%
	NS2	10%	9%	2%	0%	8%	18%	5%	13%	0%	0%	1%	17%
Nbr_entrées	NE0	41%	65%	61%	92%	72%	53%	38%	92%	90%	80%	56%	58%
	NE1	57%	34%	31%	8%	24%	16%	52%	8%	10%	20%	42%	33%
	NE2	2%	1%	7%	0%	4%	31%	10%	0%	0%	0%	1%	8%
Nbr_intersections	NI0	99%	95%	100%	0%	0%	44%	95%	92%	0%	0%	100%	100%
	NI1	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%
	NI2	1%	2%	0%	0%	0%	56%	0%	8%	0%	0%	0%	0%
	NI3	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%
	NI4	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%
	NI>=5	0%	2%	0%	0%	0%	0%	5%	0%	100%	0%	0%	0%
sens_HC/VC	HM	48%	40%	85%	50%	32%	64%	71%	58%	43%	48%	49%	50%
	VM	52%	60%	15%	50%	68%	36%	29%	42%	57%	52%	51%	50%
Type_voie	AU	88%	84%	87%	13%	0%	16%	95%	17%	0%	0%	93%	100%
	BV/AV	12%	14%	13%	54%	64%	65%	5%	71%	67%	76%	6%	0%
	Rue/Route	0%	2%	0%	33%	36%	18%	0%	13%	33%	24%	1%	0%
Acc_D	Acc_D_Non	80%	35%	11%	79%	44%	67%	33%	0%	77%	88%	8%	100%
	Acc_D_Oui	20%	65%	89%	21%	56%	33%	67%	100%	23%	12%	92%	0%
TR_D	Tr_D_Non	97%	100%	98%	25%	64%	35%	95%	100%	27%	24%	96%	100%
	Tr_D_Oui	3%	0%	2%	75%	36%	65%	5%	0%	73%	76%	4%	0%
Type_Barr_D	Absence de barrière_D	5%	30%	59%	92%	80%	80%	19%	71%	90%	92%	75%	0%
	Barrière béton_D	88%	0%	17%	0%	4%	11%	19%	17%	3%	4%	15%	0%
	Barrière métallique_D	1%	70%	20%	4%	16%	7%	62%	13%	7%	4%	9%	0%
	Mur_D	5%	0%	4%	4%	0%	2%	0%	0%	0%	0%	1%	100%
Acc_G	Acc_G_Non	98%	75%	54%	100%	100%	100%	71%	96%	100%	100%	39%	100%
	Acc_G_Oui	2%	25%	46%	0%	0%	0%	29%	4%	0%	0%	61%	0%
Tr_G	Tr_G_Non	100%	97%	98%	63%	64%	49%	100%	92%	67%	52%	99%	100%
	Tr_G_Oui	0%	3%	2%	38%	36%	51%	0%	8%	33%	48%	1%	0%
Type_Barr_G	Absence de barrière_G	5%	7%	43%	92%	96%	78%	24%	42%	97%	96%	30%	0%
	Barrière béton_G	95%	23%	44%	8%	4%	20%	62%	25%	0%	0%	66%	33%
	Barrière métallique_G	0%	70%	9%	0%	0%	2%	0%	4%	3%	0%	1%	0%
	Mur_G	0%	0%	4%	0%	0%	0%	14%	29%	0%	4%	3%	67%
sur_pont	Sur_p=Non	25%	58%	54%	83%	84%	82%	62%	100%	97%	84%	93%	100%
	Sur_p=Oui	75%	42%	46%	17%	16%	18%	38%	0%	3%	16%	7%	0%
Tunnel	Tun=Non	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	33%
	Tun=Oui	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	67%
Nbr_voies	NV1	3%	6%	0%	0%	0%	4%	0%	71%	17%	8%	0%	0%
	NV2	23%	49%	37%	58%	72%	45%	29%	4%	70%	64%	24%	8%
	NV3	74%	45%	0%	38%	28%	44%	57%	25%	10%	24%	76%	83%
	NV4	1%	0%	63%	4%	0%	7%	14%	0%	3%	4%	0%	8%
Vittesse_aut	100-60km/h	20%	65%	50%	0%	0%	13%	0%	0%	3%	0%	86%	0%
	90Km/h	0%	0%	0%	0%	0%	0%	0%	63%	0%	0%	0%	0%
	80Km/h	0%	0%	0%	8%	0%	0%	100%	0%	0%	0%	0%	0%
	70Km/h	72%	23%	48%	13%	48%	29%	0%	4%	27%	16%	12%	100%
	60km/h	0%	2%	0%	13%	20%	5%	0%	0%	3%	12%	1%	0%
	moins de 50km/h	8%	10%	2%	67%	32%	53%	0%	33%	67%	72%	1%	0%

Interprétation du tableau :

Groupe FP1 :

- Ces tronçons sont généralement des autoroutes à trois voies ayant une vitesse autorisée 70km/h.
- Ce groupe se caractérise par l'absence d'accotement et de trottoir, et par la présence d'une barrière béton.

Groupe FP2 :

- Ce groupe se caractérise par des tronçons ayant un type de voie « autoroutes » à 2 et 3 voies et ayant une vitesse autorisée égale à « 100-60km/h ».
- Il se caractérise par la présence d'accotement à droite, par une barrière métallique à droite et à gauche, et par l'absence de trottoir.

Groupe FP3 :

- Ce groupe se caractérise par des voies de type autoroute à quatre voies, par des vitesses autorisées « 70km/h » et « 100-60km/h », et par l'absence de barrière à droite

Groupe FP4 :

- La majorité des tronçons sont des boulevards et des avenues à deux voies, et se caractérisent par une vitesse autorisée moins de 50km/h et par la présence de quatre intersections
- Ce groupe se caractérise aussi par la présence de trottoir à droite, par l'absence d'accotement et de barrière.

Groupe FP5 :

- Ce groupe possède la plus grande proportion des tronçons ayant un sens de circulation vers le centre de ville de Montréal, de type « boulevard/avenue » à deux voies. La vitesse autorisée sur les tronçons de ce groupe est moins de 70km/h.
- Il se caractérise par l'absence de trottoir, d'accotement et de barrière.

Groupe FP6 :

- En majorité, les tronçons de ce groupe sont des boulevards et des avenues à deux et trois voies.
- Ce groupe se caractérise par la présence de deux intersections, par une vitesse autorisée moins de 50km/h et par un sens de circulation «hors centre-ville de Montréal».

Groupe FP7 :

- La plupart des tronçons sont des autoroutes à trois voies et qui se caractérisent par la présence d'une entrée et d'une sortie, et par une vitesse autorisée égale à 80km/h.
- Ce groupe se caractérise généralement par un sens de circulation « hors centre-ville de Montréal », par la présence d'accotement, par une barrière métallique à droite et par une barrière béton à gauche.

Groupe FP8 :

- Ce groupe se compose par des tronçons de type «boulevard/avenue » à une voie
- Il se caractérise par une vitesse autorisée égale à 90km/h et par la présence d'accotement à droite.

Groupe FP9 :

- Ce groupe possède la grande proportion des tronçons ayant cinq intersections .Ces tronçons sont des boulevards ou des avenues à deux voies
- Il se caractérise par la présence de trottoir à droite et par une vitesse autorisée moins de 50km/h.

Groupe FP10 :

- Le groupe 10, quant à lui, se caractérise par la présence de deux d'intersections, type de voies «boulevard/avenue», par l'absence de barrière et par une vitesse autorisée moins de 50km/h.

Groupe FP11 :

- En majorité, les tronçons de ce groupe sont des autoroutes à trois voies.

- Ils se caractérisent par une vitesse autorisée «100-60km/h», par la présence d'accotement, d'entrées et de sorties, par l'absence de barrière à droite, et par la présence de barrière béton à gauche.

Groupe FP12 :

- Tous les tronçons de ce groupe sont des autoroutes ayant une vitesse autorisée égale à 70km/h.
- La plupart des tronçons ont trois voies et se caractérisent par la présence d'un mur à droite et à gauche, et par la présence de tunnel.

5.4 Conclusion

Dans ce chapitre, on a essayé de mettre en évidence les particularités des tronçons échantillonnés par le MTQ. En effet, dans cette partie de travail, on a montré que cet échantillon des tronçons est hétérogène; douze groupes ont été obtenus. Cependant, il y a des groupes qui ont plus d'effectif que les autres groupes. En effet, le groupe FP11 qui regroupe plus de 30% des tronçons, se caractérise par un type de voie «autoroute» à trois voies, vitesse autorisée «100-60km /h» et par l'absence d'intersections.

CHAPITRE 6 CONFRANTATION DES RESULTATS ET PREDICTION DE L'APPARTENANCE D'UN TRONÇON À UN GROUPE TP

Une fois que le groupement des tronçons selon les facteurs physiques est achevé, ce présent chapitre traite le lien entre ce nouveau groupement et celui bâti sur les temps de parcours. En outre, dans ce chapitre, on se question si les facteurs physiques permettent d'expliquer l'état de la circulation. De plus, il s'agit de déterminer les facteurs physiques significatifs permettant d'affecter les tronçons aux groupes TP.

6.1 Confrontation des résultats (groupes FP vs groupes TP)

Dans cette première partie, il s'agit tout d'abord de faire une analyse de profil des groupes TP par rapport aux groupes FP. Cette analyse permet, entre autres, d'avoir une idée préliminaire sur la relation entre les caractéristiques physiques des segments routiers et l'état de la circulation. Par la suite, une analyse bivariée des groupes TP et FP au moyen de la méthode de khi-deux sera effectuée pour tester l'indépendance entre ces groupes.

Le tableau suivant représente simultanément la répartition des tronçons par groupe TP et FP. À partir de ce tableau, on observe qu'il y a des tronçons qui appartiennent au même groupe bâti sur les temps de parcours, mais ils n'ont pas les mêmes caractéristiques physiques, telles que les tronçons de groupe TP1. En outre, il y a des tronçons qui ont les mêmes caractéristiques physiques, mais ils n'ont pas le même état de la circulation, tel que les tronçons des groupes TP2 et TP4 qui appartiennent au groupe FP11. À partir de ces constats, il semble que l'état de la circulation et les facteurs physiques sont indépendants.

Si on examine les modalités deux à deux, on s'aperçoit le lien entre les facteurs physiques des tronçons de groupe FP11 et l'état de la circulation des tronçons de groupe TP2; plus de 20% de nombre total de tronçons appartiennent à ces groupes. Ces tronçons sont les plus fluides et qui sont des autoroutes à trois voies et ont une vitesse autorisée égale à « 100-60km/h». De même, on constate que la plupart des tronçons de groupes TP3 ont les caractéristiques de groupe FP1. Rappelons que le groupe TP3 est le deuxième groupe TP le plus fluide et que le groupe FP1 se caractérise par un type de voie «autoroute» et par une vitesse autorisée «70km/h». Cela met en

avant l'existence d'un lien entre l'état de la circulation et les facteurs physiques de certains groupes TP et FP.

Tableau 6.1: Tri croisé à l'aide de la variable « Groupe TP »

	FP1	FP2	FP3	FP4	FP5	FP6	FP7	FP8	FP9	FP10	FP11	FP12	Total
TP1	11	7	4	3	2	13	0	10	4	5	4	2	65
TP2	23	37	23	1	0	5	10	0	0	1	141	1	242
TP3	48	16	23	1	5	4	7	4	1	0	28	4	141
TP4	1	7	0	0	1	0	1	0	0	0	17	0	27
TP5	12	7	2	1	2	2	1	2	0	2	16	0	47
TP6	8	6	2	9	9	21	0	8	19	11	5	4	102
TP7	7	5	0	8	5	8	1	0	5	5	1	0	45
TP8	1	3	0	1	1	2	1	0	1	1	0	1	12
Total	111	88	54	24	25	55	21	24	30	25	212	12	681

Afin de tester l'indépendance entre les groupes TP et les groupes FP, le test de khi-deux (d'écart d'indépendance) a été utilisé. Ce test permet de vérifier, partant d'une hypothèse et d'un risque supposé au départ, l'indépendance entre les groupes TP et les groupes FP. D'un point de vue mathématique, deux variables X et Y sont indépendantes, si la répartition des modalités de X est la même pour toutes les modalités de Y. Deux hypothèses découlent de la question d'indépendance, à savoir :

H0 : Les deux variables X et Y sont indépendantes

H1 : Les deux variables X et Y ne sont pas indépendantes

La formule de khi-deux est la suivante :

$$\chi^2 = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(n_{kl} - \frac{n_k * n_l}{n})^2}{\frac{n_k * n_l}{n}}$$

Où

L est le nombre de modalités de la variable X;

K est le nombre de modalités de la variable Y;

n_l est la somme marginale de variables de X

n_k est la somme marginale de variables Y

n_{kl} est le nombre des éléments dans la $k^{éme}$ modalité de la variable Y et la $l^{éme}$ modalité de la variable X, avec l=1,.., L et k=1,..,K.

Le test de khi-deux qui suit une distribution asymétrique dont la forme dépend du nombre de degrés de libertés (DDL) qui s'écrit:

$$DDL = (K-1)*(L-1)$$

Où K= nombre de modalités de groupes TP et L= nombre de modalités de groupes FP.

L'hypothèse nulle (H_0: Les groupes TP *X* et les groupes FP *Y* sont indépendants) est rejetée lorsque la valeur du khi-deux calculée est supérieure à celle référencée à partir de la table de khi-deux pour DDL degrés de liberté et pour un risque d'erreur choisi α.

Tableau 6.2 : Tableau de khi-deux pour le cas des groupes TP et des groupes FP

	FP1	FP2	FP3	FP4	FP5	FP6	FP7	FP8	FP9	FP10	FP11	FP12	Total
TP1	0,0	0,2	0,3	0,2	0,1	11,4	2,0	25,9	0,5	2,9	13,0	0,6	57,2
TP2	6,9	1,1	0,8	6,7	8,9	10,8	0,9	8,5	10,7	7,0	57,2	2,5	121,8
TP3	27,2	0,3	12,5	3,2	0,0	4,8	1,6	0,2	4,4	5,2	5,8	0,9	66,0
TP4	2,6	3,5	2,1	1,0	0,0	2,2	0,0	1,0	1,2	1,0	8,8	0,5	23,9
TP5	2,5	0,1	0,8	0,3	0,0	0,9	0,1	0,1	2,1	0,0	0,1	0,8	7,8
TP6	4,5	3,9	4,6	8,1	7,4	19,8	3,2	5,4	46,8	14,1	22,5	2,7	142,9
TP7	0,0	0,1	3,6	25,9	6,8	5,2	0,1	1,6	4,6	6,8	12,1	0,8	67,6
TP8	0,5	1,4	1,0	0,8	0,7	1,1	1,1	0,4	0,4	0,7	3,7	2,9	14,7
Total	44,2	10,6	25,6	46,1	23,9	56,2	9,0	43,1	70,6	37,6	123,3	11,8	501,9

Appliquant la formule décrite ci-haut, la valeur du khi-deux observée est égale à 501,9 (voir tableau 6.2). Toutefois, la valeur du khi-deux théorique avec un nombre de degrés de libertés DDL= 77 et avec un risque d'erreur α=5%, est égale à 98,7. Donc, on rejette l'hypothèse nulle. Il y a une liaison entre les groupements faits selon les facteurs physiques et ceux bâtis sur les temps de parcours.

À partir de ce tableau, si on s'intéresse aux modalités ayant des contributions au khi-deux les plus élevés, on s'aperçoit qu'il y a une forte liaison entre les caractéristiques physiques des tronçons de groupe FP11 et l'état de la circulation des tronçons de groupe TP2. De plus, il y a un lien entre le temps de parcours moyen des tronçons de groupe TP6 et les caractéristiques physiques des

tronçons de groupe FP9. Cela met en avant que l'état de la circulation des groupes TP2 et TP6 peut être expliqué respectivement par les caractéristiques physiques des groupes FP11 et FP9.

6.2 Arbre de décision

Dans cette partie, l'application de l'arbre de décision permettra d'une part d'expliquer l'appartenance des tronçons aux groupes TP à partir des facteurs physiques et, d'autre part, de construire un modèle prédictif. Pour ce faire, l'algorithme C4.5, développé par Quinlan (1993), est utilisé. Cet algorithme sera appliqué à la base de données des facteurs physiques, permettant par la suite d'extraire les règles de décision. On rappelle que tous les facteurs physiques sont des variables catégorielles et que le nombre de tronçons est égal à 681.

Dans un premier essai (aucun critère d'arrêt de l'expansion de l'arbre de décision n'a été fixé), l'application de l'algorithme C4.5 a permis d'avoir un arbre de décision très touffu et de classer correctement tous les tronçons dans les groupes TP. Cet arbre de décision, nommé T_0, semble performant, mais il ne permet pas de décrire la réalité (on ne peut pas le généraliser). En effet, les règles de décisions sont vérifiées par un nombre très faible de tronçons par rapport au nombre total et l'arbre de décision s'ajuste bien aux données d'apprentissage (Saporta, 2006). Alors, il faut trouver un arbre le plus petit possible (sous-arbre de T_0) et le plus performant possible. La question qui se pose comment trouver le compromis entre la performance et la taille de l'arbre. Pour y arriver, on a proposé de diminuer en cascade la taille de l'arbre de décision, et de calculer le taux de précision en apprentissage de chaque sous-arbre obtenu, noté T_i, au moyen de la méthode de resubstitution (cette méthode est définie dans l'annexe D).

La diminution de la taille de l'arbre de décision se repose sur le critère d'arrêt de la segmentation de nœud par l'algorithme C4.5: si le nombre de tronçons dans un nœud est inférieur à un effectif minimal (E_{min}), alors la segmentation de ce nœud s'arrête à ce niveau. Il faut noter que faire augmenter la valeur E_{min} entraine la diminution de la taille de l'arbre de décision.

Le sous- arbre T_i qui sera choisi est celui qui a la meilleure capacité prédictive. Pour ce faire, la méthode de validation croisée (S sous-ensembles) a été utilisée. Cette méthode est définie dans l'annexe D. D'après Saporta (2006), S=10 est souvent préconisé. La limite de cette méthode de validation est que les données seront réparties d'une manière aléatoire sur les 10 sous-ensembles. Dans ce cas, on a refait ce test de validation quinze fois (le nombre d'essais est choisi

arbitrairement) et le taux de précision selon cette méthode est égal à la moyenne des taux de précision calculés dans tous les essais.

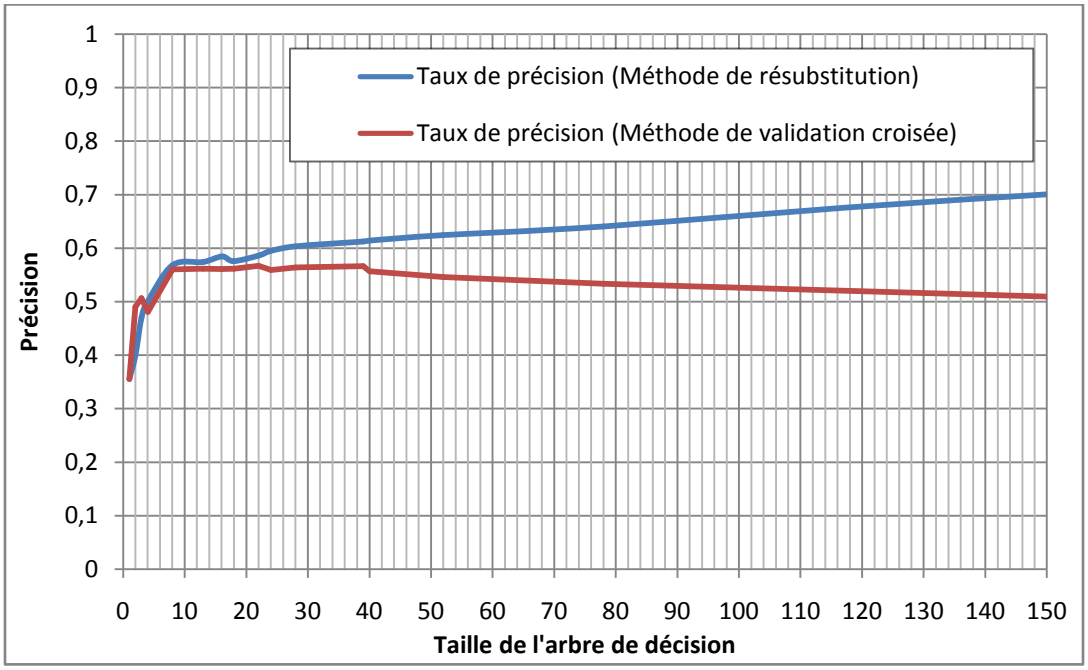

Figure 6.1: Évolution de taux de précision

La figure 6.1 illustre les taux de précision et la taille de l'arbre de décision. La courbe du taux de précision à l'apprentissage (la méthode de redistribution) croit avec la taille de l'arbre de décision. Dans la littérature, il n'y a pas une méthode rigoureuse pour choisir la taille optimale de l'arbre, mais il est préconisé de choisir celle qui correspond à la divergence entre la courbe du taux de précision en apprentissage et la courbe du taux de précision par le test de validation croisée (Saporta, 2006). Selon ce principe, on a choisi la taille de l'arbre de décision égal à 22 qui correspond à la valeur maximale du taux de précision calculé par la méthode de validation croisée (0,56). La figure 6.2 illustre l'arbre de décision choisi. Les codes des facteurs physiques sont présentés dans l'annexe E. Il faut noter que chaque feuille de l'arbre de décision est étiquetée par le groupe majoritaire.

- Type_voie = [AU]
  - Vitesse_aut = [100-60km/h]
    - sens_HC/VC = [VM]
      - sur_pont = [Sur_p=Oui] alors Groupe TP = **TP2** (53,33 % de 30 tronçons)
      - sur_pont = [Sur_p=Non]
        - Nbr_sorties = [NS0]
          - Nbr_entrées = [NE1] alors Groupe TP = **TP2** (46,15 % de 26 tronçons)
          - Nbr_entrées = [NE0] alors Groupe TP = **TP2** (59,38 % de 32 tronçons)
          - Nbr_entrées = [NE2] alors Groupe TP = **TP6** (100,00 % de 1 tronçon)
        - Nbr_sorties = [NS1] alors Groupe TP = **TP2** (59,18 % de 49 tronçons)
        - Nbr_sorties = [NS2] alors Groupe TP = **TP2** (60,00 % de 5 tronçons)
    - sens_HC/VC = [HM] alors Groupe TP = **TP2** (90,07 % de 151 tronçons)
  - Vitesse_aut = [80Km/h] alors Groupe TP = **TP2** (45,45 % de 22 tronçons)
  - Vitesse_aut = [moins de 50km/h] alors Groupe TP = **TP3** (50,00 % de 4 tronçons)
  - Vitesse_aut = [70Km/h] alors Groupe TP = **TP3** (52,11 % de 142 tronçons)
  - Vitesse_aut = [90Km/h] alors Groupe TP = **TP3** (66,67 % de 3 tronçons)
  - Vitesse_aut = [60km/h] alors Groupe TP = **TP3** (0,00 % de 0 tronçon)
- Type_voie = [BV/AV]
  - Vitesse_aut = [100-60km/h] alors Groupe TP = **TP8** (100,00 % de 1 tronçon)
  - Vitesse_aut = [80Km/h] alors Groupe TP = **TP7** (100,00 % de 1 tronçon)
  - Vitesse_aut = [moins de 50km/h]
    - Nbr_voies = [NV3] alors Groupe TP = **TP6** (37,93 % de 29 tronçons)
    - Nbr_voies = [NV2] alors Groupe TP = **TP6** (60,71 % de 56 tronçons)
    - Nbr_voies = [NV4] alors Groupe TP = **TP7** (75,00 % de 4 tronçons)
    - Nbr_voies = [NV1] alors Groupe TP = **TP1** (46,67 % de 15 tronçons)
  - Vitesse_aut = [70Km/h] alors Groupe TP = **TP3** (25,00 % de 44 tronçons)
  - Vitesse_aut = [90Km/h] alors Groupe TP = **TP6** (44,44 % de 9 tronçons)
  - Vitesse_aut = [60km/h] alors Groupe TP = **TP6** (57,14 % de 7 tronçons)
- Type_voie = [Rue/Route] alors Groupe TP = **TP6** (38,00 % de 50 tronçons)

Figure 6.2: L’arbre de décision de groupe FP

Le premier constat qu’on peut tirer à partir de l’arbre de décision choisi (figure 6.2) est que les facteurs physiques les plus pertinents pour expliquer l’état de la circulation sur les tronçons sont : type de voies, vitesse autorisée, sens de circulation, nombre de voies, et nombre d’entrées et de sorties. Le degré de la pertinence d’un facteur dépend de sa position par rapport au nœud racine. En effet, un facteur physique est plus pertinent qu’un autre facteur, s’il est le plus proche du nœud racine

Concernant les règles décision, on s’intéresse à celles qui sont les plus performantes (c'est-à-dire celles ayant les taux de précision les plus élevés) et qui couvrent un grand nombre de tronçons

par rapport aux autres règles. En effet, on s'aperçoit que le temps de parcours le plus faible caractérisant le groupe TP2 peut être expliqué par le type de voies «autoroute», la vitesse autorisée «100-60km». De plus, on peut appréhender l'état de la circulation sur les tronçons de groupe TP3 par le type de voie «autoroute» et la vitesse autorisée qui est égale à 70km/h. En outre, l'état de la circulation de groupe TP6 est expliqué par un type de voie «boulevard/avenue» à deux voies et par la vitesse autorisée «moins de 50 km/h».

Supposons que les facteurs physiques permettent de dégager les tronçons ayant une circulation normale, c'est-à-dire ceux dont le temps de parcours moyen dépend des caractéristiques physiques. Selon ce principe, l'arbre de décision permet de distinguer trois groupes dont la circulation est considérée normale, soit les groupes TP2, TP3 et TP6. En effet, l'état de la circulation sur la plupart des tronçons de groupe TP6 est considéré normal, malgré que son temps de parcours moyen soit plus élevé par rapport à celui des groupes TP2 et TP3; ce temps de parcours est expliqué par les caractéristiques facteurs de ces tronçons.

À partir de cet arbre de décision, on peut ainsi déterminer les tronçons qui sont bien et mal classés par les facteurs physiques. Ceci est présenté par la matrice de confusion (tableau 6.1) permettant d'analyser les nouveaux emplacements des tronçons aux groupes TP. Notons que les groupes sont présentés dans cette matrice en ordre croissant de temps de parcours moyen. Les cellules bleues de ce tableau représentent le nombre de tronçons bien classés et les cellules rouges ou vertes désignent le nombre de tronçons mal classés à partir des facteurs physiques. On s'aperçoit, à partir de cette matrice, que la plupart des tronçons de groupe TP4 sont assignés au groupe TP2. Ces tronçons se caractérisent par: un type de voie «autoroute», une vitesse autorisée de «100-60 km/h» et un sens de circulation «vers le centre-ville de Montréal». On constate également que les facteurs physiques ne permettent pas d'expliquer l'état de la circulation de la plupart des tronçons des groupes TP7 et TP8. En effet, les tronçons de ces groupes qui se caractérisent par un temps de parcours moyen élevé par rapport aux autres groupes TP (hormis le groupe TP4), par un type de voie «boulevard/avenue» à 2 ou 3 voies et par une vitesse autorisée égale à 60km/h ou moins de 50km/h, sont affectés au groupe TP6. On pourrait justifier cela par la présence de la congestion, en particulier, sur les tronçons des groupes TP4 parce qu'il y a une différence remarquable entre le temps de parcours moyen de ce groupe et celui du groupe TP2.

À partir de l'analyse de l'arbre de décision, on constate que lorsqu'un tronçon mal classé est, dans la plupart des cas, affecté à un groupe qui se caractérise par un temps de parcours moyen plus faible que celui de son groupe réel, à titre d'exemple, on cite le cas des tronçons de groupe TP4 qui sont affectés au groupe TP2 et aussi, le cas de la majorité des tronçons des groupes TP7 et 8 qui sont assignés au groupe TP6 (voir les cellules colorées en rouge). À priori, on peut dire qu'il y a d'autres facteurs, outre les facteurs physiques, qui peuvent expliquer l'augmentation du temps de parcours moyen de ces tronçons mal classés. De plus, on observe qu'il y a des tronçons mal classés, mais qui sont affectés cette fois, à des groupes TP caractérisés par un temps de parcours plus élevé que celui de leur groupe réel (voir les cellules colorées en vert). En effet, quelques tronçons de groupe TP2 caractérisés par un temps de parcours moyen égal à 39,17 secondes (vitesse moyenne égale à 91km/h) sont des autoroutes avec une vitesse de 70km/h, appartiennent au groupe TP3; dans ce cas, il s'agit d'un dépassement de la vitesse autorisée.

À partir de l'analyse de la matrice de confusion, on peut distinguer trois niveaux d'état de la circulation.

- Niveau A correspond à la circulation dans des conditions normales de trafic (circulation normale, voir la définition proposée dans la section 2.1 du deuxième chapitre) c'est à dire celle expliquée par les facteurs physiques, tels que le type de voies et la vitesse autorisée. Selon ce principe, la circulation sur 59% des tronçons est considérée normale (cela correspond aux cellules colorées en bleu).

- Niveau B : le temps de parcours est plus faible que celui dans la circulation normale (niveau A). Dans ce cas, les facteurs physiques ne représentent pas une contrainte pour la circulation, à titre d'exemple, on peut citer le cas d'un dépassement de la vitesse autorisée. Le pourcentage des tronçons appartenant à ce niveau de circulation est égal à 8% du nombre total de tronçons (cela correspond aux cellules colorées en vert).

- Niveau C : la circulation saturée, il s'agit d'une augmentation du temps de parcours par rapport à celui dans la circulation normale (niveau A). La proportion des tronçons de ce type de circulation est égale à 33% de nombre total de tronçons (cela correspond aux cellules colorées en rouge). Il faut noter que, dans ce niveau, on ne fait pas la distinction entre une légère augmentation du temps de parcours par rapport à la situation normale et la congestion proprement dite. La première section du deuxième chapitre met en avant la

différence entre la circulation dans des conditions normales de trafic et la circulation saturée (voir figure 2.1)

Dans la même perspective, on a séparé, dans chaque groupe TP, les tronçons selon ces trois niveaux. Ceci est présenté par la figure 6.2. À partir de celle-ci, on peut appréhender l'état de la circulation sur les tronçons des groupes TP4 et TP5 comme une congestion. En outre, on s'aperçoit que la circulation sur la majorité des tronçons de groupes TP6 est normale et que près de 27% des tronçons de ce groupe sont considérés saturés. Le groupe TP1, quant à lui, regroupe autant des tronçons congestionnés que des tronçons ayant un temps de parcours qui dépasse celui dans la circulation normale.

Il faut toujours garder à l'esprit que ces analyses se reposent sur l'hypothèse que la circulation dégagée par les facteurs physiques les plus pertinents (selon l'arbre de décision) est normale et qu'à partir de laquelle on pourrait appréhender l'état de la circulation d'un tronçon (à quel niveau de circulation, pourrions-nous l'affecter).

Dans l'espoir de vérifier cette hypothèse, on a fait une comparaison entre la vitesse autorisée qui est le deuxième facteur physique pertinent, et la vitesse moyenne (calculée à partir du temps de parcours moyen) pour chaque tronçon affecté au niveau A (circulation normale). Cela a monté que la vitesse autorisée est légèrement supérieure à la vitesse moyenne; on trouve que ceci est raisonnable.

Tableau 6.3 : Matrice de confusion des groupes TP

		Groupes TP prédits								
		TP2	TP3	TP1	TP6	TP5	TP7	TP4	TP8	Total
Groupes TP réels	TP2	225	13	3	1	0	0	0	0	242
	TP3	41	89	0	11	0	0	0	0	141
	TP1	6	25	7	27	0	0	0	0	65
	TP6	6	20	2	73	0	1	0	0	102
	TP5	15	25	0	7	0	0	0	0	47
	TP7	1	11	3	26	0	4	0	0	45
	TP4	20	5	0	2	0	0	0	0	27
	TP8	1	5	0	5	0	0	0	1	12
	Total	315	193	15	152	0	5	0	1	681

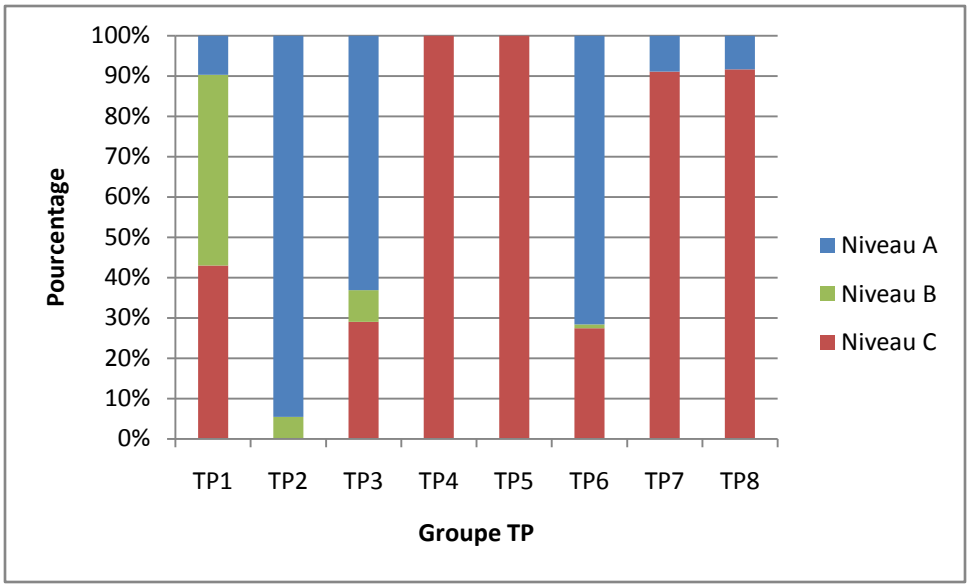

Figure 6.3: Répartition des tronçons par niveau de circulation

Jusqu'à présent, on a utilisé l'arbre de décision pour le coté descriptif. En se référant au résultat obtenu par la méthode de validation (10 sous- ensembles), on note que les facteurs physiques ne permettent pas de prédire l'appartenance d'un tronçon à un groupe TP à partir des facteurs physiques (le taux de précision selon la méthode de validation croisée égal à 0,56, voir figure 6.1). D'après le tableau 6.2, on constate que les facteurs physiques permettent de prédire l'appartenance des tronçons des groupes TP2, 3 et 6 avec des taux de précision égalent respectivement à 0,92; 0,59 et 0,69. En outre, les facteurs physiques ne permettent pas de prédire l'appartenance des tronçons aux groupes TP4, TP5 et TP8. Cela peut confirmer l'hypothèse que la circulation sur la plupart des tronçons des groupes TP2, 3 et 6 est normale.

Tableau 6.4 : Le taux de précision (prédiction des groupes TP)

	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8
méthode de validation croisée	0,10	0,92	0,59	0	0	0,69	0,05	0

6.3 Conclusion

Dans ce chapitre, nous avons vérifié la dépendance entre le regroupement fait à partir de temps de parcours et celui bâti sur les facteurs physiques au moyen de test de khi-deux et on a étudié les interactions entre les groupes TP et les groupes FP. Ensuite, au moyen de l'arbre de décision, on a essayé d'appréhender l'état de la circulation à partir des facteurs physiques des tronçons, de

dégager la circulation normale et d'exposer les facteurs physiques les plus pertinents permettant d'affecter un tronçon à un groupe TP. Enfin, on a mis en avant que les facteurs physiques ne permettent de prédire l'appartenance d'un tronçon dans un groupe TP.

CHAPITRE 7 CONCLUSIONS ET PERSPECTIVES

À l'issue de notre étude qui porte sur la méthode d'analyse et de classification des segments du réseau routier supérieur de la région de Montréal, nous avons abordé deux concepts clés : un premier relevant du thème du transport et un second plutôt orienté vers le *data mining*.

7.1 Contributions

Cette étude n'est pas la première étude liée au temps de parcours. Jusqu'à présent, peu d'analyses exposent la relation entre les caractéristiques physiques des routes et l'état de la circulation. Ces analyses avaient pour but de mesurer la congestion routière. Ce document est le premier qui porte sur la prédiction d'état de la circulation à partir des données des caractéristiques physiques des routes au moyen des techniques du *data mining*.

Cette étude fait suite aux travaux de Loustau *et al.*(2009). Durant cette recherche, on a essayé de répondre à certaines questions posées par ces auteurs ou par le MTQ sur les différentes conditions routières et les facteurs ayant des répercutions non seulement sur la mesure de la congestion routière, mais aussi sur la mesure de la fiabilité du réseau routier montréalais.

Dans ce travail, on a dressé le portrait général des segments routiers échantillonnés par le MTQ. Pour en exposer les différentes particularités, les tronçons ont été regroupés en segments homogènes à partir de leurs facteurs physiques. Douze groupes ont été obtenus dans cette étude. Ensuite, une analyse bivariée a permis de vérifier la relation entre les groupements faits à partir des facteurs physiques et ceux bâtis sur les temps de parcours. Cela met en avant l'opportunité au MTQ d'intégrer des données des facteurs physiques dans le prochain plan d'échantillonnage des segments routiers pour collecter de nouveau des données du temps de parcours.

Dans l'espoir d'améliorer au MTQ la connaissance sur l'état de la circulation sur les segments routiers qu'il a échantillonnés, à partir des facteurs physiques, un arbre de décision a été utilisé. Cet outil du *data mining* a été employé pour la prédiction et aussi pour le côté descriptif. Selon certaines hypothèses, cela a permis de dégager la circulation normale et, par la suite, de mieux appréhender l'état de la circulation sur les tronçons. Dans cette étude, les facteurs physiques ne permettent pas de prédire l'appartenance d'un tronçon aux groupes TP.

7.2 Limitations

Plusieurs limitations ont été soulignées au cours de ce travail. Premièrement, il y a le problème de la qualité de la base de données des facteurs physiques. En effet, dans certains cas, le même tronçon est caractérisé par deux modalités différentes, tel que la vitesse autorisée. Dans ce cas, on a affecté au tronçon la modalité qui lui couvre plus en termes de longueur. De plus, on n'a pas pu caractériser certains tronçons à partir de Google Street View. Ces tronçons ont été exclus de cette étude.

Parmi les problèmes qu'on a rencontrés pour prédire l'appartenance des tronçons aux groupes TP à partir des facteurs physiques, est le grand nombre de groupes TP et le faible effectif des tronçons dans certains groupes, tels que les groupes TP4, 7 et 8. De plus, certaines règles de décision sont vérifiées par un faible nombre de tronçons. En outre, le dégagement de la circulation normale dépend du choix de la taille de l'arbre de décision, des facteurs physiques qui sont considérés pertinents pour l'affectation des tronçons aux groupes TP, de la particularité des tronçons échantillonnés par le MTQ et le groupement effectué par Loustau *et al.(*2009).

7.3 Perspectives

Cette étude présente une première approche pour expliquer le groupement des tronçons fait selon les temps de parcours (travaux de Loustau *et al.*, 2009). Elle dévoile qu'il y a des tronçons qui appartiennent au même groupe TP n'ont pas le même niveau de la circulation. Pour améliorer la compréhension de l'état de la circulation sur le réseau routier de la grande région de Montréal, il s'agit de regrouper, pour chaque groupe FP, les segments routiers selon les distributions fréquentielles des relevés de temps de parcours. Cette approche permettra de distinguer les segments routiers les plus problématiques sans avoir de se soucier de leurs caractéristiques physiques.

Vu qu'elles ne sont pas disponibles lors de la réalisation du projet, d'autres données des caractéristiques physiques des segments routiers peuvent être ajoutées à notre base de données, telles que la capacité d'un tronçon et sa localisation par rapport au centre-ville de Montréal, le type de zone et la géométrie des tronçons. Ces caractéristiques physiques peuvent être pertinentes lorsqu'il est question de dégager la circulation normale et de prédiction des groupes TP.

Ce travail peut servir aux études menées sur les incidents. En effet, ceci permettra d'exposer les nouvelles particularités des tronçons échantillonnés par le MTQ. De plus, on pourra expliquer la dégradation du niveau de circulation sur certains segments du réseau routier supérieur de la région montréalaise à partir du taux d'incident, et aussi par les facteurs physiques. À après avoir dégagé la circulation normale, il est possible de se servir des données d'incidents pour mieux appréhender l'état de la circulation et pour distinguer entre une congestion récurrente et non récurrente.

BIBLIOGRAPHIE

Aichour, B. (2006). Les problèmes des transports urbains et leur impact sur la circulation à Constantine. Cahiers Scientifiques du Transport N° 50/2006, pp35-60.

Agard, B., and Kusiak, A. (2005). Exploration des Bases de Données Industrielles à l'aide du Data Mining – Perspectives, 9ème Colloque National AIP PRIMECA, La Plagne5-8 Avril 2005. Consultée le 10 Novembre 2009, tiré de http://www.simagi.polymtl.ca

Agard, B., Trépanier, M., and Du Parc, N. Études des générateurs de déplacement à l'aide de données de cartes à puces, Communication présentée au 44e congrès de l'Association Québécoise du Transport et des Routes, Montréal. Consulté en janvier 2010, tiré de http://www.simagi.polymtl.ca

Alberto, M. F.M., and Andrew, P.t. (2005). Speed Factors on Two-line Rural Highways in free-Flow Conditions. Geometric design and the effects on Traffic Operations. Transportation Research Board, N°.1912, Washington, pp.39-46.

Archer, J., Fotheringham, N., Symmons, M., Corben. B. (2008). *The Impact of Lowered Speed Limits in Urban Areas*. Report 279 Transport Accident Commission. P.71.

Babin, A., and Gourvil, L. (2007). La base de données sur les relevés-tronçons des voitures flottantes du MTQ – Document technique, Ministère des transports du Québec (Service de la modélisation des systèmes de transport), Janvier 2007.

Baccini, A., and Besse, P. (2004). Data Mining: Exploration Statistique, Laboratoire de Statistique et Probabilités — UMR CNRS C5583, Université Paul Sabatier, Consulté le 15 Janvier 2010, tiré de http://wwwirma.ustrasbg.fr

Banister, D., and Lichfield, N. (1995). Chapitre1: The key Issue in Transport and Urban Development. Transport and Urban Development. Edited by David Banister, published by E& FN SPON, pp 1-16.

Bramer. M. (2007). Principales of Data MINING. Undergraduate Topics in computer Science, Springer-Verlag London. Page 344.

Mining Methods and Geographic Information, Technical Report NCSA-ALG03-0002. Consulté le 15 Novembre 2009, tiré de http://isda.ncsa.illinois.edu

Benzécri, F. (1980). Pratique de l'analyse de données: analyse des correspondances, dunod. P424.

Chen, Z., and Johm, N.I. (2005). Effects of Geometric Characteristics on Heard-On Crash Incidence on Two-Lane Roads in Connecticut. Statistical; Highway Safety Data, Analysis, and Evaluation; Occupant Protection; Systematics Revieux and Meta-analysis. Transportation Research Broad N°.1908, Washington, pp.159-164.

Büchner, A.G., Anand, S.S. and Hughes, J.G. (1997). Data Mining in Manufacturing Environments: Goals, Techniques and Applications, *I*n Studies in Informatics and Control. Consultée le octobre 2009, tiré de http://www.infj.ulst.ac.uk

CEMT. (1998). La congestion routière en Europe, Rapport de la cent deuxième Table Ronde d'économie de Transport, tenue le à Paris les 12 et 13 1998.

ECMT. (2007). Managing Urban Traffic Congestion. OCDE. p.31.

Comité Interrégional pour le Transport Des marchandises. (1999). La congestion routière et les le transport des marchandises : diagnostic. p.57.

Cornuéjols, A., and Miclet, L. (2003). Apprentissage artificiel: concept et algorithmes. Enrolles. p.591.

Cheng, Y.,Zhang, Y.; Hu, J. and Li, L. (2007). Mining for Similarities in Urban Traffic Flow Using Wavelets, Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference Seattle, WA, USA, Sept. 30 - Oct. 3. Consulté en novembre 2009, tire de http://ieeexplore.ieee.org.

Deakin, E., and Harvey, G. (1993). A manual of regional transportation modeling practice for air quality analysis, National Association of Regional Councils, Washington, D.C. Consulté le 17 octobre 2009, tiré de http://tmip.fhwa.dot.gov

Desbois, D. (2008). L'analyse des correspondances multiples « À la Hollandaise » : Introduction à l'analyse d'Homogénéité, Revue Modulad. Consulté Nombre Novembre 2010, tiré de http://www-rocq.inria.fr

Devéze, B. and Fouquin, M. (2004). Data Mining C4.5 – DBSCAN, EPITA-SCIA. Consulté Novembre 2010, tiré de http://devezeb.free.fr

Escofier, B., Pagès, J.(1998). Analyse factorielles simples et multiples: objectifs, méthodes et interprétation. Dunod, 3ème édition. Paris. PP284.

Fayyad, U.; Piatetsky,- Shapiro, G. and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communication Of The ACM, Vol. 39, No. 11, pp. 27-34.

Feuilloy, M. (2009). Étude algorithmiques d'apprentissage artificiel pour la prédiction de la syncope chez l'homme. Thèse de doctorat, spécialité informatique, École Doctorale STIM. Consulté le 10 décembre 2010, tirée de http://tel.archives-ouvertes.fr/docs/00/46/50/08/PDF/Manuscrit_These_Feuilloy_temp.pdf

FHWA. (2003). Freeway Management and Operations Handbook. Rapport FHWA-OP-04-003. Washington, DC.

FHWA (2005). Traffic Control systems Handbook. Report FHWA-HOP-06-006. Washington, DC.

Foucart, T. (1984). Analyse Factorielle de Tableau Multiple. Deuxième édition. Masson, Paris. P.234.

Giudici. P. (2003). Applied Data Mining, statistical Methods for Business and industry.John Wiley & Sons Ltd, England. Page364.

Gourvil, L., Joubert, F. (2004). Évaluation de la congestion routière dans la région de Montréal, Ministère des transports du Québec, *Collection Études et Recherches en Transports* – N° RTQ-04. Consulté le 25 octobre 2009, tiré de http://www.mtq.gouv.qc.ca

Gouvernement du Québec. (1995). Planification des transports et révision des schémas d'aménagement. Consultée en Aoute 2009, tirée de http://www.mamrot.gouv.qc.ca

Grangé, D., Lebart, L. (1996).Traitement statistiques des enquêtes. Dunod, Paris. PP 255.

Haccoun, R.R., Cousineau,D (2010). Statistiques : Concepts et application. Deuxième édition Revue et Augmentée. Les presses de l'université de Montréal, Québec. P.456.

Hanks, J.W and Lomax, T.J..(1990). Roadway congestion in Major Urban areas: 1982 to 1988. Report 1131-3. Texas Transportation Institue. Consulté en Décembre 2009, tiré de http://tti.tamu.edu

Hanks, J.W and Lomax, T.J. (1992). 1989 Roadway congestion Estimates and Trends, Research Report N 113 1-4, cooperative research program, Texas Transportation Institute, The Texas A&M university systems, College Stations, Texas.

Han. J., and Kamber. M. (2001). Data Mining Concepts and Techniques. Morgan Kaufmann Publishers. Page 550.

Jambu, M.(1989). Exploration informatique et statistique des données. Collection technique et scientifique des télécommunications, Dunod. Paris. PP505.

Jiang, W., Vaidya, J., Balaporia, Z., Clifton, C. and Banich, B.(2005). Knowledge Discovery from Transportation Network Data, In the 21st International Conference on Data Engineering (ICDE 2005), Tokyo, Japan, pp.1061-1072. Consulté en decembre 2010, tiré de http://cimic.rutgers.edu

Kou, Y., Lu, C.T., Santos Jr., R.F.D.: Spatial Outlier Detection: A Graph-based Approach. In: 19th IEEE International Conference on Tools with Artificial Intelligence, pp. 281—288. IEEE, 2007.

Laffly, D. (2009). Analyse Bivariée de Variables Qualitatives : Le Test de Chi2, Laboratoire Société Environnement Territoire, UMR 5603 du CNRS et Université de Pau. Consulté en Avril 2010, tiré de http://web.univpau.fr

Larose, D. T. (2005). Des Données à la connaissance : une introduction au data mining, traduction, traduction et adaptation de Thierry Vallaud. Éditions Vuibert Informatique, Paris. P.223.

Ludovic, L. (2008). L'analyse des données des origines à 1980 : quelques éléments. Journal Électronique d'Histoire des Probabilités et de la Statistique. Vol 4, n°2. Consulté le 8 Février 2010 tirée de www.jehps.net

Le Centre Pour un Transport Durable. (2002). Définition et Vision du Transport durable, consulté en Aoute 2009, tirée de http://cst.uwinnipeg.ca

Léfébure, R., et Venturi. V.,(1998). *Le Data Mining.* Eyrolls: Paris

Le rapport de Gouvernement du Québec (1995). Planification des transports et révision de schémas d'aménagement. Consulté le 10 Octobre 2009, tiré de http://www.mamrot.gouv.qc.ca/

Loustau. P.(2009). Modélisation des temps de parcours sur un réseau routier à l'aide de données de véhicules flottants. M. Sc. A, école polytechnique de Montréal, QC, Canada.

Loustau.P., Morency. C., Trépanier.M, and Gourvil.L. (2009). Portrait de la fiabilité des temps de parcours sur le réseau autoroutier montréalais. Consulté le 5 décembre 2009, tiré de http://www.aqtr.qc.ca

Luo, Q.(2008). Transportation Data Analyzing by Using Data Mining Method, IEEE, International Symposiums on Information Processing, pp. 766-767.

Mary, J. (2005). Étude de l'Apprentissage Actif, Application à la Conduite d'Expériences, Ph.D., Université Paris XI. Paris : France. Consulté en Mars 2010, tiré de http://tao.lri.fr

MIRO (2006). Analyse des temps de parcours sur le réseau routier de la grande région de Montréal : Rapport final, Projet N° Q94231, Ministère des transports du Québec.

NCHRP (1997). *Report 398: Quantifying Congestion Final Report:* Volume 1, Rapport finale. Transportation research Broad, National Research Council. Washington, DC: National Cooperative Highway research Program.

NCHRP (2001). *Economic Implication of Congestion.*, National Cooperative Highway Research Program Rapport 463. Transportation Research Board, National Research Concil, Washington,D.C.

Nakache, J.P., Confais, J.(2003). Statistique explicative appliquée. Édition Technip. P.296.

Nesamani, K.S., Chu, L.Y., MacNally, M.G., Jayakrishnan,R. (2005). Estimation of Vehicular Emissions by Capturing Traffic Variations. Consulté le 10 Octobre 2009, tiré de http://www.escholarship.org

Pouliot, M. et Danserea, N.(1998), Transports et développement économique, Chapitre 8 : Concepte 1. *Site Web Géographie des Transports*, Hofstra University: Department of Economics and Geography. Consulté le 20 juillet 2009, tirée de http://www.geog.umontreal.ca

Prud'homme R., Ming S. Y.(2000). Les coûts économique de la congestion du périphérique Parisien : une approche désagrégées », Les cahiers scientifiques de transport N°37/2000, pp. 59-73.

Prud'homme R., Darbera R., Newbery D., Diekman A., Elbeck B.(1999). Notre système de transport actuel est il durable? = Is our present transport système sustainable ?. Presses de l'école nationale des ponts et chaussées : France.

Pyle. D., (1999). Data Preparation for Data Mining. Morgan Kaufmann, USA.

Quinlan, J., R. (1993). Programme forMachine learning., Morgan Kaufman, USA

Rakotomalala, R. (2005).Tanagra : une plate-forme d'expérimentation pour la fouille de données, Revue Modulad, Numéro 32, 70-85, 2005. Consulté en avril 2010, tiré de http://www-rocq.inria.fr

Reymond, M.(2005). La tarification de la congestion automobile : acceptabilité sociale et redistribution des recettes du péage. Thèse de doctorat, Université Montpellier I, décembre 2005, pp. 14-16.

Robitaille, M., Nguyen, T. (2003). Évaluation de la congestion « de la théorie à la pratique »: Réseau routier de l'agglomération de Montréal. Communication présentée au *Congrès 2003 de l'Association des Transports du Canada*, St John's – Terre-Neuve. Consulté 10 octobre2009, tiré de http://www.tac-atc.ca

Saporta, G., (2006). Probabilités, analyse de données et statistique, 2ème édition révisée et argumentée, Paris. P.622.

Sarah B.M., Michael J.D, 2003, « Développement of congestion performance Measures using its information », Rapport Final, Virginia Transportation Research Council, 03-R1, janvier 2003

Sholom, M. W., and Nitin, I. (1998). *Predictive Data Mining*. San Francisco, California: Morgan Kaufmann Publishers, Inc.

Tufféry, S. (2007). Data Mining et statistique décisionnelle : l'intelligence des données. Édition Technip, Paris. P.533.

Tufféry, S. (2009). Data Mining et statistique décisionnelle : l'intelligence des données. Édition Technip, Troisieme edition actualisée et augmentée, Paris. P.705.

Wang, Y.J., Yu, Z.C., He, S.B., Cheng, J.L. and Zhang, Z.J.(2009). A Data-mining-based Study On Road Traffic Information Analysis And Decision Support, Second Pacific-Asia Conference on Web Mining and Web-based Application, pp.24-27. Consulté novembre 2009, tire de http://ieeexplore.ieee.org

Westphal, C. and Blaxton, T. (1998*). Data Mining Solutions Methods and Tools for Solving Real-World Problems*. Wiley: New York.

Whitehead, D., Fournier, P. (2000). Développement d'indicateurs descriptifs de la congestion routière – Document de travail, ministère des Transports du Québec (Service de la modélisation des systèmes de transport), Juillet 2000.

Xu, R. and Wunsch II, D. (2005). Survey of Clustering Algorithms, IEEE Transactions On Neural Networks, VOL. 16, NO. 3, pp. 645 – 678.

Zighed, D.A,. Rakotomalala, R. (2002). Extraction de connaissance à partir de données (ECD). Techniques de ingénieurs, Traité Informatique. p.24.

Zhang, H.S., Zhang, Y., Li, Z.H. and Hu D.C. (2004). Spatial–Temporal Traffic Data Analysis Based on Global Data Management Using MAS, IEEE Transactions on Intelligent Transportation Systems, Vol. 5, N° 4, pp. 267- 275.

Zuindeau, B. (1999). L'analyse des externalités environnementales : un essai régulationniste. Issue d'une communication effectuée aux sixièmes journées de L'IFRESI à Lille (22avril 1999).

ANNEXE A : ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES

L'analyse factorielle des correspondances multiples (AFCM) est une méthode statistique de l'analyse des données. Elle a le même principe que l'analyse factorielle des correspondances (AFC). Toutefois, cette méthode est utilisée dans le cas où il y a plus de deux variables qualitatives à étudier. Elle est également recommandée lorsque la multiplicité des modalités de ces variables peut engendrer certains problèmes d'analyse. Cette méthode vise à étudier les relations entre les différentes modalités et permet de synthétiser et d'agréger les données, et de les représenter dans un plan multidimensionnel (Tufféry, 2007). Dans leur livre « Traitement Statistiques des Enquêtes », Grangé et Lebart (1996) ont noté que: « la possibilité de réduire la diminution provient de l'existence de corrélations entres les variables ».

- **Notation**

On note $A_1,..,.A_s,..,A_S$ les variables qualitatives étudiées décrivant les caractéristiques de l'ensemble d'éléments E, S est le nombre total de variables et n est le nombre total d'éléments. Chaque variable A_s est considérée comme un vecteur :

$$A_s = [X_s(1), \dots, X_s\,(j), \dots\dots X_s\,(P_s)\,]$$

Où P_s représente le nombre de modalités de A_s et $X_s\,(j)$ est la modalité j de la variable A_s.

Les données se représentent sous la forme d'un tableau T appelé tableau disjonctif complet.

Tableau A.7-1 : Tableau disjonctif complet T (adapté de Foucart, 1984)

	A_1					....	A_s					....	A_S				
	$X_1(1)$	....	$X_1(p)$	....	$X_1(P_1)$	....	$X_s(1)$	....	$X_s(p)$	....	$X_s(P_s)$	....	$X_S(1)$	....	$X_S(p)$	....	$X_S\,(P_S)$
1																	
⋮																	
i	0	0	1	0	0	....	0	1	0	0	0	....	0	0	0	0	1
⋮																	
n																	

La ligne i du tableau T caractérise l'élément i et les colonnes représentent les modalités des attributs.

Les cellules de ce tableau ne contiennent que des valeurs binaires (0,1). Si la modalité a été prise par l'élément i alors la valeur de la colonne de cette modalité est égale à 1, sinon la valeur est égale à 0. On note aussi que pour chaque variable A_s, une seule modalité qualitative a la valeur 1 et par conséquent :

- L'effectif marginal en ligne est égal au nombre d'attributs (S).
- L'effectif marginal en colonne correspond au nombre d'éléments qui vérifient la modalité j.
- L'effectif total de chaque attribut A_s est égal à n
- La somme du tableau T est égale à nS.

Le tableau T peut être aussi représenté d'une manière plus simple.

Tableau A.7-2 : Tableau disjonctif complet

	m_1	....	m_j	....	m_J	Total
1	a_{11}	....	a_{1j}	....	a_{1J}	$a_{1.}$
⋮		....		....		⋮
		....		....		
i	a_{i1}	....	a_{ij}	....	a_{iJ}	$a_{i.}$
⋮		....		....		⋮
		....		....		
n	a_{n1}	....	a_{nj}	....	a_{nJ}	$a_{n.}$
Total	$a_{.1}$	....	$a_{.j}$	....	$a_{.J}$	n*S

Soit $M = \{m_1, \dots, m_J\}$ l'ensemble de modalités de tous les attributs et J est l'effectif de l'ensemble M.

La ligne i du tableau T caractérise l'élément i et la colonne j caractérise la modalité m_j, avec i=1…,n et j=1….,J. a_{ij} est la valeur de la cellule (i,j) et qui prend la valeur 0 ou 1; si la modalité m_j a été prise par l'élément i alors a_{ij} est égale à 1, sinon elle prend la valeur 0.

On note que $a_{.j}$ est la somme de la $j^{ème}$ colonne et $a_{i.}$ est la somme de la $i^{éme}$ ligne du tableau T et qui est égal à S.

D'après Tuffèry (2010), le poids de chaque modalité j est égal à :

$$w_j = \frac{a_{.j}}{n*S} \quad (A.1)$$

Calcul de la distance

Il y a plusieurs méthodes pour calculer la distance entre deux points. D'après Benzécri (1980) et Saporta (2006), la métrique utilisée dans l'analyse factorielle des correspondances multiples est la distance de khi-deux(X^2). La distance, que ce soit entre deux modalités ou entre deux éléments, est calculée comme suit.

La distance entre deux modalités j et j' vaut :

$$d(j,j') = \sum_{i=1}^{n} \frac{n}{S} \left(\frac{a_{ij}}{a_{.j}} - \frac{a_{ij'}}{a_{.j'}}\right)^2 \quad (A.2)$$

La distance entre deux éléments i et i' s'écrit:

$$d(i,i') = \sum_{j=1}^{J} \frac{n}{a_{.j}} (a_{ij} - a_{i'j})^2 \quad (A.3)$$

La distance entre la modalité j et le centre de gravité g du nuage des modalités s'écrit :

$$d(j,g) = n \sum_{i=1}^{n} \left(\frac{a_{ij}}{a.j} - \frac{1}{n}\right)^2 = \frac{n}{a.j} - 1 \quad (A.4)$$

Avec le centre de gravité du nuage des modalités j est un vecteur de dimension n et s'écrit:

$$g = [\frac{1}{n}, \dots \frac{1}{n}, \dots \frac{1}{n}] \quad (4.5)$$

L'inertie du nuage des modalités

- L'inertie $I_n(j)$ de la modalité j représente sa contribution dans l'inertie totale :

$$I_n(j) = W_j \; d(j,g) = \frac{1}{S}\left(1 - \frac{a_{.j}}{n}\right) \quad (A.6)$$

Rappelons que w_j est le poids de la modalité j, $a_{.j}$ est l'effectif de la modalité j et S est le nombre d'attributs.

- L'inertie de l'attribut A_s qui a P_s modalités vaut donc :

$$I_n(A_s) = \sum_{p=1}^{P_s} I_n(p) = \frac{1}{S} (P_s - 1) \quad (A.7)$$

- L'inertie totale vaut :

$$I_T = \sum_s I_n(A_s) = \frac{J}{S} - 1 \qquad (A.8)$$

Rappelons que J est le nombre total de modalités et S est le nombre d'attributs.

On constate que l'inertie totale ne dépend pas de la distance entre deux variables, elle s'exprime uniquement par le nombre de variables et de modalités.

Matrice des composantes principales et détermination des axes factoriels

Cette partie consiste à déterminer la matrice des composantes principales et les équations des axes principaux, que ce soit pour les attributs ou pour les modalités.

Soit D la matrice diagonale des effectifs marginaux des modalités.

$$D = \frac{1}{n}\, diag\left(a_{.1}, \dots . a_{.J}\right) \qquad (A.9)$$

Soit V la matrice d'inertie de nuage de points :

$$V = \frac{1}{S}\, T' T D^{-1} \qquad (A.10)$$

Où, T'est la matrice transposée de T.

Pour déterminer les axes factoriels, il s'agit de trouver les vecteurs propres U_j et les valeurs propres θ_J de cette matrice symétrique V (équation A.10). En d'autres termes, il s'agit de calculer les θ_J vérifiant :

$$\det\left(V - \theta_J I_J\right) = 0 \qquad (A.11)$$

Où I_J est la matrice unitaire d'ordre (J, J).

D'après (Tuffèry, 2007), il faut savoir que les valeurs propres sont non trivialement égales à 0 et 1, alors le nombre d'axes est égal à :

$$k = J - S \qquad (A.12)$$

Les k valeurs propres seront rangées dans une matrice diagonale M.

Alors dans le plan $R^{K,}$ l'équation du k-ième axe factoriel s'écrit :

$$\frac{1}{S}\, T' T\, D^{-1} U_k = \theta_k U_k \qquad (A.13)$$

La matrice des vecteurs propres aura la forme suivante :

$$U = [U_1 \dots U_r \dots U_k]$$

On obtient alors k vecteurs propres $\vec{U}_1, \dots \vec{U}_r, , \dots, \vec{U}_k$ qui forment une nouvelle base orthonormée de R^k. Ces vecteurs propres sont les vecteurs directeurs des différents axes factoriels, notés $\Delta_1, \dots \Delta_r, \dots, \Delta_k$, avec r=1,..,k.

Après avoir construit les axes factoriels, il s'agit désormais de déterminer les nouvelles coordonnées des éléments (ω_{ir}) dans le plan R^k, appelées composantes principales. En d'autres termes, il s'agit de déterminer les nouvelles coordonnées des éléments i sur chaque axe factoriel Δ_r du vecteur $\vec{U}_r$. Ainsi, la matrice des composantes principales des éléments (C_e) s'écrit :

$$C_e = T\, D^{-1} U \tag{A.14}$$

À pares avoir présenté les éléments dans le plan R^k, on va présenter les modalités dans le plan R^n. D'après Escofier et Pagés (1998), le calcul des axes des facteurs du nuage des modalités et absolument identique à celui du nuage des éléments. Alors, tous les résultats seront déduisent de ceux obtenus pour le nuage des éléments. Il suffit donc de représenter les vecteurs propres U dans le plan R^k pour déterminer les vecteurs propres dans le plan R^n.

Sur le plan R^n, la matrice des vecteurs propres s'écrit:

$$Z = T\, D^{-1} U\, M^{1/2} \tag{A.15}$$

Ainsi, la matrice des composantes principales des modalités (C_m) s'écrit :

$$C_m = D^{-1} T' Z \tag{A.16}$$

Une fois que les axes factoriels sont déterminés et le calcul des coordonnées des modalités est achevé, il est intéressant d'examiner la contribution de chaque modalité dans la construction de ces axes et de repérer celle ayant une forte contribution. Ceci sert à étudier les relations entre les différentes modalités (Escofier et Pagès, 1998; Tuffèry, 2010). La contribution d'une modalité j dans la construction d'un axe r est calculée comme suit.

$$\text{Contribution}\,(\mathrm{j}, \mathrm{r}) = \frac{1}{\theta_r} \frac{a_{.j}}{n * S} V_{jr} \tag{A.18}$$

Où V_{jr} et θ_r sont respectivement la coordonnée de la modalité j sur l'axe factoriel r et la valeur propre de l'axe factoriel r; avec j=1,..,J et r=1.,..,k. D'après Escofier et Pagès (1998), il faut

s’intéresser aux modalités ayant la contribution qui est supérieure au poids (le poids d’une modalité est déjà défini par la formule A.1).

ANNEXE B : METHODE DE CLASSIFICATION PAR PARTITIONNEMENT (K-MOYENNES)

L'idée principale de la méthode de partitionnement est de répartir n éléments sur un nombre de groupes donné noté k (Bramer, 2007). L'algorithme des k-moyennes est l'un des algorithmes de partitionnement les plus utilisés. D'après Ludovic (2008), il est difficile d'identifier avec certitude le premier utilisateur de cette méthode. En fait, il y a plusieurs chercheurs qui ont utilisés la méthode des k-moyennes dans leurs études sans faire une publication officielle des résultats obtenus. Par ailleurs, l'algorithme utilisé dans certains travaux a été imputé à MacQueen (1967).

- **Algorithme**

D'après Bramer (2007), l'algorithme de la méthode k-moyennes est le suivant.

- Initialisation : choisir la valeur k (combien de groupes les éléments doivent être partitionnés)
- Sélection : choisir arbitrairement les k éléments qui représentent les centres initiaux de k groupes.
- Affectation : assigner chaque élément restant au centre de groupe le plus proche.
- Mise à jour : calculer le nouveau centre de chaque groupe.
- Répétition : répéter les étapes « Affectation » et « Mise à jour », l'algorithme s'arrête lorsque tous les éléments gardent leurs groupes.

D'après Tuffèry (2010), il est possible d'ajouter un autre critère pour arrêter l'algorithme en fixant un nombre maximal d'itérations.

- **Métrique :**

La distance euclidienne peut être utilisée dans le processus du calcul de la distance entre un élément et un centre de groupe:

$$D(X, C_k) = \sqrt{\sum_{i=1}^{p} (x_i - c_{ki})^2} \qquad \text{(B.1)}$$

Où $X=x_1,x_2\ldots,x_P$ et $C_k= c_{k1},c_{k2}\ldots c_{kP}$ sont deux vecteurs de dimensions P et qui représentent respectivement les coordonnées de l'élément X et celles du centre de groupe C_k.

- **Evaluation de la qualité de répartition**

La méthode de k-moyennes consiste à regrouper les tronçons en minimisant la dispersion intragroupe et en maximisant la dispersion intergroupe. D'après Giudici (2003), la valeur de dispersion intragroupe s'écrit :

$$W = \sum_{k=1}^{K} W_k \qquad \text{(B.2)}$$

Avec W_k est la valeur de variation de k^{ème} groupe et qui est égale à la somme des carrés des écarts d'éléments au centre de leur groupe k.

$$W_k = \sum_{j=1}^{n_k} \sum_{i=1}^{P} (x_{ij} - C_{ki})^2 \qquad \text{(B.3)}$$

Où $C_k =[C_{k1},..,C_{ki},..,C_{kp}]$ est le centre du k^{éme} groupe, P est le nombre de variables d'un élément j et n_k est le nombre d'éléments dans le k^{ème} groupe.

La valeur de dispersion totale (T) est égale à la somme de la valeur de dispersion intragroupe W et intergroupe (BSS).

$$T = W + BSS \qquad \text{(B.4)}$$

$$BSS = \sum_{k=1}^{K} \sum_{i=1}^{p} n_k (x_{ik} - C_i)^2 \qquad \text{(B.6)}$$

Où $C=[C_1,\ldots,C_i\ldots,C_p]$ est le centre de tous les éléments et K est le nombre total de groupes.

D'après Guidici (2003), la qualité de la réparation est mesurée par R-carré qui varie entre 0 et 1:

$$\text{R-carré} = \mathbf{1} - \frac{\mathbf{W}}{\mathbf{T}} = \frac{\mathbf{BSS}}{\mathbf{T}} \qquad \text{(B.7)}$$

La qualité de partitionnement est bonne lorsque la valeur de R-carré s'approche de 1. En effet, les tronçons dans chaque groupe sont similaires (faible valeur de W) et les groupes sont hétérogènes entre eux (la valeur de dispersion intergroupe augmente). La qualité de partitionnement se dégrade lorsque la valeur de R-carré s'approche de 0. Il faut noter que

la valeur R-carré égale à 0 lorsque tous les éléments appartiennent à un seul groupe, et égale à 1 lorsque chaque élément représente un groupe.

Un autre critère peut être utilisé conjointement à avec l'indicateur R-carré, soit pseudo-F. Ce critère d'évaluation a été proposé par Calinski et Harabasz (1947) pour prendre la décision de fusion de deux groupes.

$$\text{Pseudo-F} = \frac{\mathrm{BSS}\,(\mathrm{K})/(\mathrm{K}-1)}{\mathrm{W}\,(\mathrm{K})/(\mathrm{n}-1)} \tag{B.8}$$

Où K est le nombre de groupes et n est le nombre total d'éléments.

ANNEXE C: ARBRE DE DECISION C4.5

L'arbre de décision permet de prédire l'appartenance d'un élément à un groupe en s'appuyant sur ses attributs. Une des caractéristiques les plus attirantes des arbres de décision réside dans leur interprétation, notamment la construction des règles de décision (Larose, 2005).

- **Processus général de construction d'un arbre de décision**

Han et Kamber (2001) ont présenté dans leur livre « Data Mining, concepts and Techniques » le processus général de construction d'un 'arbre de décision et qui est le suivant :

- L'arbre de décision commence par un nœud initial qui renferme tous les éléments de la base de données P (étape 1).
- Si les éléments de P appartiennent au même groupe, alors le nœud se transforme en feuille et sera étiqueté par ce groupe (étapes 2).
- Sinon, ce nœud sera découpé, d'une manière successive, en sous-ensembles. Ce découpage s'appuie sur des indicateurs de mesure de la qualité du partitionnement. Ces indicateurs visent à donner une note à un partitionnement; autrement dit, ils représentent un moyen pour identifier l'attribut qui permet une meilleure division des données. (étape 3)
- Pour chaque nouveau nœud intermédiaire, une branche sera ajoutée et une nouvelle évaluation sera faite (étapes 4).

À chaque répartition, l'algorithme utilise le même processus récursif pour continuer à construire l'arbre de décision.

D'après Han & Kamber (2001), la récursivité s'arrête dans les cas suivants :

- Lorsque les éléments dans un groupe sont homogènes ;
- s'il ne reste pas d'attributs pour faire la répartition ;
- s'il n'y a pas d'éléments avec la valeur d'attribut.

- **L'algorithme C4.5**

L'algorithme C4.5, développé par Quinlan (1993), prend en compte à la fois des variables continues et discrètes, des valeurs manquantes, des poids des observations, et il se caractérise par l'élagage qui facilite la décision finale (Devéze et Fouquin, 2004). Afin de sélectionner l'attribut permettant de faire un meilleur partitionnement, cet algorithme utilise le ratio du gain informationnel. Cet indicateur tient compte du nombre de valeurs d'une variable et de la proportion de ces valeurs dans les données. Il est égal à la différence entre le gain informationnel et l'information potentielle générée par la partition des éléments Ces deux indicateurs sont définis par Quinlan (1993) dans son livre «*Programme forMachine learning*».

Soit D un ensemble de données dans un nœud qui sera testé, C_i est la classe i avec i =1,…, m. D'après Han et Kamber (2006), la mesure de l'entropie de D s'écrit :

$$Info(D) = -\sum_{i=1}^{m} p_i \ \log_2(p_i) \qquad \text{(C.1)}$$

Avec p_i est la fréquence relative de la classe i dans l'ensemble de données D (pourcentage d'éléments de la classe i dans D).

Supposant que l'attribut A est choisi pour faire la division (avoir la valeur d'entropie la plus élevée) et qu'il peut prendre v valeurs distinctes (a_1, a_2,…, a_v). Cet attribut peut être utilisé pour repartir D en v sous-ensembles notés (D_1,..., D_j,…, D_v), où D_j est constitué par les données de D ayant la valeur a_j de l'attribut A. La mesure de l'entropie de D après sa division par l'attribut A est définie par la formule suivante :

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} - Info(D_j) \qquad \text{(C.2)}$$

La valeur $\frac{|D_j|}{|D|}$ représente le poids de la j^{éme} partition.

Le gain informationnel d'une partition en A est égal à la différence en la valeur de l'entropie de D avant et après la division de nœud par l'attribut A. Il est représenté par la formule suivante :

$$Gain = Info(D) - Info_A(D) \qquad \text{(C.3)}$$

L'information potentielle générée par la partition des éléments en A est calculée comme suit :

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2(\frac{|D_j|}{|D|}) \quad \text{(C.4)}$$

Il faut calculer pour chaque attribut le ratio du gain informationnel pour choisir finalement celui qui a la valeur la plus élevée. La formule de cet indicateur est la suivante:

$$RatioGain = gain(D) - SplitInfo_A(D) \quad \text{(C.5)}$$

Ainsi, un nœud sera créé et étiqueté par l'attribut choisi, et des branches contenant les valeurs de cet attribut seront construites.

ANNEXE D : ÉVALUATION ET COMPARAISON DES MODÈLES

Cette annexe consiste à définir la méthode de validation croisée (S sous-ensembles), la méthode de resubstitition et la matrice de confusion.

- La méthode de resubstitution et la méthode de validation croisée

D'après Bramer (2007), l'estimation de la précision d'un classificateur est une étape importante. Il s'agit de mesurer le degré de la précision d'un modèle à classifier correctement un élément non échantillonné dans un groupe. La première méthode qui peut être effectué en utilisant les données d'apprentissage, est la méthode de resubstitution (Feuilloy, 2009). Cette évaluation est optimiste, car elle consiste à mesurer la performance du modèle sur les données utilisées pour l'apprentissage (Nakache et Confais, 2003). Dans la littérature, il y a plusieurs méthodes qui estiment le taux d'erreur, telles que la méthode de validation croisée. Cette méthode est l'une des méthodes de mesure de la performance du modèle. D'après Nakache et Confais (2003), cette méthode est utilisée dans le cas où toute la base de données a été utilisée pour l'apprentissage. D'après Han & Kamber (2001) et Bramer (2007), la base de données totale est divisée, équitablement, en S sous-ensembles, notées $E_1,..,E_i,\ldots,E_S$. Parmi S itérations, chaque sous-ensemble sera testé une seule fois et appartiendra à la base d'apprentissage S-1 fois. En effet, dans la $i^{ème}$ itération, le sous-ensemble E_i représentera la base de données test et les S-1 sous-ensembles seront utilisés dans la phase d'apprentissage. L'erreur estimée finale (R) est donnée par la moyenne des erreurs mesurées (R_i) :

$$R = \frac{1}{S}\sum_{i=1}^{S} R_i \qquad (D.1)$$

- **La matrice de confusion**

Pour évaluer la performance d'un modèle, une matrice de confusion doit être construite permettant de comparer les classes réelles et les classe prédites par le modèle (Lefébure et Venturi, 1998). Elle contient le nombre d'éléments qui sont correctement classés ou mal classés pour chaque classe.

Tableau D.1: Matrice de confusion (adapté de Bramer, 2007)

		Classe prédite				
		$C_{1.pred}$	$C_{2.pred}$	$C_{j.pred}$	$C_{k.pred}$	Total
Classe réelle	$C_{1.réelle}$	n_{11}	n_{12}	n_{1j}	n_{1k}	
	$C_{2.réelle}$	n_{21}	n_{22}	n_{2j}	n_{2k}	
	$C_{i.réelle}$	n_{i1}	n_{i2}	n_{ij}	n_{ik}	
	$C_{k.réelle}$	n_{k1}	n_{k2}	n_{kj}	n_{kk}	
	Total					N

Où :

k est le nombre de classes et **N** est le nombre total d'éléments.

$C_{j.pred}$ est la $j^{éme}$ colonne de la matrice de confusion et qui représente la situation prévue.

$C_{i.réelle}$ est la $i^{éme}$ ligne et qui représente la situation réelle.

n_{ij} est la valeur dans la cellule (i,j) de la matrice de confusion et il indique le nombre d'éléments affectés dans la classe réelle $C_{i.réelle}$ et la classe prédite $C_{j.pred}$.

La somme des éléments diagonaux de cette matrice représente le nombre d'éléments qui sont correctement classés, noté N_{bien}. À partir de la matrice de confusion, il est possible de calculer le taux d'erreur qui est égal au rapport entre le nombre d'éléments mal classés, noté N_{mal}, et le nombre total d'éléments N.

$$\text{Taux d'erreur}(\%) = \frac{N_{mal}}{N} \tag{D.2}$$

$$N_{mal} = N - N_{bien}\,. \tag{D.3}$$

ANNEXE E : RECODAGE

Tableau E.1: Les codes des modalités

Champs	Modalités avant recodage	Modalités après recodage
Nbr_sorties	0	NS0
	1	NS1
	2	NS2
Nbr_entrées	0	NE0
	1	NE1
	2	NE2
Nbr_intersections	0	NI0
	1	NI1
	2	NI2
	3	NI3
	4	NI4
	5	NI>=5
	6	NI>=5
	7	NI>=5
sens_HC/VC	VC	VC
	HC	HC
Type_voie	Autoroute	AU
	Boulevard/Avenue	BV/AV
	Rue/Route	Rue/Route
Acc_D	Oui	Acc_D_Oui
	Non	Acc_D_Non
Tr_D	Oui	Tr_D_Oui
	Non	Tr_D_Non
Type_Barr_D	Barrière béton	Barrière béton_D
	Présence d'un mur	Mur_D
	Barrière métallique	Barrière métallique_D
	Absence de barrière	Absence de barrière_D
Acc_G	Oui	Acc_G_Oui
	Non	Acc_G_Non
Tr_G	Oui	Tr_G_Oui
	Non	Tr_G_Non
Type_Barr_G	Barrière béton	Barrière béton_G
	Présence d'un mur	Mûr_G
	Barrière métallique	Barrière métallique_G
	Absence de barrière	Absence de barrière_G
sur_pont	Non	Sur_p=non

	Oui	Sur_p=oui
Tunnel	Non	Tunnel=non
	Oui	Tunnel=oui
Nbr_voies	1	NV1
	2	NV2
	3	NV3
	4	NV4
Vittesse_aut	35km/h	moins de 50km/h
	45km/h	moins de 50km/h
	50km/h	moins de 50km/h
	60km/h	60km/h
	65km/h	70km/h
	70km/h	70Km/h
	80km/h	80Km/h
	90km/h	90Km/h
	100-60 km/h	100-60km/h