

Étude de cas:

Analyse de marché du transport aérien canadien avec R

Atelier d'introduction à R

CABRAL CRUZ, SAMUEL

Avec la collaboration de

BEAUCHEMIN, DAVID

GOULET, VINCENT

Dans le cadre du colloque R à Québec

25 mai 2017

© 2017 David Beauchemin, Samuel Cabral Cruz et Vincent Goulet



Cette création est mise à disposition selon le contrat [Attribution-Partage dans les mêmes conditions 4.0 International](#) de Creative Commons. En vertu de ce contrat, vous êtes libre de :

- ▶ **partager** — reproduire, distribuer et communiquer l'œuvre ;
- ▶ **remixer** — adapter l'œuvre ;
- ▶ utiliser cette œuvre à des fins commerciales.

Selon les conditions suivantes :



Attribution — Vous devez créditer l'œuvre, intégrer un lien vers le contrat et indiquer si des modifications ont été effectuées à l'œuvre. Vous devez indiquer ces informations par tous les moyens possibles, mais vous ne pouvez suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son œuvre.



Partage dans les mêmes conditions — Dans le cas où vous modifiez, transformez ou créez à partir du matériel composant l'œuvre originale, vous devez diffuser l'œuvre modifiée dans les même conditions, c'est à dire avec le même contrat avec lequel l'œuvre originale a été diffusée.

Table des matières

Table des figures	3
Liste des codes sources	5
Liste des tableaux	6
Préface	7
Introduction	8
Étude de cas	11
2.1 Extraction, traitement, visualisation et analyse des données	11
2.1.1 Extraction	11
2.1.2 Traitement	13
2.1.3 Visualisation et analyse des données	20
2.2 Création de fonctions utilitaires	26
2.3 Conception de graphiques en R	33
2.4 Outils d'analyse statistique en R	39
2.5 Ajustement de distributions statistiques sur données empiriques	45
2.6 Calcul stochastique en R	52
Conclusion	57
A Code source du projet	61

Table des figures

1.1	Interface de l'outil OpenFlights	8
2.1	Extrait du fichier airports.dat	12
2.2	Structure des fichiers de données géospatiales	16
2.3	Exemple de carte géographique produite avec <i>ggmap</i>	22
2.4	Exemple de carte géographique produite avec <i>leaflet</i>	23
2.5	Densité de la population canadienne	26
2.6	Passage de paramètres graphiques à la commande <i>plot</i>	35
2.7	Tracer une courbe avec la commande <i>plot</i>	36
2.8	Tracer une courbe avec la commande <i>curve</i>	37
2.9	Distribution des altitudes des aéroports canadiens	38
2.10	Représentation graphique de la fonction f3	48
2.11	Comparaison des résultats d'une analyse stochastique à 6 réplicats . .	54
2.12	Classement <i>RedMonk</i> des différents langages de programmation	57

Liste des codes sources

1.1	Environnement de travail	10
2.1	Extraction des données	13
2.2	Filtrer les données	14
2.3	Traitement standard de données géospatiales en R	17
2.4	Exemple de requête SQL	18
2.5	Fonctionnalités avancées de SQL	19
2.6	Fonctionnalités avancées de SQL	19
2.7	Fonctions de visualisation de données	21
2.8	Générer une carte du trafic aérien avec <i>ggmap</i>	21
2.9	Générer une carte du trafic aérien avec <i>leaflet</i>	22
2.10	Structure pour la définition d'une fonction	26
2.11	L'instruction <i>return</i> et le retour standard d'une fonction R	27
2.12	Définir des valeurs par défauts dans les fonctions utilitaires	27
2.13	Passage d'arguments à une fonction	29
2.14	L'assignation et les valeurs par défaut	30
2.15	Retour multiple par l'entremise d'une liste	30
2.16	Gestion des erreurs sous R	31
2.17	Utilisation de la commande <i>plot</i>	34
2.18	Utilisation de la commande <i>curve</i>	34
2.19	<i>hist</i> , <i>density</i> , <i>lines</i> , <i>abline</i> , <i>legend</i> et <i>mtext</i>	38
2.20	Fonctions relatives à la distribution Normale	40
2.21	Génération de nombres aléatoires	41
2.22	Fonctions de densité et de répartition empiriques	42
2.23	Tests d'indépendance et de corrélation entre distributions	43
2.24	Régression linéaire sur données empiriques	44
2.25	Optimisation générique avec R	46
2.26	Maximisation d'une fonction avec <i>optim</i>	47
2.27	Ajustement de distribution sur données empiriques	49
2.28	Réplicat maison de la fonction <i>fitdistr</i>	49
2.29	Exemple d'utilisation de la fonction <i>distFit</i>	52
2.30	Pige aléatoire sur support vectoriel	53
2.31	Replication d'une analyse stochastique	54
A.1	Benchmark.R	61
A.2	CaseStudyDevQ1.R	63
A.3	CaseStudyDevQ2.R	68
A.4	CaseStudyDevQ3.R	73
A.5	CaseStudyDevQ4.R	74
A.6	CaseStudyDevQ5.R	75

A.7 CaseStudyDevQ6.R	79
--------------------------------	----

Liste des tableaux

2.1	Liste des distributions statistiques disponibles en R	39
2.2	Comparaison entre les coefficients réels et estimés par régression linéaire	45

Préface

Dans le cadre du colloque "R à Québec" qui se tiendra le 25 et 26 mai 2017 sur le campus de l'Université Laval, une séance d'introduction au langage de programmation R sera offert aux participants. Cette séance vise principalement la compréhension et la pratique permettant de maîtriser les rudiments de cet environnement de programmation. [13] Cette séance sera divisée en deux parties. En ce qui concerne la première partie, les fondements du langage seront visités d'une manière théorique sous la forme d'un exposé magistral. La deuxième partie, tant qu'à elle, se concentrera davantage sur la mise en pratique des notions abordées lors de la première partie grâce à la complétion d'une étude de cas cherchant à faire l'analyse de marché du transport aérien canadien. Ce document correspond en fait à la documentation complète de cette deuxième partie de formation.

Étant donné qu'il s'agit tout de même d'une formation pour débutants, la majorité du code sera déjà fournie, mais il n'en vaut pas moins la peine de parcourir ce projet si ce n'est que pour constater la puissance et la simplicité du langage. De plus, il est souvent difficile de mettre en perspective les innombrables fonctionnalités d'un langage lorsque nous commençons à l'utiliser. Cette étude de cas nous fournit ainsi un bel exemple d'enchaînement de traitements jusqu'à l'aboutissement ultime qui consiste à répondre aux questions que nous nous posions avant même d'amorcer l'analyse.

D'autre part, il est important de préciser que le code qui sera présenté ne correspond pas toujours à la manière la plus efficiente d'accomplir une tâche donnée. L'objectif principal étant ici la transmission de connaissances dans un dessin éducatif plutôt que d'une réelle analyse de marché. Nous tenons aussi à mentionner que bien qu'il s'agisse d'une formation s'adressant à des débutants, plusieurs notions qui seront mises en valeur font plutôt état de niveau intermédiaire et avancé, mais apporté toujours de manière simplifiée et accessible à quiconque qui n'aurait jamais travaillé avec R.

Nous tenons à remercier Vincent Goulet de nous avoir fait confiance dans l'élaboration de cette partie de la formation ainsi que tous les membres du comité organisationnel de l'évènement. Nous croyons sincèrement que R est un langage d'actualité qui se révèle un atout à quiconque oeuvrant dans un domaine relié de près ou de loin aux mathématiques.

Introduction

Dans le cadre de cette étude de cas, nous nous placerons dans la peau d’un analyste du département de la tarification oeuvrant au sein d’une compagnie canadienne se spécialisant dans le transport de colis par voies aériennes en mettant à profit le jeu de données d’*OpenFlights*. [8]



FIGURE 1.1 – Interface de l’outil OpenFlights

Parmi les bases de données disponibles, nous retrouvons :

- airports.dat** Données relatives à tous les aéroports du monde [25]
- routes.dat** Données relatives à tous les trajets possibles dans le monde [24]

Ainsi, notre mandat consistera, dans un premier temps, à analyser les bases de données mises à notre disposition afin de créer des fonctions utilitaires qui permettront de facilement intégrer les informations qu’elles contiennent lors de la tarification d’une livraison spécifique. Une fois cette tarification complétée, nous devrons fournir des chartes pour facilement estimer les prix d’une livraison qui s’avèreront être des outils indispensables au département de marketing et au reste de la direction. Après avoir transmis les documents en question, votre gestionnaire voulant s’assurer que la nouvelle tarification sera efficace et profitable vous demandera d’analyser les prix de la concurrence pour en extrapoler leur tarification. Finalement, vous serez appelé à comparer ces deux tarifications et la compétitivité de votre nouvelle tarification com-

parativement au reste du marché en procédant à une analyse stochastique.



OpenFlights

OpenFlights est un outil en ligne permettant de visualiser, chercher et filtrer tous les vols aériens dans le monde. Il s'agit d'un projet libre entretenu par la communauté via GitHub. [5] L'information disponible y est étonnamment très complète et facile d'approche. Ce qui en rend ce jeu de données très intéressant pour quiconque qui désire s'initier à l'analyse statistique.
<https://openflights.org/>

Bien qu'on n'en soit toujours qu'à l'introduction, nous tenons dès lors à introduire des notions de programmation qui comparativement à celles qui suivront sont d'autre un peu plus général. Tout d'abord, afin de maximiser la portabilité des scripts que vous créerez dans le futur, il est important de rendre votre environnement de travail indépendant de la structure de dossier dans laquelle il se trouve. Pour ce faire, nous devons donc utiliser le principe de liens relatifs plutôt qu'absolus. En R, deux fonctions bien spécifiques nous fournissent les outils afin de rendre cette tâche possible. Il s'agit de *getwd* et *setwd* [34]. Comme leurs noms l'indiquent, elles servent respectivement à extraire le chemin vers l'environnement de travail et à le modifier.

De manière similaire qu'au sein d'un invité de commandes traditionnel, il est possible d'utiliser "." (cd ..) afin de revenir à un niveau supérieur dans la structure de dossiers. Dans la plupart des cas, le code source d'un projet sera souvent isolé du reste du projet en le plaçant dans un sous-dossier dédié.¹

Bref, comme le code source du présent projet se retrouve à l'intérieur du sous-dossier *dev* [7] et que nous pourrions vouloir avoir accès à d'autres parties du répertoire au sein du code, le code suivant nous permettra de placer notre racine de projet à un niveau de dossier supérieur et de stocker ce chemin dans la variable *path*. Avec cette variable, tous les appels subséquents à des portions du répertoire pourront donc se faire de manière relative puisque c'est cette variable *path* qui changera d'une architecture à un autre, tandis que la structure du répertoire restera toujours la même.

La deuxième notion que nous tenons à introduire immédiatement est celle de reproductibilité d'une analyse statistique. Comme vous le savez probablement, l'aléatoire pur n'existe pas en informatique, d'où la raison pour laquelle nous utiliserons plutôt le terme de nombres pseudo-aléatoires. Bien que cela peut sembler étrange à première vue, il existe tout de même un point positif à tout ceci, soit la possibilité de fixer une racine au générateur de nombre pseudo-aléatoires (GNPA) ce qui aura comme impact de toujours produire les mêmes résultats pour autant que le GNPA utilisé soit le même. Comme nous pouvons le voir dans le code ci-dessous, l'instruction *set.seed* [42] se chargera de fournir une valeur de départ aux calculs du GNPA.

1. Il s'agit ici d'une excellente pratique de programmation et je dirais même indispensable si vous utilisez un gestionnaire de versions.

Code Source 1.1 – Environnement de travail

```
1 ##### Setting working directory properly #####  
2 setwd('.')  
3 path <- getwd()  
4 set.seed(31459)
```



Code source du projet

Le code source du projet se retrouve dans son intégrité en annexe à ce document. N'hésitez pas à vous y référer au besoin.

Étude de cas

2.1 Extraction, traitement, visualisation et analyse des données

Cette section est certainement la plus importante de toutes, elle vise à faire un traitement adéquat et pertinent des données afin de pouvoir les réutiliser facilement dans les sections suivantes. Une mauvaise application des concepts d'extraction, de traitement et de visualisation des données peut entraîner des interprétations aberrantes des phénomènes que nous cherchons à analyser.

2.1.1 Extraction

Les données d'OpenFlights possèdent l'avantage d'être téléchargeables directement via le web pour les rendre disponibles à notre environnement de travail. Pour ce faire, nous mettons à profit la fonction *read.csv* [29]. Bien que le nom de la fonction indique qu'elle permet de lire un fichier présenté dans un format *.csv*, nous pouvons tout aussi bien utiliser cette fonction pour extraire des fichiers *.dat*. La différence principale entre ces deux types de fichiers et que les fichiers *.csv* utilisent un caractère d'encadrement des informations qui se trouve à être les doubles guillemets dans la majorité des cas. De plus, les fichiers *.csv* utiliseront comme leur nom l'indique la virgule à titre de séparateur bien que celui-ci puisse être modifié pour un symbole différent.[2] Lorsque nous jetons un coup d'oeil à la structure des fichiers *.dat* disponibles à la Figure 2.1, nous constatons que ceux-ci respectent à la fois les deux caractéristiques que nous venons de mentionner rendant ainsi l'utilisation de la fonction *read.csv* si naturelle.

Dans la même figure, on constate aussi l'absence d'une ligne servant à présenter les en-têtes de colonnes. Ceci pourra dans certains cas vous jouer de mauvais tours en ignorant la première ligne de données ou encore de considérer les titres comme étant des entrées en soi.¹ Bien qu'il serait possible de travailler avec des données sans nom, il s'agit ici d'une très mauvaise pratique. Pour remédier à la situation, nous assignerons donc des noms aux colonnes grâce à l'attribut *colnames* d'un objet *data.frame* en lui passant un vecteur de noms.

Par défaut, lors de l'importation, la fonction *read.csv* retournera un *data.frame* en transformant les chaînes de caractères sous la forme de facteurs (*factors*). Cette action sera complètement transparente à l'utilisateur puisque l'affichage des variables ne sera

1. La deuxième situation étant bien moins dramatique et plus facilement identifiable.

```

1,"Goroka Airport","Goroka","Papua New Guinea","GKA","AYGA",-6.081689834590001,145.391998291,5282,
2,"Madang Airport","Madang","Papua New Guinea","MAG","AYMD",-5.20707988739,145.789001465,20,10,"U"
3,"Mount Hagen Kagamuga Airport","Mount Hagen","Papua New Guinea","HGU","AYMH",-5.826789855957031,
4,"Nadzab Airport","Nadzab","Papua New Guinea","LAE","AYNZ",-6.569803,146.725977,239,10,"U","Pacif
5,"Port Moresby Jacksons International Airport","Port Moresby","Papua New Guinea","POM","APPY",-9.4
6,"Wewak International Airport","Wewak","Papua New Guinea","WWK","AYWK",-3.58383011818,143.6690063,
7,"Narsarsuaq Airport","Narsarsuaq","Greenland","UAK","BGBW",61.1604995728,-45.4259986877,112,-3
8,"Godthaab / Nuuk Airport","Godthaab","Greenland","GOH","BGGH",64.19090271,-51.6781005859,283,-3,
9,"Kangerlussuaq Airport","Sondrestrom","Greenland","SFJ","BGSF",67.0122218992,-50.7116031647,165,
10,"Thule Air Base","Thule","Greenland","THU","BGTU",76.5311965942,-68.7032012939,251,-4,"E","Amer
11,"Akureyri Airport","Akureyri","Iceland","AEY","BIAR",65.66000366210938,-18.07270050048828,6,0,"I
12,"Egilsstaðir Airport","Egilsstaðir","Iceland","EGS","BIEG",65.2833023071289,-14.401399612426758
13,"Hornafjörður Airport","Hofn","Iceland","HFN","BIHN",64.295601,-15.2272,24,0,"N","Atlantic/Reyk
14,"Húsavík Airport","Húsavík","Iceland","HZK","BIHU",65.952301,-17.426001,48,0,"N","Atlantic/Reyk
15,"Ísafjörður Airport","Ísafjörður","Iceland","IFJ","BIIS",66.05809783935547,-23.135299682617188,

```

FIGURE 2.1 – Extrait du fichier airports.dat

pas impacté étant donné que R aura créé des formats d’affichage qui associe à chaque facteur la valeur unique correspondante. Le seul impact réel réside dans la possibilité d’utiliser des fonctions à caractères mathématiques sur les données, peu importe si ces dernières sont numériques ou non. Parmi ce genre de fonctions, nous pouvons penser à des fonctions d’agrégation (*clustering*) ou tout simplement à l’utilisation de la fonction *summary* [40] permettant d’afficher des informations génériques sur le contenu d’un objet. Il est important de comprendre que les données ne sont toutefois plus représentées comme des chaînes de caractères, mais bien pas un indice référant à la valeur textuelle correspondante.

La manière de représenter des valeurs manquantes variera souvent d’une base de données à une autre. Une fonctionnalité très intéressante de la fonction *read.csv* est de pouvoir automatiquement convertir ces chaînes de caractères symboliques en *NA* ayant une signification particulière dans R. Dans le cas présent, les valeurs manquantes sont représentées par `\\n` ou `" "` correspondant à un simple retour de chariot et un espace vide respectivement. Il suffit donc de passer cette liste de valeurs à l’argument *na.strings*.



read.csv

La fonction *read.csv* possède plusieurs autres arguments très intéressants dans des situations plus pointues. Pour en savoir plus, nous vous invitons à consulter la documentation officielle.
<https://stat.ethz.ch/R-manual/R-devel/library/utils/html/read.table.html>

Comme nous venons de le démontrer, l’extraction des données peut facilement devenir une tâche ingrate si nous n’avons aucune connaissance sur la manière dont l’information y a été entreposée. La règle d’or est donc de toujours avoir une idée globale de ce que nous cherchons à importer afin de bien paramétrer les fonctions.

Si nous assemblons les différents aspects que nous venons d’aborder, nous aboutissons donc au code suivant :

Code Source 2.1 – Extraction des données

```
1 airports <- read.csv("https://raw.githubusercontent.com/jpatokal/  
  openflights/master/data/airports.dat", header = FALSE, na.  
  strings=c('\n',''))
```

2.1.2 Traitement

Une fois en possession du jeu de données, il fut nécessaire de nettoyer ce dernier pour en rendre son utilisation plus simple selon nos besoins. Parmi les différentes modifications apportées, nous retrouvons :

- Conserver que les observations relatives aux aéroports Canadiens.
- Filtrer les variables qui seront pertinentes dans le cadre de l’analyse que nous menons.²
- Alimentation des valeurs manquantes avec des sources de données externes (si possible) ou appliquer un traitement approximatif justifiable en documentant les impacts possibles sur le restant de l’analyse.

Nous considérons pertinent d’apporter quelques précisions sur le fonctionnement de R avant d’explicitier les traitements susmentionnés. Tout d’abord, R est un langage interprété orienté objet à caractère fonctionnel optimisé pour le traitement vectoriel. Ces simples mots ne sont pas à prendre à la légère puisque ce n’est qu’en s’appropriant ce mode de penser que les futurs développeurs que vous êtes parviendront à utiliser R dans toute sa puissance, sa simplicité et son élégance. Par sa sémantique objet, R permet de définir des attributs aux objets créés. Comme il sera possible de le voir plus loin, l’accès à ces attributs se fera à l’aide de l’opérateur `$`. Vous vous demandez probablement : Comment savoir si nous sommes en présence d’un objet ? C’est simple, tout dans R est un objet ! Le langage R permet aussi de mimer le paradigme fonctionnel puisque les fonctions (qui sont en fait des objets) sont des valeurs à part entière qui peuvent être argument ou valeur d’une autre fonction. De plus, il est possible de définir des fonctions dites anonymes qui se révéleront très pratiques. Finalement, par son caractère vectoriel, la notion de scalaire n’existe tout simplement pas en R. C’est pour cette raison que l’utilisation de boucles est à proscrire (ou du moins à minimiser le plus possible). En effet, l’utilisation d’une boucle revient en quelque sorte à la création d’un nouveau vecteur et à la mise en place de processus itératifs afin d’exécuter la tâche demandée. Heureusement, par un raisonnement vectoriel, il est très simple de convertir ces traitements sous une forme vectorielle dans la plupart des cas. [6] Pour accéder à une valeur précise d’un vecteur, nous utiliserons l’opérateur `[]` en spécifiant les indices correspondants aux valeurs désirées, un vecteur booléen d’inclusion/exclusion ou encore un vecteur contenant les noms des attributs nommés qui nous intéressent.

2. On ne devrait jamais travailler avec des informations superflues. Faire une pré-sélection de l’information ne fait qu’alléger les traitements et augmente de manière significative la compréhensibilité du programme.

Avec ces outils en mains, il devient ainsi très facile de filtrer les aéroports canadiens à l'aide de l'attribut que nous avons nommé *country* du *data.frame* *airports*. Par un raisonnement connexe, la fonction *subset* [46] nous offre aussi la possibilité de conserver que certaines variables contenues dans une table tout en appliquant des contraintes sur les observations à conserver. Le ?? dévoile au grand jour la dualité qui peut exister entre la multitude des fonctionnalités présentent en R.

Code Source 2.2 – Filtrer les données

```
1 airportsCanada <- airports[airports$country=='Canada',]
2 airportsCanada2 <- subset(airports, country == 'Canada')
3 all.equal(airportsCanada, airportsCanada2)
4
5 airportsCanada[is.na(airportsCanada$IATA), c("airportID", "name", "
6   IATA", "ICAO")]
7 subset(airportsCanada, is.na(IATA), select = c("airportID", "name",
8   "IATA", "ICAO"))
```

Nous ne devons pas être surpris qu'il y ait autant de possibilités différentes de parvenir au même résultat. Il s'agit là d'une des principales caractéristiques d'un logiciel libre, puisque la responsabilité du développement continu ne dépend plus que d'une seule personne ou entité, mais bien de la communauté d'utilisateurs au complet. Ceci peut toutefois sembler mélangeant pour des nouveaux utilisateurs et la question suivante arrivera assez rapidement lorsque vous commencerez à développer vos propres applications : Quelle est la meilleure manière d'accomplir une tâche X ? La bonne réponse est tout aussi décevante que la prémisse étant donné que chaque fonction aura été développée dans un besoin précis si ce n'est que de rendre l'utilisation de fonctionnalité de base plus aisée et facile d'approche... C'est pourquoi nous conseillons plutôt d'adopter un mode de pensée itératif, créatif et ouvert qui consiste à utiliser les fonctions qui vous semblent, à la fois, les plus simples, les plus versatiles et les plus efficaces. À partir du moment où vous constaterez qu'une de ces trois caractéristiques n'est plus au rendez-vous, il suffira d'amorcer des recherches pour bonifier vos connaissances et améliorer vos techniques. C'est un peu le but de ce document de vous faire faire une visite guidée pour vous offrir un coffre d'outil qui facilitera vos premiers pas en R.



subset

Bien que la fonction *subset* simplifie énormément l'écriture de requêtes afin de manipuler des bases de données, celle-ci souffre par le fait même de devenir rapidement inefficace lors de traitements plus complexes. D'autres packages tels que *dplyr* et *sqldf* deviendront dans ces situations des alternatives beaucoup plus efficaces.

<https://www.rdocumentation.org/packages/raster/versions/2.5-8/topics/subset>

Après avoir fait une présélection des données qui nous seront utiles dans le reste

de l'analyse, nous avons constaté que certaines variables n'étaient pas complètes. Tout d'abord, la variable IATA n'était pas toujours définie pour tous les aéroports canadiens contrairement à la variable ICAO. Étant donné la faible proportion des valeurs manquantes et du fait qu'une valeur fictive n'aurait qu'un impact minimal dans le cas de l'analyse, nous avons décidé de remplacer les valeurs manquantes par les 3 dernières lettres du code ICAO. En regardant les aéroports canadiens possédant les deux codes, nous observons que cette relation est respectée dans plus de 80% des cas. Une autre alternative aurait été de simplement prendre le code ICAO, mais le code IATA semblait beaucoup plus universel puisqu'il s'agit du code communément utilisé pour le transport des particuliers.

Le vrai problème au niveau des données résidait davantage dans l'absence d'informations sur les fuseaux horaires de certains aéroports ainsi qu'un accès indirect à la province de correspondance de tous les aéroports. Heureusement, ce genre d'information ne dépend que de l'emplacement de l'entité dans le monde, ce qui rend la tâche beaucoup plus simple lorsque nous avons accès aux coordonnées géospatiales.



Adresses et coordonnées géospatiales

Dans la situation où seule l'adresse de l'entité aurait été disponible, nous aurions été contraints d'utiliser des techniques de géocodage qui permettent de transformer une adresse en coordonnées longitude/latitude et parfois même altitude. Ce genre de transformation est devenu beaucoup plus accessible avec l'avancement de la technologie et plusieurs APIs sont disponibles gratuitement sur le web pour procéder à ce genre de transformation. Encore une fois, il vaut mieux bien se renseigner pour identifier l'interface qui répondra le mieux à nos besoins en considérant notamment :

- ▶ Format de l'intrant
- ▶ Format de retour
- ▶ Limitation du nombre de requêtes sur une période de temps donnée
- ▶ Efficacité de l'outil
- ▶ Méthode d'interpolation
- ▶ Précision des valeurs

<https://www.programmableweb.com/news/7-free-geocoding-apis-google-bing-yahoo-and-mapquest/2012/06/21>

Bien que nous savons qu'il est possible de combler les valeurs manquantes à l'aide de données géographiques encore faut-il disposer de ses dites données. Encore une fois, grâce à de bonnes recherches vous parviendrez à trouver une source qui contiendra ce dont vous cherchez ou du moins un élément de réponse qui vous permettra d'en extrapoler la valeur ce qui sera déjà préférable à des données manquantes. Statistiques Canada possède une bibliothèque géographique très garnie et c'est notamment sur

leur site que nous avons pris le fichier *.shp* qui définit les provinces et territoires du Canada. [19] En ce qui concerne les fuseaux horaires, nous avons trouvé ceux-ci sur un site dédié à cette fin qui mentionne ne plus être maintenu à jour, mais dont la dernière mise à jour a été faite le 28 mai 2016. Étant donné que les fuseaux horaires n'ont pas tendance à changer souvent dans les pays industrialisés comme le Canada, ceci ne consistait pas en un enjeu majeur. [23]



ArcGIS et les fichiers *.shp*

Le premier fichier ayant l'extension *.shp* fut créé dans le but d'être utilisé conjointement avec la suite de logiciel ArcGIS. Il s'agit de la première suite logicielle commercialisable visant le traitement des données géospatiales. Étant des pionniers dans le domaine, plusieurs aspects des outils visant à faire des traitements géospatiaux proviendront directement de leurs travaux. Les fichiers *.shp* sont aujourd'hui vue comme un standard pour transporter ce type de données.

<https://www.arcgis.com/features/index.html>

Pour être en mesure de bien travailler avec ce genre de fichier, nous devons en comprendre leur fonctionnement. Tout d'abord, lorsque vous téléchargerez un *.zip* de données géospatiales, vous devrez toujours obtenir la structure suivante de fichiers :





Name	Date modified	Type	Size
 gpr_000b11a_e.dbf	5/13/2017 10:48 PM	DBF File	2 KB
 gpr_000b11a_e.prj	5/13/2017 10:48 PM	PRJ File	1 KB
 gpr_000b11a_e.shp	5/13/2017 10:48 PM	SHP File	53,066 KB
 gpr_000b11a_e.shx	5/13/2017 10:48 PM	SHX File	1 KB

FIGURE 2.2 – Structure des fichiers de données géospatiales

Tel qu'illustré à la Figure 2.2, un dossier de données géospatiales se divisera minimalement sous la forme de quatre fichiers :

- .shp*** Contient l'information géographique représentée sous la forme de points, segments et/ou polygones.
- .dbf*** Contient l'information rattachée à toutes les entités définies dans le *.shp*.
- .prj*** Contient les informations sur la projection associée (le modèle mathématique permettant d'interpréter les informations du *.shp* [11]).
- .shx*** Contient les index des enregistrements du *.shp*.

Cette structure peut donner l'impression que leur utilisation conjointement avec R sera compliquée, mais c'est loin d'être le cas grâce aux paquets *rgdal* [18] et *sp*[26]. Pour conclure sur ce point, notons que la désignation *ShapeFile* au sens large désigne

l'ensemble de la structure de fichier et non pas seulement le *.shp* lui-même. [1]

Le paquetage *rgdal* n'aura qu'une utilité bien précise, soit celle d'aller extraire les informations contenues dans le *ShapeFile*. Cependant, il possède des dépendances directement dans le paquetage *sp* ce qui explique pourquoi le seul appel de *rgdal* entraîne du même coup l'appel de *sp*. Les rôles de *sp* sont plutôt de transformer les informations des objets R sous une forme compatible au *ShapeFile* que nous aurons lu. Notez bien la transformation de la projection sous une base commune en passant ainsi de *NA* vers

```
" +proj=longlat"
```

(projection choisie en fonction des données contenues) à

```
" +proj=longlat +datum=NAD83 +no_defs +ellps=GRS80 +towgs84=0,0,0"
```

soit la projection du *ShapeFile*. La nécessité que nos points soit sous la même projection que celle du *ShapeFile* provient du fait que nous voulons superposer ces derniers pour ensuite en extraire l'information correspondante. Les deux fonctions indispensables ici sont *CRS* qui retourne un objet de classe *Coordinate Reference System* à partir d'une chaîne de caractère passée en argument et *over* qui se chargera de faire la superposition de points géométriques sur une couche (correspondant ici au *ShapeFile* vu selon une certaine projection) qui contient les attributs envers lesquels nous avons un intérêt. Le retour de la fonction *over* sera finalement un *data.frame* de même longueur que le nombre de points donnés en argument que nous pourrions facilement combiner le jeu de données initial. Cette recette ne risque pas de varier beaucoup d'un *ShapeFile* à un autre, vous pourrez donc littéralement reprendre le code ci-dessous.

Code Source 2.3 – Traitement standard de données géospatiales en R

```
1  # Step 1 – Import the Packages
2  library(sp)
3  library(rgdal)
4  # Step 2 – Read the ShapeFile
5  prov_terr.shape <- readOGR(dsn=paste(path,"/Reference/prov_terr",
6  sep=""), layer="gpr_000b11a_e")
7  # Step 3 – Create the Spatial Points to be overlaid
8  unknown_prov <- airportsCanada[,c("airportID", "city", "longitude",
9  "latitude")]
10 sppts <- SpatialPoints(unknown_prov[,c("longitude", "latitude")])
11 # Step 4 – Set the Spatial Points on the same projection as the
12   ShapeFile
13 proj4string(sppts) <- CRS("+proj=longlat")
14 sppts <- spTransform(sppts, proj4string(prov_terr.shape))
15 # Step 5 – Extract the Desired Information by overlaying the
16   Spatial Points on the ShapeFile
17 merged_prov <- cbind(airportsCanada, over(sppts, prov_terr.shape))
```

Maintenant que nous disposons de l'information requise pour compléter notre base de données, nous devons combiner la table primaire avec les sous-tables créées lors de nos extractions et refaire un dernier filtre final pour se débarrasser de tout ce qui ne sera plus utile. Bien que les fonctionnalités de base de R vous permettraient d'accomplir la tâche, nous profitons de cette étape du processus pour vous présenter les

paquetages *sqldf* [21] et *dplyr* [47].

Le langage SQL (*Structured Query Language*) fut inventé en 1974 et ce dernier fut normalisé en 1986 devenant ainsi un standard dans l’exploitation des bases de données relationnelles. Devenir familier avec les langages normalisés tels que le SQL ne peut qu’être à votre avantage puisque ceux-ci vous permettront d’écrire des tronçons de code qui pourront facilement être transportés avec peu de modifications d’un environnement à un autre. Leur caractère normalisé impose aux environnements voulant respecter les standards de l’industrie d’être en mesure de compiler ces instructions bien qu’il y ait des fonctionnalités permettant de répliquer leur comportement ou du moins offrir un paquetage permettant leur interprétation. [15] Nous conseillons fortement à tous les analystes de données de s’approprier les rudiments du SQL très tôt dans leur cheminement en raison de sa simplicité et sa flexibilité. Les requêtes SQL sont habituellement constituées des quatre instructions suivantes :

Select	Déclare les variables que nous voulons conserver
From	Indique la source des données
Where	Mentionne les conditions que les observations doivent respecter pour se retrouver dans l’extrait
Order by	Spécifie la manière de trier l’extrait

La syntaxe rudimentaire rend sa compréhension presque immédiate, et ce, même à des personnes ignorant même qu’il s’agit en fait d’une requête SQL. Dépendamment des noms de variables contenues dans les relations exploitées, les requêtes peuvent parfois se lire aussi bien qu’une liste d’épicerie écrite en anglais. Le [Code Source 2.4](#) fournit un exemple de l’utilisation du langage SQL avec R rendu disponible par le paquetage *sqldf*.

Code Source 2.4 – Exemple de requête SQL

```
1 library(sqldf)
2 sqldf("SELECT name,IATA,altitude,province
3       FROM airportsCanada
4       WHERE province = 'New Brunswick'
5       ORDER BY name")
```

En nous fiant à la requête ci-dessus, nous pourrions la transformer de manière textuelle sous la forme suivante :

1. Sélectionne les variables *name*, *IATA*, *altitude* et *province*
2. Dans la relation(table) *airportsCanada*
3. Dont la province est *New Brunswick*
4. En triant le tout par *name*

Toutefois, les fonctionnalités de SQL ne s’arrêtent pas ici. Grâce à des instructions très compactes, nous pourrions rendre le comportement de la requête bien plus complexe. Parmi les fonctionnalités qui font partie de notre quotidien, nous retrouvons * qui lorsqu’utiliser au sein de l’instruction *select* permettra d’extraire l’ensemble des variables de la relation sans avoir à les écrire une à la fois. La fonction *coalesce* servira à extraire la première valeur non manquante parmi une liste de variables fournie

en argument. Nous attirons au passage votre attention sur le mot clé *as* qui a pour effet d'attribuer un nom à l'expression sous-jacente. Finalement, le bon vieux *left join* rendant si simple la fusion conditionnelle de deux tables en conservant toutefois les observations de la relation mère malgré le fait qu'il n'y ait pas eu correspondance dans la table à fusionner. Les conditions de cette fusion seront explicitées avec l'instruction *on* qui n'aura pas de signification tangible sans la présence de *join*. Le [Code Source 2.5](#) présente une requête combinant toutes ces fonctionnalités que vous serez en mesure de retrouver dans le code source du projet.

Code Source 2.5 – Fonctionnalités avancées de SQL

```
1 airportsCanada <- sqldf("
2   SELECT
3     a.*,
4     COALESCE(a.tzFormat,b.TZID) AS tzMerged,
5     c.PRENAME AS provMerged
6   FROM airportsCanada a
7   LEFT JOIN merged_tz b
8     ON a.airportID = b.airportID
9   LEFT JOIN merged_prov c
10     ON a.airportID = c.airportID
11  ORDER BY a.airportID")
```

Il serait faux de dire que ceci correspond à une bonne introduction à SQL sans parler de la capacité d'imbriquer des requêtes SQL. C'est à ce moment que toute la puissance du langage se révèle à nous. Le [Code Source 2.6](#) montre un exemple standard d'imbrication qui a été exploité pour créer la relation *routesCanada* en ne conservant que les routes aériennes empruntées pour les vols internes au Canada.^{3 4}

Code Source 2.6 – Fonctionnalités avancées de SQL

```
1 routesCanada <- sqldf("
2   SELECT *
3   FROM routes
4   WHERE sourceAirportID IN (SELECT DISTINCT airportID
5                             FROM airportsCanada)
6     AND destinationAirportID IN (SELECT DISTINCT airportID
7                                  FROM airportsCanada)")
```

3. Le mot clé *DISTINCT* spécifie de ne conserver qu'une seule observation pour chaque modalité retrouvée

4. L'utilisation de la case dans les exemples n'a été utilisée que pour bien faire la différence entre les instructions SQL des informations spécifiques aux relations traitées. Le SQL n'est pas sensible à la case.



Structured Query Language (SQL)

Le langage SQL regorge de plusieurs autres fonctionnalités qui ne seront pas abordées dans ce document. Parmi ces dernières, nous retrouvons *GROUP BY*, *HAVING*, les fonctions d'agrégation numérique tel quel *SUM*, *AVG*, *MIN*, *MAX*, etc. et nous pourrions continuer ainsi encore longtemps.

<https://www.w3schools.com/sql/>

Avant de passer à la prochaine section, il serait injuste de présenter *squdf* avec autant de précisions sans toucher un mot sur les paquetages *plyr* et *dplyr*. Ces derniers visent à reproduire les opérations permises par le langage SQL avec une notation aussi simpliste, mais en optimisant ces opérations en tenant compte du fonctionnement intrinsèque de R, soit le calcul matriciel. Une différence majeure avec le SQL provient du mode de pensée se rapprochant davantage d'un mode procédural pour *plyr* que du mode fonctionnel pour le SQL. Ces packages deviendront des outils très pertinents lorsque vous commencerez à faire face à des problèmes de temps d'exécution irraisonnables.



plyr ou *dplyr* ?

Le paquetage *dplyr* est en fait une seconde version du paquetage *plyr* visant à optimiser le temps de calcul, simplifier son utilisation à l'aide d'une syntaxe plus intuitive et à rendre ses fonctions plus cohérentes entre elles. De plus, *dplyr* concentre son développement autour de la classe objet *data.frame*. Pour toutes ces raisons, l'utilisation de *dplyr* serait à préconiser si vous travaillez avec des *data.frame* qui consistent du même coup en la classe standard de R pour représenter les bases de données...

<https://blog.rstudio.org/2014/01/17/introducing-dplyr/>

2.1.3 Visualisation et analyse des données

La visualisation des données est une étape cruciale dans l'interprétation de ces dernières. En effet, seule une connaissance approfondie des données vous permettra d'en percevoir les secrets les plus précieux qui y résident. Afin de visualiser les données directement contenues dans une relation, le langage R met à notre disposition différentes fonctions qui sont décrites ci-dessous.

<i>View</i>	Permet d'ouvrir la relation dans l'outil de visualisation de RStudio. Ce dernier permettra aussi d'appliquer des transformations de faible complexité comme le filtre sur un variable ou le tri. [36]
<i>head</i>	Renvoie en console un extrait des premières observations d'une relation (par défaut, 10 observations sont renvoyées). [44]
<i>summary</i>	Compilation de statistiques pertinentes au sujet des différentes variables contenues dans une relation. Pour les variables quantitatives, le minimum, le 1 ^{er} quintile, la moyenne, la médiane, le 3 ^{ème} quintile et le

maximum seront calculés, tandis qu’une simple analyse de fréquence des différentes modalités sera produite dans le cas d’une variable qualitative.

table Au même titre que le comportement de *summary* pour les variables qualitative, la fonction *table* renvoie un vecteur comptabilisant le nombre d’occurrences de chaque valeur unique. [28]

Code Source 2.7 – Fonctions de visualisation de données

```
1 View(airportsCanada)
2 head(airportsCanada)
3 summary(airportsCanada)
4 nbAirportCity <- table(airportsCanada$city)
```

Ces fonctions ressemblent beaucoup plus à des outils pour optimiser le temps de développement qu’à des traitements que nous chercherons à laisser en production compte tenu de leur affichage très rudimentaire. De plus, il sera facile de se perdre dans le contenu présenté plus la relation possèdera des variables. Pour contrer ce problème, la production de graphiques sera la plupart du temps une solution plus qu’intéressante. Cependant, toujours dans un objectif de cohérence avec la structure du code source du projet, nous n’aborderons pas immédiatement la création de graphiques en R. Nous nous contenterons plutôt d’introduire les méthodes de visualisation de données géospatiales pour faire le pont avec la [sous-section 2.1.2](#).

Au moment de l’analyse, deux paquetages ont retenu notre attention pour la production de cartes géographiques qui faciliteront la transmission de connaissances sommaires au sujet du jeu de données. Nos critères de sélection étaient encore une fois la simplicité des requêtes, la beauté de l’extrant final et la flexibilité des instructions pour les adapter à un contexte précis.

Le paquetage *ggmap*, nous a permis de produire la [Figure 2.3](#). Si cette dernière vous semble familière, ce n’est pas sans raison ! Le paquetage *ggmap* vise en fait à rendre la visualisation de données géospatiales sur des supports statiques disponibles en ligne tels que ceux de *Google Maps* en les combinant avec la puissance des fonctionnalités du paquetage *ggplot2*. [22]

En jetant un coup d’œil au [Code Source 2.8](#), nous voyons qu’il est possible des cartes très rapidement avec seulement quelques lignes de code. Malgré la facilité d’utilisation de *ggmap*, nous ressentons rapidement ses limitations lorsque nous espérons produire des cartes interactives similaires à celles que nous retrouvons dans la plupart des applications web et mobiles modernes.

Code Source 2.8 – Générer une carte du trafic aérien avec *ggmap*

```
1 # install.packages("ggmap")
2 library(ggmap)
3 map <- get_map(location = "Canada", zoom = 3)
4 TrafficData <- subset(airportsCanada, as.numeric(paste(combinedIndex)
5 ) > 0.05)
6 lon <- as.numeric(paste(TrafficData$longitude))
7 lat <- as.numeric(paste(TrafficData$latitude))
```

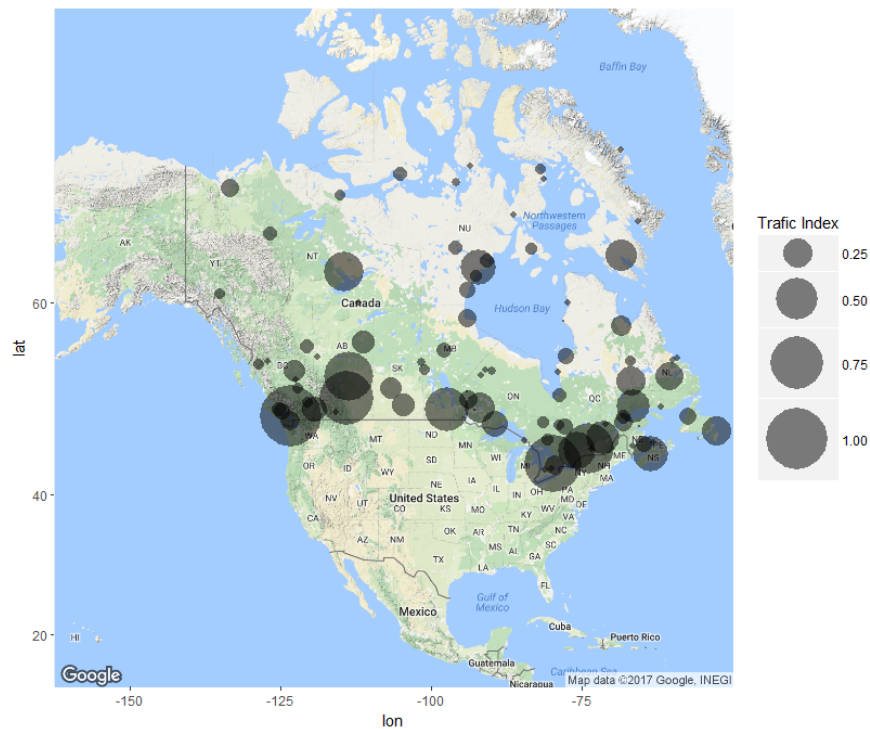


FIGURE 2.3 – Exemple de carte géographique produite avec *ggmap*

```

7 size <- as.numeric(paste(TrafficData$combinedIndex))
8 airportsCoord <- as.data.frame(cbind(lon, lat, size))
9 mapPoints <-
10   ggmap(map) +
11   geom_point(data=TrafficData, aes(x=lon, y=lat, size=size), alpha=0.5,
12             shape=16)
13 (mapTraffic <-
14   mapPoints +
15   scale_size_continuous(range = c(0, 20), name = "Traffic Index"))

```

Pour ce faire, le paquetage *leaflet* [9] viendra à notre secours avec un faible coût en complexité compte tenu de la flexibilité impressionnante rajoutée. Le [Code Source 2.9](#) est à l'origine de la vue statique présentée à la ?? extraite de la carte interactive qu'il génère.

Code Source 2.9 – Générer une carte du trafic aérien avec *leaflet*

```

1 # install.package("leaflet")
2 library(leaflet)
3 url <- "http://hiking.waymarkedtrails.org/en/routebrowser/1225378/
4       gpx"

```

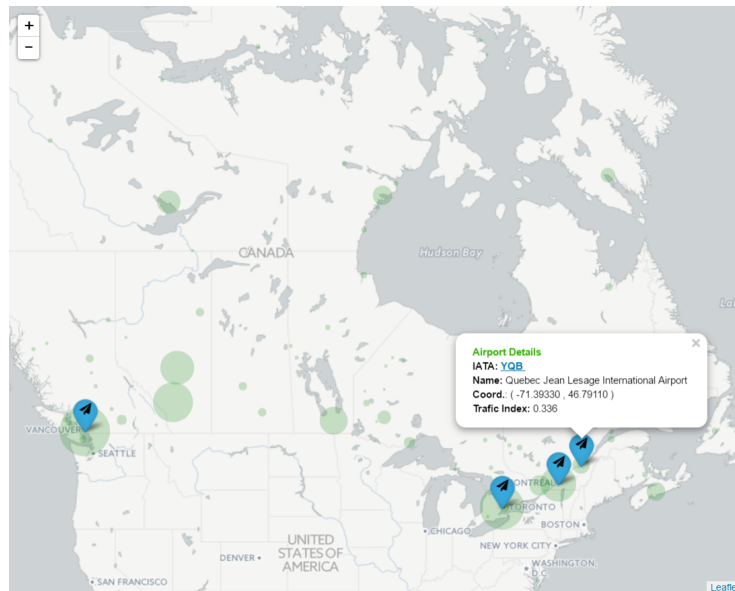


FIGURE 2.4 – Exemple de carte géographique produite avec *leaflet*

```

4 download.file(url, destfile = paste(path, "/Reference/worldRoutes.
  gpx", sep=""), method = "wget")
5 worldRoutes <- readOGR(paste(path, "/Reference/worldRoutes.gpx", sep=
  ""), layer = "tracks")
6 markersData <- subset(airportsCanada, IATA %in% c('YUL', 'YVR', 'YYZ',
  'YQB'))
7 markersWeb <- c("https://www.aeroportoquebec.com/fr/pages/accueil"
  ,
  "http://www.admtl.com/",
  "http://www.yvr.ca/en/passengers",
  "https://www.torontopearson.com/")
8
9
10
11
12 # Defining the description text to be displayed by the markers
13 descriptions <- paste("<b><FONT COLOR=#31B404> Airport Details</FONT
  ></b> <br>",
14                       "<b>IATA: <a href=", markersWeb, ">", markersData$I
  IATA, "</a></b><br>",
15                       "<b>Name:</b>", markersData$name, "<br>",
16                       "<b>Coord.</b>: (", markersData$longitude, ",",
  markersData$latitude, ") <br>",
17                       "<b>Traffic Index:</b>", markersData$
  combinedIndex)
18
19 # Defining the icon to be add on the markers from fontawesome
  library
20 icons <- awesomeIcons(icon = 'paper-plane',
21                       iconColor = 'black',
22                       library = 'fa')
23

```



```

24 # Combinaison of the different components in order to create a
    standalone map
25 (mapTraffic <- leaflet(worldRoutes) %>%
26   addTiles(urlTemplate = "http://{s}.basemaps.cartocdn.com/light_
    all/{z}/{x}/{y}.png") %>%
27   addCircleMarkers(stroke = FALSE, data = TrafficData, ~as.numeric(
    paste(longitude)), ~as.numeric(paste(latitude)),
28     color = 'black', fillColor = 'green',
29     radius = ~as.numeric(paste(combinedIndex))*30,
    opacity = 0.5) %>%
30   addAwesomeMarkers(data = markersData, ~as.numeric(paste(
    longitude)), ~as.numeric(paste(latitude)), popup =
    descriptions, icon=icons))
31
32 # Resizing of the map
33 mapTraffic$width <- 874
34 mapTraffic$height <- 700
35
36 # Export of the map into html format
37 # install.packages("htmlwidgets")
38 library(htmlwidgets)
39 saveWidget(mapTraffic, paste(path, "/Reference/leafletTraffic.html",
    sep = ""), selfcontained = TRUE)

```

Le fonctionnement des deux paquetages est sensiblement le même. Nous commençons par extraire une carte qui servira de support directement à partir du web. Nous passons ensuite les informations géographiques nécessaires au constructeur du paquetage utilisé pour créer une instance. Nous ajoutons ensuite des composantes à cette instance à l'aide de méthode conçue spécifiquement à cette fin. Sans entrer davantage dans les détails, il est intéressant de mentionner les particularités que le paquetage *leaflet* offre en sus des fonctionnalités graphiques traditionnelles. Tout d'abord, les *markers* peuvent être personnalisées de fond en comble. Dans l'exemple présent, nous avons mis à profit la banque de symboles et d'outils CSS (*Cascading Style Sheets*) *fontawesome* [4] qui est célèbres auprès des utilisateurs L^AT_EX pour la diversité et la qualité de ses icônes. Un autre aspect encore plus pratique est la présentation d'informations supplémentaires lorsque l'utilisateur appuie sur le marqueur offrant ainsi une manière simple de stockée beaucoup d'information au sein du même objet sans alourdir indument sa lisibilité. L'ajout de ces informations et le formatage se résument par le passage de commande *html* directement à l'argument *popup*. Vous savez maintenant comment nous avons procédé pour exposer le code IATA, le nom, les coordonnées géographiques ainsi que l'indice de trafic aérien sur chacun des marqueurs auxquels l'icône *fa-paper-plane* a été assignée. Le dernier point intéressant de *leaflet* est la capacité de créer des *widgets html* indépendants rendant le partage de l'information encore plus simple sans nécessiter de recompiler le code source à chaque fois qu'un utilisateur aura envie de visionner l'objet. [9]

i**Est-ce tout ce que peut accomplir *leaflet* ?**

Bien entendu, les exemples présentés dans ce document font l'éloge que de deux applications grossières des capacités de ces deux paquetages. Vous serez en mesure d'aisément trouver plusieurs autres exemples d'applications sur le web. Pour l'instant, voici quelques pages d'intérêt qui ont servi à créer la carte interactive pour amorcer vos recherches :

<https://rstudio.github.io/leaflet/>
<https://rstudio.github.io/leaflet/markers.html>
<https://rstudio.github.io/leaflet/popups.html>
<http://rgeomatic.hypotheses.org/550>
<https://www.r-bloggers.com/interactive-mapping-with-leaflet-in-r/>
<http://stackoverflow.com/questions/38837112/how-to-specify-radius-units-in-addcirclemarkers-when-using-leaflet-in-r>
<http://stackoverflow.com/questions/31562383/using-leaflet-library-to-output-multiple-popup-values>
<https://gis.stackexchange.com/questions/171827/generate-html-file-with-r-using-leaflet>

En terminant, il est possible de valider nos résultats en comparant ceux-ci avec la densité de la population canadienne. Nous devrions être en mesure d'observer une augmentation du trafic aérien dans les zones où la densité de population est plus intense. Ainsi, nos cartographies sont cohérentes avec la [Figure 2.5](#).

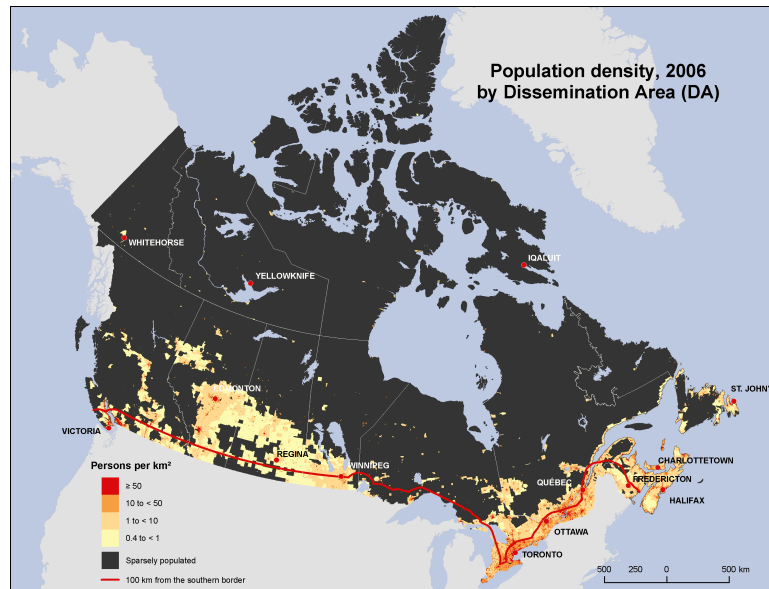


FIGURE 2.5 – Densité de la population canadienne

2.2 Création de fonctions utilitaires

Cette section servira principalement à faire la revue des concepts les plus importants dans la création de fonctions utilitaires. Lorsque nous parlons de fonctions utilitaires, nous faisons référence à des fonctions définies par l'utilisateur afin de favoriser la compréhension de son programme et favoriser la réutilisation de tronçons de programme. Dans le cadre du projet, nous avons pris l'initiative de construire les trois fonctions suivantes :

- airportsDist** Calculer la distance en Km entre deux aéroports
- arrivalTime** Calculer l'heure d'arrivée d'un colis posté au moment du calcul
- shippingCost** Calculer le coût d'une livraison

Lorsque nous voulons définir une fonction, la structure présentée par le [Code Source 2.10](#) sera toujours utilisée.

Code Source 2.10 – Structure pour la définition d'une fonction

```

1 # nom_de_la_fonction <- function( liste_des_arguments )
2 # {
3 #   corps_de_la_fonction
4 #   ...
5 #   valeur_retournee_par_la_fonction
6 # }
```

À partir du [Code Source 2.10](#), nous pouvons dès lors déduire plusieurs éléments de théorie. Tout d’abord, le mot clé *function* sera toujours nécessaire pour mentionner à R que nous sommes en train de définir une fonction, et ce, qu’elle soit anonyme ou non. D’autre part, la valeur retournée par une fonction sera toujours la valeur de la dernière expression évaluée au sein de son corps qui sera délimitée par les accolades. Bien entendu, il est possible de contourner ce processus standard en introduisant le mot clé *return* qui aura pour effet d’entreprendre les processus de retour à l’exécution du programme principal tout en ignorant le reste de l’exécution que la fonction aurait pu engendrer. C’est exactement ce que le [Code Source 2.11](#) cherche à expliciter. Bien que la seule différence entre les deux fonctions soit la présence de l’instruction *return*, ces deux fonctions auront un comportement bien différent puisque la première retournera l’addition des deux paramètres qu’elle aura reçus pendant que la seconde arrêtera son exécution au croisement de l’instruction *return* pour renvoyer la valeur du premier argument, soit 5 et 2 respectivement. En théorie, nous chercherons donc à éviter l’utilisation du *return* ou d’autres modificateurs de flux du même genre et il n’y aura donc qu’une entrée et une sortie possible pour chaque fonction. En pratique, ce genre d’instructions peuvent simplifier grandement l’écriture du code, mais leur utilisation restera réservée à des situations bien particulières.

Code Source 2.11 – L’instruction *return* et le retour standard d’une fonction R

```
1 ftest1 <- function(a,b)
2 {
3   a+b
4 }
5 ftest2 <- function(a,b)
6 {
7   return(a)
8   a+b
9 }
10 ftest1(2,3)
11 ftest2(2,3)
```

L’exemple du [Code Source 2.11](#) combiné à la structure générique présentée précédemment nous accorde un environnement idéal pour introduire les notions d’argumentation. Comme mentionné ci-dessus, le passage des arguments se fera à l’intérieur des parenthèses suivant le nom des fonctions. Il s’agit en fait de la même syntaxe que toutes les autres fonctions que nous avons déjà utilisées dans la section précédente. En d’autres mots, une fois une fonction utilitaire définie correctement par l’utilisateur, celle-ci sera équivalente aux autres fonctions rendues disponibles par les différents paquets. Si nous examinons le [Code Source 2.11](#), nous voyons que la fonction *ftest1* et *ftest2* prennent 2 paramètres à titre d’arguments nommés *a* et *b*. Une fois les arguments déclarés dans l’en-tête de fonction, nous pourrions les utiliser comme bon nous semble à l’intérieur du corps en utilisant leur étiquette.

Code Source 2.12 – Définir des valeurs par défauts dans les fonctions utilitaires

```
1 ftest3 <- function(a=2,b=3)
2 {
3   a+b
4 }
5 ftest3()
```

Comme plusieurs autres langages de programmation, la méthode entreprise pour définir des paramètres par défaut revient simplement à en faire la définition directement dans l'en-tête de la fonction grâce à l'opérateur d'égalité. Bien que la définition de paramètre par défaut peut sembler anodine pour un nouveau programmeur, vous apprendrez rapidement au cours de votre carrière que vos programmes ne doivent jamais contenir de chiffres magiques. Nous désignons par chiffre magique, tout nombre (et par extension toute expression) constant présent dans un programme sur lesquels un utilisateur donné ne pourrait avoir une influence sur celui-ci sans directement modifier le code source. Malgré le fait que vous soyez convaincus que votre programme ne vous sera jamais utile dans un autre dessein que celui pour lequel qui vous a initialement amené à le créer, vous serez souvent influencé par le contexte dans lequel vous opérez. En plus d'être inefficace, ce genre de pratique va directement à l'encontre du but premier de la définition de fonction au sens élargi, soit la réutilisation du code. Un moyen simple d'ajouter de la flexibilité à une fonction sera alors la définition de paramètres par défaut. Vous ne pourrez que retirer du positif d'adopté des bonnes pratiques de programmation dès vos débuts dans le domaine. Sur le long terme et à l'aide d'une documentation adéquate de vos programmes (et fonctions), vous ne pourrez que bénéficier de votre rigueur même si cette dernière vous aura fait perdre du temps précieux au début de votre apprentissage.

D'accord, mais qu'entendons-nous par documentation "adéquate" ? Trop souvent, la mauvaise documentation d'un programme ne vient pas d'un mal intentionnellement causé par le développeur, mais bien d'une mauvaise éducation sur ce qui caractérise une bonne documentation. Premièrement, le fait qui vous semble le plus évident au moment du procédé de documentation ne le sera pas nécessairement au futur utilisateur. Par le fait même, une documentation devrait être aussi monotone à lire qu'à écrire. Deuxièmement, une documentation ne devrait pas correspondre à un paragraphe sans structure précise ou encore à un enchaînement de faits complètement désorganisés qui n'auront un sens logique que pour celui qui les aura écrits. Troisièmement, un utilisateur s'attendra à retrouver le même type d'information dans la documentation de deux entités différentes qui sont toutefois du même genre.

Lorsque nous mettons ces considérations en perspective, on vient donc rapidement à la conclusion qu'une structure standard devrait toujours être utilisée. En plus d'offrir un cadre rigide sur la manière de créer notre documentation, ces outils auront l'avantage de produire des fichiers de référence complets qui posséderont tous les aspects pratiques d'une documentation professionnelle. Un bon exemple de ce genre d'outils est *Doxygen* [3] qui est très populaire pour la documentation de script écrit en C/C++. Le principe derrière cet outil a justement été repris pour l'adapter au code R dans le cadre du développement du paquetage *roxygen2* [14]. Nous croyons que l'utilisation de ces balises est indispensable même si aucune documentation officielle ne sera jamais générée. Il s'agit simplement d'une excellente habitude de travail et cela vous aidera à structurer votre documentation selon un modèle standard et reconnu par la communauté.



Doxygen et Roxygen, ça respire quoi en hiver ?

Le principe de ces outils est extrêmement rudimentaire. De manière intuitive, nous utiliserons les commentaires afin de faire la documentation de nos programmes. Ce sera toujours le cas ! La principale différence provient de l'introduction de balises qui guideront la présentation de l'information lors de la production de la documentation officielle disponible sous plusieurs formats (html, PDF, \LaTeX , etc.) À titre d'exemple, nous utiliserons la balise `param` pour décrire un paramètre, `return` pour décrire le retour et `examples` pour donner des exemples d'utilisation dans le cadre de la documentation d'une fonction. Dans bien des cas, \LaTeX sera derrière le formatage de cette documentation. Il est bon de savoir que l'écriture d'une telle documentation sera un prérequis à tous ceux qui seraient tentés de créer un paquetage et de le publier sur *Comprehensive \TeX Archive Network (CTAN)*.
<https://cran.r-project.org/doc/manuals/R-exts.html#Marking-text>

En reprenant les fonctions *ftest1*, *ftest2* et *ftest3*, nous pouvons faire quelques tests en variant le nombre d'arguments envoyés et le comportement résultant.

Code Source 2.13 – Passage d'arguments à une fonction

```
1 > ftest1(3)
2 Error in ftest1(3) : argument "b" is missing, with no default
3 > ftest2(3)
4 [1] 3
5 > ftest2(b=5)
6 Error in ftest2(b = 5) : argument "a" is missing, with no default
7 > ftest3(3)
8 [1] 6
9 > ftest3(3,5)
10 [1] 8
11 > ftest3(b=5)
12 [1] 7
13 > ftest3(b=5,3)
14 [1] 8
15 > ftest3(3,5,4)
16 Error in ftest3(3, 5, 4) : unused argument (4)
```

Comme le montre le [Code Source 2.13](#), nous pourrions admettre comme règle que tout argument ne possédant pas de valeur par défaut doit absolument avoir une valeur d'attribuer lors de l'appel de la fonction. De plus, nous observons que la notion d'argument nommé n'a pas vraiment de signification en R. Ainsi, tous les arguments seront traités de manière positionnelle à moins d'indication contraire par la spécification du nom de l'argument dans l'appel de la fonction. Nous pouvons toutefois remarquer un cas particulier avec l'appel de `ftest2(3)` qui fournira bel et bien la valeur de 3 même si aucune valeur n'a été fournie pour le paramètre `b` et qu'il n'aille aucune valeur par

défaut. Ceci s'explique par le fait que R détectera une erreur de valeur manquante qu'au moment de l'exécution plutôt qu'au moment de l'appel de la fonction. Ainsi, puisque cette `ftest2` retournera la valeur de `a` et que son exécution n'ira jamais évaluer la commande `a+b`, R n'aura jamais remarqué l'absence d'une valeur pour `b`. De manière similaire, une erreur sera produite si nous fournissons à `fctest2` qu'une valeur à `b`. L'appel `fctest3(b=5,3)` expose la flexibilité tout aussi incroyable que dangereuse des procédés d'assignation de valeurs lors de l'appel de fonction en R. Cette flexibilité de pouvoir alterné l'ordre pour spécifier les valeurs à nos paramètres vient du fait que R traitera ces deux processus d'assignation de manière indépendante. Dans un premier temps, l'ensemble des valeurs assignées à des paramètres en spécifiant leur nom sera extrait du vecteur de paramètres fourni par l'appel et les valeurs restantes seront attribuées de manière positionnelle sur les arguments n'ayant pas reçu de valeur. Il faut toutefois faire attention à ceci puisque aucune discrimination ne sera effectuée par rapport aux paramètres ayant des valeurs par défaut comme illustré par le [Code Source 2.14](#).

Code Source 2.14 – L'assignation et les valeurs par défaut

```
1 > ftest4 <- function(a,b=3,c,d)
2 + {
3 +   a+b+c+d
4 + }
5 > ftest4(c=2,1,3)
6 Error in ftest4(c = 2, 1, 3) : argument "d" is missing, with no
   default
```

Votre oeil déjà très aguerri a probablement remarqué que les fonctions définies dans le cadre de cette étude de cas utilisaient une technique de retour multiple par l'entremise d'une liste. Cette technique deviendra intéressante dans les cas où une fonction doit effectuer plusieurs sous-calculs correspondant à des entités distinctes d'un calcul donné. À titre d'exemple, bien qu'une fonction soit destinée à exécuter une tâche précise, son utilisateur pourrait parfois être intéressé par la valeur d'un des calculs intermédiaires réalisés. L'avantage de la liste est la possibilité intrinsèque d'attribuer des noms aux différentes valeurs renvoyées. En plus d'ajouter beaucoup de valeur à vos fonctions sans nécessairement rendre le code source beaucoup plus complexe, ce type de retour vous aidera grandement dans le débogage de ces dernières lors de leur développement. Cette technique possède toutefois les désavantages d'imposer une certaine rigueur au niveau de leur utilisation en obligeant l'utilisateur à récupérer la liste dans un objet et ensuite de faire l'extraction de la valeur désirée avec l'opérateur `$`. Le [Code Source 2.15](#) offre un exemple concret de cette notion de retour multiple.

Code Source 2.15 – Retour multiple par l'entremise d'une liste

```
1 > ftest5 <- function(a,b=3,c,d)
2 + {
3 +   returningList <- list()
4 +   returningList$value <- a+b+c+d
5 +   returningList$params <- c(a,b,c,d)
6 +   returningList
7 + }
8 > (x <- ftest5(c=2,1,3,4))
9 $value
```

```

10 [1] 10
11
12 $params
13 [1] 1 3 2 4
14
15 > x$value
16 [1] 10

```

Le dernier thème qu'il nous reste à aborder au sujet des fonctions consiste en la gestion des erreurs. Lorsque nous voulons définir les limites d'utilisation d'une fonction, il est préférable de parfaitement connaître ce qu'elle ne peut accomplir. Nous définirons ensuite des validations pour s'assurer que nous ne sommes pas de ce genre de cas particuliers et nous renverrons à l'utilisateur un message lui permettant de corriger son appel. La simplicité de R pour générer ce genre de traitement enlève toute raison possible de ne pas le faire. Ce procédé se résume en quatre étapes qui sont :

1. Identifier une limitation du programme ;
2. Faire la validation nécessaire pour détecter la survenance de cette limitation ;
3. Composer un message concis fournissant toute l'information nécessaire pour corriger l'appel ;
4. Soulever l'erreur à l'exécution à l'aide de l'instruction *stop* en fournissant le message composer à l'étape précédente en argument.

Code Source 2.16 – Gestion des erreurs sous R

```

1 > ftest6 <- function(a,b)
2 + {
3 +   if(b == 0)
4 +   {
5 +     stop("The value of b is not valid. A division by 0 would be
      generated.")
6 +   }
7 +   a/b
8 + }
9 > ftest6(3,4)
10 [1] 0.75
11 > ftest6(3,0)
12 Error in ftest6(3, 0) :
13   The value of b is not valid. A division by 0 would be generated.

```



Comment jouer avec le feu sans se brûler ?

Il arrivera parfois où la génération d'erreurs sera inévitable, mais pour lesquelles nous voudrions appliquer un traitement particulier. Nous appelons ce processus la gestion d'exception. Similairement à la majorité des autres langages de programmation, R inclut des méthodes *try/catch* pour pallier au problème. Nous avons mis cette technique en pratique dans la dernière partie de cette étude de cas.

<https://stat.ethz.ch/R-manual/R-devel/library/base/html/try.html>

<https://stat.ethz.ch/R-manual/R-devel/library/base/html/try.html>

2.3 Conception de graphiques en R

Avant même d’aborder les fonctionnalités graphiques de R, nous devons préciser qu’elles sont quasi infinies. C’est donc pour cette raison que nous nous contenterons de ne faire qu’une revue globale des types de graphiques qui combleront amplement vos besoins pour faire vos premiers pas. Advenant le cas où ces connaissances ne seront plus suffisantes, il existe énormément d’exemples sur les forums de la communauté pour apaiser votre curiosité.

Pour débiter, la fonction *plot* [33] est de loin la fonction la plus rudimentaire de faire un graphique avec R. Cette fonction ne possède que trois arguments : *x*, *y* et *...*. Naturellement, nous devons fournir des valeurs d’abscisse et d’ordonnée à la commande *plot* via les arguments *x* et *y*. Par la suite, la fonction s’occupera de produire un graphique à points traditionnel. En partant directement du jeu de données *airports.dat*, nous pouvons être tentés d’essayer cette commande en représentant les couples longitude/latitude de chaque aéroport dans le monde. Bien entendu, le résultat obtenu sera peu élégant ne représentant que l’essentiel.

C’est à ce moment que l’argument *...* entre en scène. Nous n’avons pas discuté de ce type d’argument dans la section précédente puisque nous considérions plus intuitif de le présenter à l’aide d’un exemple de son utilisation la plus commune, le passage d’options graphiques au sein de la commande *plot*. Il ne sera toutefois pas rare de retrouver cet argument dans bon nombre de fonctions, mais sa nécessité sera souvent moindre que dans le cas de la création de graphiques. Cet argument possède la propriété particulière d’absorber tous les paramètres qui seront passés à la fonction et qui n’auront pas été assignés à un argument. Ces mêmes paramètres pourront donc ensuite être transmis à une autre fonction au sein du corps de la fonction.

C’est exactement ce qui se produit dans le cas de la commande *plot* qui enverra tous les paramètres supplémentaires à la fonction *par* [45] étant la commande gérant tous les aspects des graphiques en R. Heureusement, il existera des comportements par défaut pour tous les arguments de cette fonction. Il sera inconcevable et surtout inutile à quiconque d’apprendre l’ordre réel dans lequel ses arguments se présentent. Le passage des paramètres se fera donc en nommant chaque argument sur lequel nous voulons imposer un comportement différent.



par magie !

La fonction *par* vous sera de grands secours à plusieurs reprises. Une utilisation fréquente de cette fonction est de modifier la division de la fenêtre d’affichage de R. En modifiant la valeur de l’argument *mfrow*, nous pourrions ainsi combiner plusieurs graphiques intimement reliés sur la même fenêtre graphique facilitant du même coup leur comparaison.

Par exemple, *par(mfrow = c(2,2))* divisera la fenêtre en 2 lignes et 2 colonnes pour ainsi accueillir 4 graphiques distincts.

C’est précisément ce que nous avons fait dans la deuxième version de notre gra-

phique ([Figure 2.3](#)) en spécifiant le nom des axes (*xlab* et *ylab*) ainsi qu'un titre au graphique (*main*). Nous avons aussi modifié le type de point pour passer de points vides à des points remplis (*pch*) tout en réduisant la taille de ces derniers pour obtenir une meilleure résolution (*cex*). Finalement, nous avons utilisé une police en gras pour le titre du graphique et les axes (*font* et *font.lab*) en plus de venir augmenter la taille de ces derniers (*cex.main* et *cex.lab*). Référez-vous au [Code Source 2.17](#) pour plus de détails.

Code Source 2.17 – Utilisation de la commande *plot*

```
1 plot(airports$longitude, airports$latitude)
2 plot(airports$longitude, airports$latitude, cex = 0.1, xlab="Longitude
   ", ylab="Latitude", main="Spatial Coordinates of All the Airports
   ", pch = 20, font = 2, cex.main = 1.5, font.lab = 2, cex.lab = 1.5)
```

Dans le cas où nous aurions plutôt voulu faire la représentation d'une fonction continue, nous pourrions encore une fois utiliser la commande *plot* en modifiant l'argument *type*. Bien que cette pratique peut nous sembler justifiée, elle pourra jouer de mauvais tours à un utilisateur non averti. Comme le montre la [Figure 2.7](#), dépendamment de l'espacement des valeurs des points calculés, nous pourrions perdre toute l'information sur l'allure réelle de la courbe que nous cherchons à visualiser.

Il sera donc préférable d'utiliser la commande *curve* [30] pour ce genre de tâche afin de simplifier le code source en ne précisant que les extrêmes de l'étendue sur lequel nous voulons tracer la fonction en spécifiant au besoin le nombre de valeurs à calculer dans l'intervalle.

Code Source 2.18 – Utilisation de la commande *curve*

```
1 fquad <- function(x, a=2, b=3, c=4)
2 {
3   a*x**2+b*x+c
4 }
5 fquad(2)
6 par(mfrow = c(2,2))
7 plot(x <- seq(-10,10,10), fquad(x,2,3,4), type = "l", ylab = "fquad(x)"
   ", xlab = "x", main = "dx = 10")
8 plot(x <- seq(-10,10,5), fquad(x,2,3,4), type = "l", ylab = "fquad(x)"
   ", xlab = "x", main = "dx = 5")
9 plot(x <- seq(-10,10,2), fquad(x,2,3,4), type = "l", ylab = "fquad(x)"
   ", xlab = "x", main = "dx = 2")
10 plot(x <- seq(-10,10), fquad(x,2,3,4), type = "l", ylab = "fquad(x)"
   ", xlab = "x", main = "dx = 1")
11
12 par(mfrow = c(1,1))
13 curve(fquad(x), from = -10, to = 10)
```

Un autre type de graphique fréquemment utilisé dans les analyses statistiques sont les histogrammes. Ces derniers permettent de rapidement avoir une idée globale sur le type de distribution à laquelle nous sommes confrontés. L'argument *breaks* de la

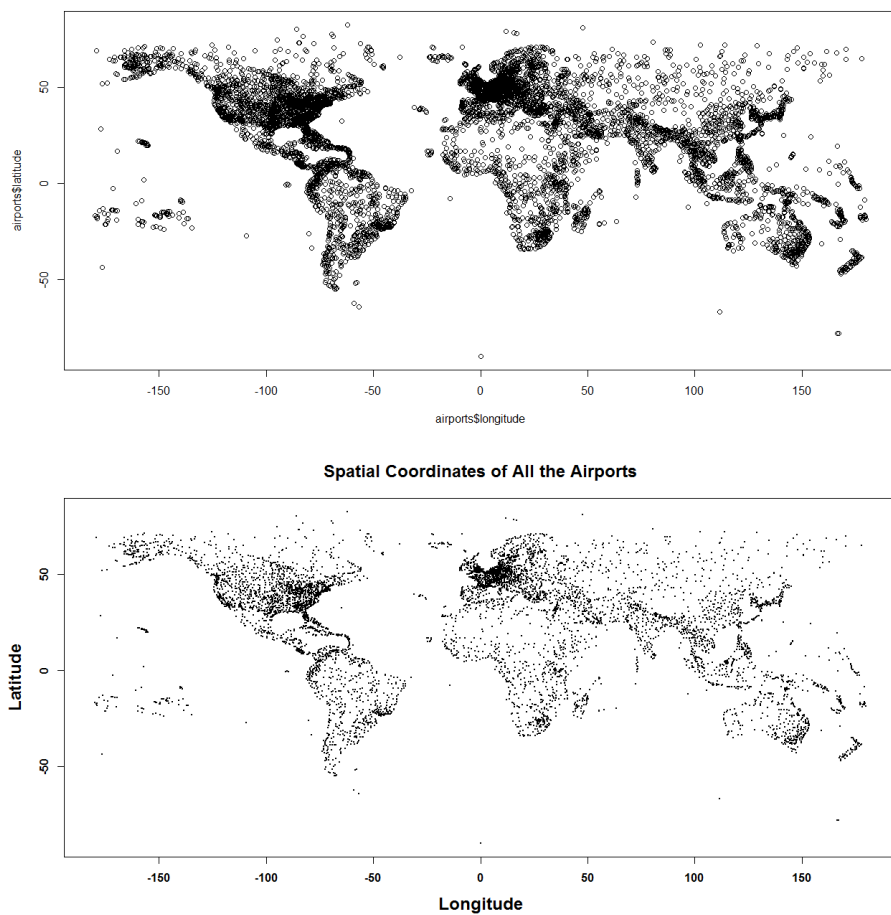


FIGURE 2.6 – Passage de paramètres graphiques à la commande *plot*

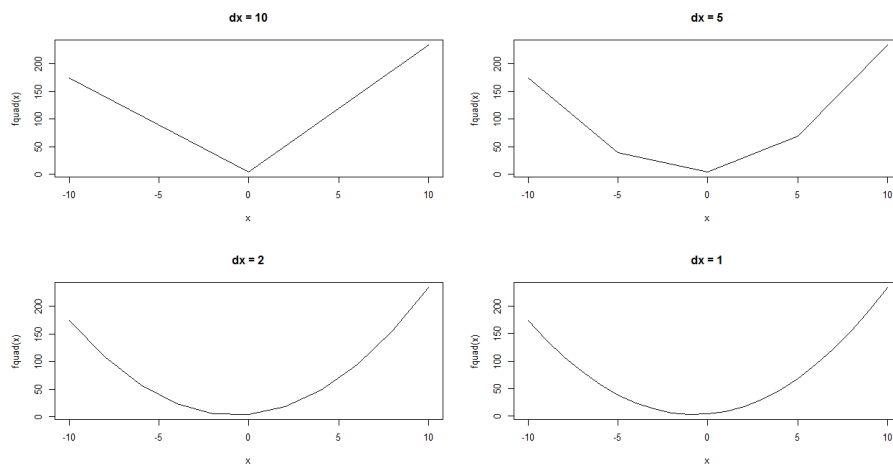


FIGURE 2.7 – Tracer une courbe avec la commande *plot*

commande *hist* [35] est de loin le plus important puisqu'il permettra d'obtenir un visuel beaucoup plus précis de la situation en réduisant la taille des regroupements effectués. En ne spécifiant qu'un seul nombre à cet argument, nous indiquons à R de diviser les données pour obtenir ce même nombre de groupes d'étendue équivalente. Dans le cas où un vecteur de nombre lui serait fourni, R comprendra plutôt qu'il doit regrouper les données en utilisant ces nombres à titre de bornes pour les différents intervalles. Un autre argument bien intéressant est *freq*. Cet argument booléen contrôlera l'affichage de la hauteur des colonnes de l'histogramme. Le nombre d'observations sera affiché si sa valeur est vraie (valeur par défaut) ou sous la forme d'une probabilité.

i

Excel et les histogrammes

Si vous êtes habitués de travailler avec *Excel*, vous avez probablement une mauvaise impression de la valeur ajoutée d'utiliser des histogrammes. Ceci vient du fait qu' *Excel* travaille plutôt avec des graphiques à bâtons. La différence entre ces deux types de graphique réside dans le fait que les colonnes d'un histogramme posséderont à la fois une largeur et une hauteur, tandis que les diagrammes à bâtons ne possèdent qu'une notion de hauteur et sont plutôt destinés à représenter la distribution d'une variable qualitative.

La fonction *density* [37] est aussi très intéressante d'un côté pratique pour estimer la fonction de densité empirique sous-jacente. Cette fonction possède un argument *adj* avec lequel nous contrôlerons le degré de lissage employé. La valeur par défaut de cet argument est 1 et plus sa valeur sera faible, plus nous nous rapprochons de la distribution discrète, tandis qu'une valeur supérieure aura pour effet de lisser davantage

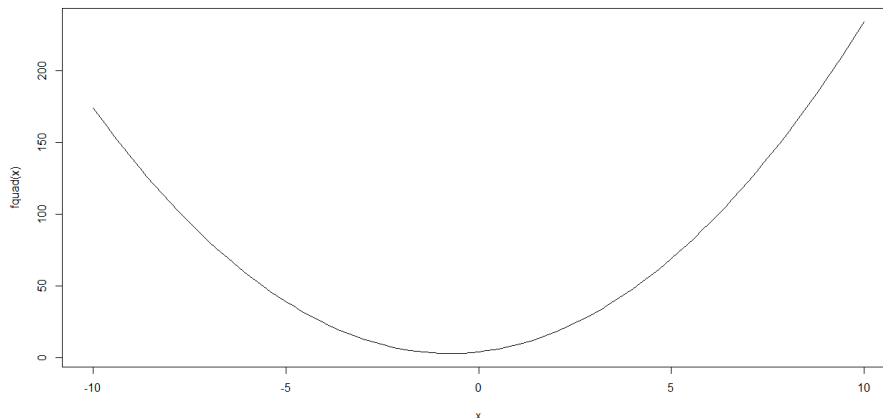


FIGURE 2.8 – Tracer une courbe avec la commande *curve*

la fonction obtenue.

Bon nombre des fonctionnalités graphiques de R peuvent être combinées au sein d'un même graphique. Il s'agira d'un comportement natif dans certains cas (les commandes *points* et *lines*) ou d'un comportement induit par l'argument *add* comme c'est possible de le faire avec *curve*. Il sera possible de facilement tracer la fonction de densité renvoyée par *density* grâce à la commande *lines*.

La commande *abline* [27] simplifiera grandement l'affichage de fonctions linéaires. L'utilisation de celle-ci pourra se faire de trois manières différentes. La première consiste à spécifier les arguments *a* et *b* pour produire la représentation d'une droite d'équation $y = ax + b$. La deuxième permettra plutôt de tracer une droite d'équation $y = h$ en attribuant une valeur à l'argument *h*. La dernière et non la moindre qui est, selon moi, la plus commode d'entre toutes permet de créer des droites d'équation $x = v$. L'ajout de ce genre de droites permettra de faire ressortir des valeurs d'abscisses ayant une signification particulière dans le cadre de votre analyse.

Certaines autres fonctions vous permettront de rajouter de l'information afin de faciliter la lecture de vos graphiques. Parmi ces fonctions, la plus importante sera *legend* qui comme son nom l'indique, s'occupera de générer une légende au graphique que nous venons de produire. Cette fonction est tout autant paramétrisable que le graphique sous-jacent. Nous pouvons tout de même identifier des arguments plus communs que d'autres. L'argument *bty* permettra de supprimer l'encadrement de la légende en lui attribuant la valeur "n". Nous préciserons aussi un type de points avec *pch* ou un type de ligne avec *lty* sur lesquels nous pourrons affecter la même couleur que la courbe correspondante à l'aide de *col*. La fonction *mtext* s'occupera plutôt d'ajouter du texte à des endroits précis sur le graphique pour noter des observations ou ajouter des explications sur des aspects qui nous semblent plus surprenants.

L'ensemble des points discutés ci-dessus ont été repris dans le [Code Source 2.19](#) pour produire la [Figure 2.9](#).

Code Source 2.19 – *hist*, *density*, *lines*, *abline*, *legend* et *mtext*

```

1 Altitude <- as.numeric(paste(airportsCanada$altitude))
2 hist(Altitude)
3 hist(Altitude,xlim = c(0,5000))
4 hist(Altitude,xlim = c(0,5000),breaks = 100)
5 hist(Altitude,xlim = c(0,5000),breaks = 100,freq = FALSE,col = "
  gray",border = grey(0.8),font = 2,font.lab = 2)
6 lines(density(Altitude,adjust = 4),lwd = 2,col = "blue")
7 lines(density(Altitude,adjust = 1),lwd = 2,col = "purple")
8 lines(density(Altitude,adjust = 0.25),lwd = 2,col = "red")
9 altitudeAvg <- round(mean(Altitude),1)
10 abline(v = altitudeAvg,lwd = 2)
11 legend(2500,0.0015,legend = c("4","1","0.25"),title = "Density
  Adjustment \n Factor",col = c("blue","purple","red"),bty = "n",
  title.col = "black",lty = 1,lwd = 3,y.intersp = 0.5,cex = 1.25)
12 mtext(paste("Average: \n",altitudeAvg),at = altitudeAvg,cex = 0.75)

```

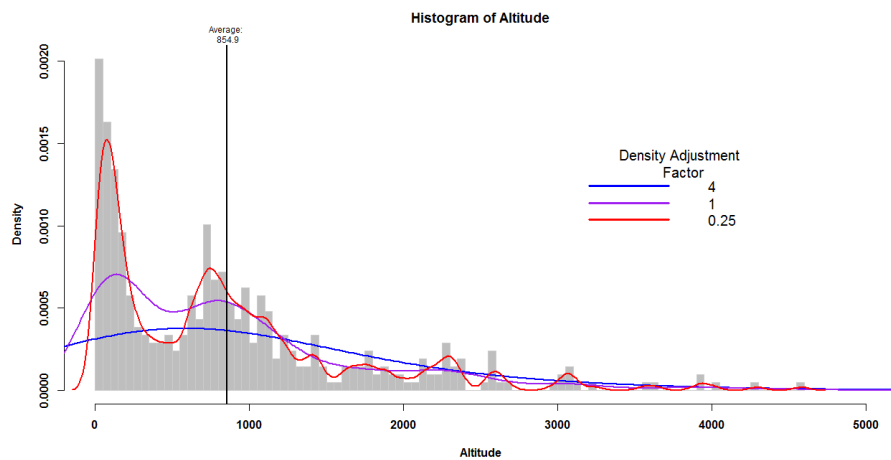


FIGURE 2.9 – Distribution des altitudes des aéroports canadiens

i**Vers l'infini et plus loin encore !**

Vous aurez compris qu'il ne s'agit que d'un TRÈS bref aperçu des capacités graphiques de R. Il existe des structures standard pour générer d'autres types de graphique tels que les diagrammes en pointes de tarte (*pie*) ou encore les boîtes à moustaches (*boxplot*). Certains d'entre vous trouveront peut-être que la génération de graphique est un processus lent et ardu, mais il s'agit ici du coût à payer pour avoir autant de flexibilité. Ces mêmes personnes seront toutefois heureuses d'apprendre que plusieurs paquets intègrent des modules de visualisation standard pour les objets qui leur sont propres. Il serait par contre un peu prétentieux de définir et de modifier les options d'affichage par défaut des objets dont l'existence ne dépend aucunement de leur utilisation.

2.4 Outils d'analyse statistique en R

Un des aspects du langage R sur lequel sa réputation s'est bâtie est la variété des outils statistiques qu'il place à la disposition de son utilisateur. Sans même avoir à importer une quelconque librairie à partir de CRAN, plusieurs distributions statistiques sont disponibles. La [Tableau 2.4](#) fait la revue des ces distributions et de leur identifiant R correspondant. [10]

Distribution	identifiant R
Bêta	beta
Binomiale	binom
Binomiale négative	nbinom
Chi Deux	chisq
Exponentielle	exp
Fisher	f
Gamma	gamma
Géométrique	geom
Hypergéométrique	hyper
Normale	norm
Poisson	pois
Student	t
Uniforme	unif
Weibull	weibull

TABLE 2.1 – Liste des distributions statistiques disponibles en R

D'autres distributions deviendront aussi disponibles via des paquets dédiés à cette fin. Le paquetage **actuar** [20] donne accès à plusieurs distributions supplémentaires communément utilisées en actuariat. La distribution Pareto en est un bon

exemple.

Un aspect particulièrement intéressant de ces implémentations de distribution statistique (qu'elles soient disponibles par défaut en R ou via l'importation d'un paquetage) est la constance dans la structure de ces fonctions. Pour chacune des distributions, nous retrouverons en outre les trois fonctions qui suivent :

$d\langle ID_R \rangle$	Calcule la valeur de la fonction de densité de la distribution ayant l'identifiant R $\langle ID_R \rangle$.
$p\langle ID_R \rangle$	Calcule la valeur de la fonction de répartition de la distribution ayant l'identifiant R $\langle ID_R \rangle$.
$q\langle ID_R \rangle$	Renvoie le quantile associé à la valeur fournie en argument selon la fonction de répartition de la distribution ayant l'identifiant R $\langle ID_R \rangle$.
$r\langle ID_R \rangle$	Permet de générer des valeurs aléatoires suivants la distribution ayant l'identifiant R $\langle ID_R \rangle$.

De plus, les arguments de ces fonctions se présenteront toujours sous le même format. Nous devrons soit fournir la valeur à laquelle nous voulons évaluer la fonction ou encore un nombre d'observations à générer dans le cas des fonctions préfixées par "r" et les paramètres de la loi utilisée. À des fins d'optimisation des performances, le logarithme de ces fonctions sera souvent nécessaire et c'est ce qui explique la présence de l'argument `log`.⁵ Finalement, nous serons parfois intéressés par la fonction de survie d'une distribution donnée correspondant au complément de la fonction de répartition. En attribuant la valeur `FALSE` à l'argument `lower.tail`, les fonctions préfixées par "p" renverront ainsi la valeur de la fonction de survie. Un exemple d'utilisation de ces fonctions est présenté par le [Code Source 2.20](#).

Code Source 2.20 – Fonctions relatives à la distribution Normale

```
1 > set.seed(2017)
2 > mean <- 6
3 > sd <- 2
4 > x <- 0:12
5 > dnorm(x, mean, sd)
6 [1] 0.002215924 0.008764150 0.026995483 0.064758798
7 [5] 0.120985362 0.176032663 0.199471140 0.176032663
8 [9] 0.120985362 0.064758798 0.026995483 0.008764150
9 [13] 0.002215924
10 > pnorm(x, mean, sd)
11 [1] 0.001349898 0.006209665 0.022750132 0.066807201
12 [5] 0.158655254 0.308537539 0.500000000 0.691462461
13 [9] 0.841344746 0.933192799 0.977249868 0.993790335
14 [13] 0.998650102
15 > r <- seq(0,1,0.1)
16 > qnorm(r, mean, sd)
17 [1] -Inf 3.436897 4.316758 4.951199 5.493306
18 [6] 6.000000 6.506694 7.048801 7.683242 8.563103
19 [11] Inf
20 > rnorm(10, mean, sd)
```

5. Plusieurs propriétés statistiques découlent du logarithme des fonctions de densité et de répartition tel que la fonction génératrice de moments pour ne nommer que cette dernière.

```

21 [1] 8.868403 5.845416 7.478274 2.482791 5.860350
22 [6] 6.903811 2.083267 5.996951 5.469328 9.126445

```

Ceux qui sont familiers avec les distributions statistiques auront remarqué qu'à l'aide des fonctions décrites ci-dessus nous aurons donc deux manières de générer des nombres aléatoires. La première qui est aussi la plus évidente sera d'utiliser les fonctions préfixées avec "r". La seconde utilisera le théorème de la réciproque consistant à générer des valeurs aléatoires suivant une loi uniforme de paramètre $a := 0$ et $b := 1$ pour ensuite trouver le quantile correspondant de la fonction de répartition de la loi pour laquelle nous voulons générer des nombres aléatoires grâce aux fonctions préfixées par "q". Ces deux techniques sont mises à profit dans le [Code Source 2.21](#).

Code Source 2.21 – Génération de nombres aléatoires

```

1 > y1 <- rnorm(1000, mean, sd)
2 > summary(y1)
3      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
4  0.07041  4.70800   6.02800   6.06200   7.35500  12.59000
5 > sd(y1)
6 [1] 1.96455
7 > r <- runif(1000)
8 > y2 <- qnorm(r, mean, sd)
9 > summary(y2)
10      Min. 1st Qu.  Median     Mean 3rd Qu.    Max.
11 -0.1347  4.7670   6.0830   6.0910  7.5070  12.2400
12 > sd(y2)
13 [1] 1.966951

```



Théorème de la réciproque

Ce sont les 4 propriétés des fonctions de répartition qui rendent possible l'application du théorème de la réciproque. Ces propriétés sont définies comme suit (où F désigne la fonction de répartition d'une variable aléatoire X quelconque) :

1. F_X est croissante
2. Elle est partout continue à droite
3. $\lim_{x \rightarrow -\infty} F_X(x) = 0$
4. $\lim_{x \rightarrow \infty} F_X(x) = 1$

Étant donné que ces propriétés seront toujours respectées pour toute fonction de répartition, nous pourrons appliquer cette méthode, peu importe la distribution qu'elle soit clairement définie ou non !

https://fr.wikipedia.org/wiki/Fonction_de_r%C3%A9partition#Th.C3.A9or.C3.A8me_de_la_r.C3.A9ciproque

En présence de données empiriques, la première étape d'une analyse statistique sera de dresser le portrait statistique de ces données. Nous avons déjà parlé de la

fonction `summary` à la [sous-section 2.1.1](#). Nous rajouterons ici les fonctions `mean` et `sd` retournant respectivement la moyenne et l'écart-type d'un jeu de données empiriques comme nous l'avons fait montré dans le [Code Source 2.21](#).

Afin de valider l'ajustement d'une distribution sur les données empiriques, nous serons souvent contraints à identifier les fonctions de densité et de répartition sous-jacentes. Il existe plusieurs façons de faire. Celle qui nous semble toutefois la plus pertinente et polyvalente exploite le comportement de la fonction `ecdf`. Cette dernière permet de construire une fonction de répartition empirique à partir des observations fournies en argument. Nous pouvons ensuite construire une fonction de densité empirique en évaluant cette fonction de répartition à deux points autour de la valeur désirée et en divisant ensuite le résultat par la largeur de l'intervalle évalué. Les instructions permettant de construire ces fonctions sont fournies par [Code Source 2.22](#).

Code Source 2.22 – Fonctions de densité et de répartition empiriques

```
1 empCDF <- ecdf(compData$weight)
2 empPDF <- function(x, delta=0.01)
3 {
4   (empCDF(x+delta/2)-empCDF(x-delta/2))/delta
5 }
```

En plus de dresser le portrait statistique des données, on peut aussi vouloir faire des tests statistiques à partir de celles-ci. Parmi les tests disponibles, nous retrouvons notamment :

- ▶ Test de normalité (Test de Shapiro-Wilk)
- ▶ Test de comparaison de deux variances (Test F)
- ▶ Test de Student
- ▶ Test du Khi carré
- ▶ Test de Wilcoxon
- ▶ ANOVA (Analyse de variance)
- ▶ Test de corrélation

Il n'est toutefois pas indispensable de connaître l'utilité de tous ces tests, les situations dans lesquelles ils devront être utilisés ni la mécanique mathématique sous-entendue puisque la plupart des méthodes statistiques incluront déjà les appels nécessaires de ceux-ci. Ce sera le cas de la fonction `lm` comme nous le verrons plus loin. [\[16\]](#)

Dans le cadre de notre étude de cas, nous avons performé les tests du Khi carré et de corrélation afin de s'assurer que les variables explicatives du poids et de la distance soient indépendantes et sans corrélation. Dans le cas où ce genre de phénomène serait apparu entre nos variables, nous aurions été contraints d'utiliser des modèles de régression plus complexes tels que les modèles linéaires généralisés.

Lorsque nous effectuons un test statistique, nous cherchons toujours à répondre à une question binaire représentée sous la forme de deux hypothèses H_0 et H_1 complémentaires. Une valeur nommée la **p-value** sera ensuite calculée en acceptant l'hypothèse H_0 comme vraie. Cette valeur correspondra à la probabilité d'observer un

résultat équivalent ou supérieur du test que nous venons d'exécuter en considérant l'hypothèse nulle comme vraie. En d'autres mots, cette valeur nous indiquera la probabilité de se tromper en rejetant l'hypothèse nulle en considérant l'hypothèse nulle comme vraie initialement. Ainsi, à partir du moment où la **p-value** sera inférieure au seuil de crédibilité que l'on s'était fixé (habituellement 5%), nous considérerons l'hypothèse nulle comme fausse.

Dans le cas du test du Khi carré, l'hypothèse nulle suppose que les deux distributions sont indépendantes. Le test de corrélation suppose tant qu'à lui que la valeur théorique de corrélation est équivalente à 0. Comme nous pouvons le voir avec le [Code Source 2.23](#), nous ne pouvons pas rejeter ces deux hypothèses.

Code Source 2.23 – Tests d'indépendance et de corrélation entre distributions

```

1 > weightsBinded <- as.numeric(cut(compData$weight,25))
2 > distancesBinded <- as.numeric(cut(compData$distance,25))
3 > contingencyTable <- table(weightsBinded, distancesBinded)
4 > chisq.test(contingencyTable)
5
6   Pearson's Chi-squared test
7
8 data:  contingencyTable
9 X-squared = 248.38, df = 391, p-value = 1
10
11 Warning message:
12 In chisq.test(contingencyTable) :
13   Chi-squared approximation may be incorrect
14 > contingencyTable <- rbind(contingencyTable[1:4,], colSums(
15   contingencyTable[5:18,]))
16 > contingencyTable <- cbind(contingencyTable[,1:14], rowSums(
17   contingencyTable[,15:24]))
18 > (independencyTest <- chisq.test(contingencyTable))
19
20   Pearson's Chi-squared test
21
22 data:  contingencyTable
23 X-squared = 72.814, df = 56, p-value = 0.06495
24
25 > cor.test(compData$weight, compData$distance, method = "pearson")
26
27   Pearson's product-moment correlation
28
29 data:  compData$weight and compData$distance
30 t = -0.7801, df = 99998, p-value = 0.4353
31 alternative hypothesis: true correlation is not equal to 0
32 95 percent confidence interval:
33 -0.008664731  0.003731121
34 sample estimates:
35      cor
36 -0.0024669

```

Il est pertinent de remarquer que le test du Khi carré possède des limitations importantes dans le cas de distributions devenant un peu trop clairsemées. Ce test nécessite une efficacité statistique d'au minimum 5 observations à toutes les intersections

des deux variables catégoriques. C'est pour cette même raison que nous combinons les dernières lignes et colonnes de la table de contingence. Malgré tout, le test offre toujours une **p-value** d'environ 6% ce qui reste supérieur à notre seuil de 5% et nous ne pouvons donc pas rejeter notre hypothèse nulle. Dans le cas du test de corrélation, nous voyons que la valeur 0 est comprise dans notre intervalle de confiance autour de la valeur de corrélation empirique déterminée, ce qui nous permet d'affirmer qu'aucune corrélation n'existe entre ces deux variables. La **p-value** de 43% aurait été suffisante pour arriver à la même conclusion.

Pour terminer cette section, jetons un coup d'oeil à la régression linéaire qui fut accomplie dans le but de modéliser la distribution ayant mené à générer les données (Code Source 2.24).

Code Source 2.24 – Régression linéaire sur données empiriques

```

1 > profitMargin <- 1.12
2 > avgTaxRate <- sum(table(airportsCanada$province)*as.numeric(paste
  (taxRates$taxRate)))/length(airportsCanada$province)
3 > compModel <- lm(price/(profitMargin*avgTaxRate) ~ distance +
  weight, compData)
4 > summary(compModel)
5
6 Call:
7 lm(formula = price/(profitMargin * avgTaxRate) ~ distance + weight,
8     data = compData)
9
10 Residuals:
11      Min       1Q   Median       3Q      Max
12 -30.7903  -4.6585   0.0305   4.6462  29.9563
13
14 Coefficients:
15             Estimate Std. Error t value Pr(>|t|)
16 (Intercept)  3.227e+01  7.509e-02  429.7   <2e-16 ***
17 distance     2.820e-02  9.206e-05   306.4   <2e-16 ***
18 weight       7.252e-01  9.479e-03    76.5   <2e-16 ***
19 ---
20 Signif. codes:
21 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
22
23 Residual standard error: 6.89 on 99997 degrees of freedom
24 Multiple R-squared:  0.499, Adjusted R-squared:  0.499
25 F-statistic: 4.98e+04 on 2 and 99997 DF, p-value: < 2.2e-16

```

L'appel de la fonction `lm` [31] est assez rudimentaire. Il suffit de fournir une formule de régression contenant les variables explicatives avec lesquelles nous tenons à faire la régression et nous spécifions le nom de la table contenant ces variables. Nous remarquons ici la technique du retour multiple abordée à la 2.2. Nous voyons aussi que pour chaque coefficient un test de Student a été effectué pour déterminer à quel point l'estimation était significativement différente de 0. D'autre part, le test de Fisher permet de savoir s'il existe réellement une relation entre les variables explicatives choisies et la variable réponse analysée. [12]

Lorsque l'on compare les valeurs réellement utilisées dans le A et les coefficients es-

timés, nous voyons que ces derniers sont très proches les uns des autres. La [Tableau 2.2](#) fait la revue de ces valeurs.

Variable	Valeur réelle	Valeur estimée
distance	0.0275	0.0282
poids	0.7	0.7252

TABLE 2.2 – Comparaison entre les coefficients réels et estimés par régression linéaire



Lire des tables directement sur le web

Afin de récupérer les valeurs sur les niveaux de taxe pour chaque province canadienne, nous avons pris l'initiative de passer directement via le web. Cette méthode possède l'avantage de se mettre à jour directement avec l'information la plus récente si la structure de la page n'est pas modifiée. Afin de parvenir à ce résultat, les paquetages `XML`, `RCurl` et `rlist` fournissent des fonctions permettant d'interpréter la structure HTML d'une page web spécifiée par le passage du chemin `url` en argument à la fonction `readHTMLTable` pour y détecter les occurrences de balises du genre

`<table><it>...</it></table>`

. http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/XML/html/readHTMLTable.html

2.5 Ajustement de distributions statistiques sur données empiriques

En plus des capacités statistiques impressionnantes que nous avons survolées à la section précédente, R dispose d'une vaste gamme d'outils d'optimisation. Pour ne pas trop nous écarter du but premier de cette documentation, soit de faire une revue globale des fonctionnalités de R en utilisant une étude de cas à titre de support de présentation, nous concentrerons la discussion autour des fonctions `optim` [32] et `fitdistr` [38]. Nous terminerons en présentant comment répliquer le comportement de la fonction `fitdistr` dans le cadre d'une fonction utilitaire.

La fonction `optim` est un excellent choix de fonction pour aborder tout problème d'optimisation. Contrairement à bien d'autres outils, cette fonction permettra d'optimiser plusieurs paramètres à la fois. Elle imposera tout de même quelques limitations telles que l'impossibilité de facilement préciser un intervalle d'optimisation, le fait qu'elle cherchera toujours le minimum et que nous devrons lui fournir un point de

départ. Tous ces désavantages seront toutefois contrebalancés par la possibilité d'optimiser plusieurs paramètres à la fois. [41]

Parmi les arguments de la fonction `optim`, nous devons minimalement désigner les valeurs de départ à nos paramètres avec `par` et fournir à `fn` la fonction qui devra être minimisée. Il sera possible de définir des bornes aux valeurs optimisées grâce aux arguments `lower` et `upper`. Le [Code Source 2.25](#) illustre une application standard de cette fonction. Vous ne serez pas surpris de rencontrer à nouveau la technique du retour multiple au sein d'une liste. De cette liste, nous utiliserons principalement les attributs `par` et `value`. Ceux-ci nous donneront accès aux paramètres optimisés et à la valeur de convergence obtenue. Il sera conseillé de garder un oeil sur `convergence` qui indiquera si l'optimisation s'est terminée de manière conforme (valeur de 0) ou que le processus d'optimisation n'est pas parvenu à converger (valeur de 1). La valeur de `counts` témoigne du nombre d'itérations effectuées afin d'arriver au résultat affiché. Par défaut, la fonction `optim` arrêtera au compte de 501 itérations après quoi les valeurs actuelles de l'optimisation seront renvoyées en plaçant toutefois la valeur de l'attribut `convergence` à 1.

Code Source 2.25 – Optimisation générique avec R

```
1 > f1 <- function(x,y) 5*x**2 - 7*y + 10
2 > f2 <- function(x,y) 10*x**2 + 30*y -2
3 > foptim <- function(x1,x2) (f1(x1,x2) - f2(x1,x2))**2
4 > (results <- optim(par = c(4,5), function(par) foptim(par[1], par
5   [2])))
6 $par
7 [1] 0.4532121 0.2968149
8 $value
9 [1] 8.385268e-05
10
11 $counts
12 function gradient
13      57      NA
14
15 $convergence
16 [1] 0
17
18 $message
19 NULL
20
21 > f1(results$par[1], results$par[2])
22 [1] 8.949302
23 > f2(results$par[1], results$par[2])
24 [1] 8.958459
```

Malgré le fait que nous ayons mentionné des limitations à la fonction `optim`, cela ne signifie pas pour autant que nous ne pourrions pas imaginer des manières de contourner ces dernières. En effet, une maximisation revient tout simplement à trouver la valeur minimale de l'inverse de la fonction étudiée. Ainsi, le simple ajout d'un signe de négation devant la fonction passée à l'argument `fn` nous permettra d'effectuer une maximisation plutôt qu'une minimisation. Il s'agit là de la stratégie que nous avons

empruntée dans le [Code Source 2.26](#).

Il n'est pas rare que plus d'une solution soit viable aux yeux d'un processus d'optimisation dépendamment du problème éludé. Nous appelons ces nombreuses solutions des extremums locaux. C'est l'existence de ces extremums qui rend les valeurs initiales de l'optimisation si sensibles. Lorsque possible, il sera donc fortement conseillé de procéder à des techniques de validation graphique comme nous l'avons fait dans le cadre du [Code Source 2.26](#) (voir [Figure 2.10](#)).

Code Source 2.26 – Maximisation d'une fonction avec `optim`

```
1 > f3 <- function(x,y) -x**2 - 2*y**2 + 3*x + 4*y - 5
2 > (results <- optim(par = c(0,0), function(par) -f3(par[1],par[2]))
3   )
4 $par
5 [1] 1.5001064 0.9999031
6 $value
7 [1] 0.75
8
9 $counts
10 function gradient
11      69      NA
12
13 $convergence
14 [1] 0
15
16 $message
17 NULL
18
19 > # install.packages("rgl")
20 > library(rgl)
21 > persp3d(f3,xlim = c(-5,5),ylim = c(-5,5))
```



Contrôler l'incontrôlable

Bien que vous n'aurez pas à modifier le comportement par défaut de la fonction `optim` pour parvenir à vos fins, il est important de savoir que la fonction propose plusieurs arguments qui permettent d'influencer la manière que l'optimisation sera effectuée. Nous pouvons rapidement citer les arguments `method` et `control`. Veuillez vous référer à la documentation officielle pour de plus amples détails à leur sujet.

<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html>

Une autre fonction d'intérêt lorsque nous travaillons avec des distributions statistiques est `fitdistr` provenant du paquetage `MASS`. Celle-ci permet de facilement ajuster une distribution donnée à un jeu de données empirique. Évidemment, nous pourrions très bien passer par `optim` pour réaliser le même travail moyennant un cer-

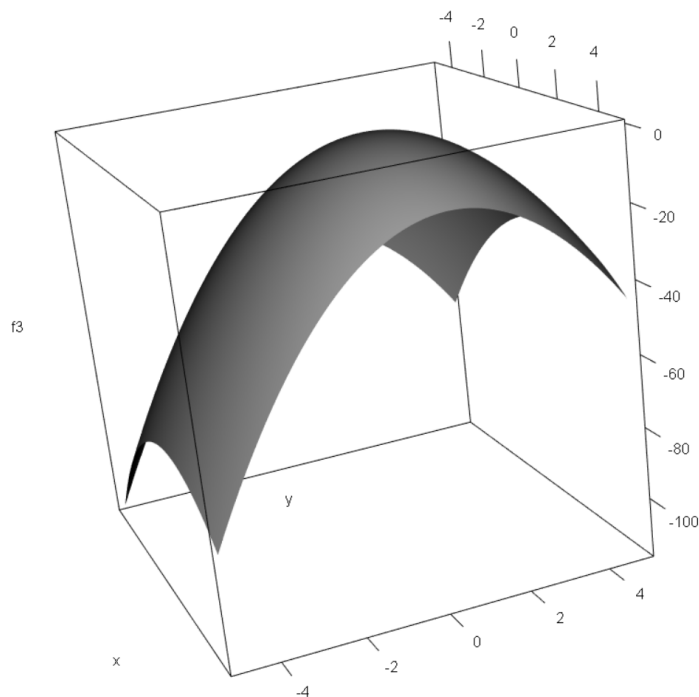


FIGURE 2.10 – Représentation graphique de la fonction **f3**

tain coût de complexité. Or, ce mal sera parfois nécessaire puisque la fonction **fitdistr** n'est définie que pour les distributions suivantes : [39]

- | | |
|----------------------|---------------|
| ► Bêta | ► Géométrique |
| ► Binomiale négative | ► Log-Normale |
| ► Cauchy | ► Logistique |
| ► Khi carrée | ► Normale |
| ► Exponentielle | ► Poisson |
| ► F (Fisher) | ► T (Student) |
| ► Gamma | ► Weibull |

Il arrivera donc dans certains cas que nous devrons procéder à l'ajustement des distributions par la méthode du maximum de vraisemblance directement avec **optim**. Nous prioriserons toutefois l'utilisation de **fitdistr**.

L'appel de **fitdistr** se fera la majorité du temps en passant un vecteur de données sur lesquelles ajuster la distribution et en précisant le nom de la distribution à ajuster. Ce qui présente un net avantage en terme de simplicité par rapport à l'appel

de la fonction `optim` qui permettra d'accomplir le même travail. Le [Code Source 2.27](#) présente l'utilisation de ces deux méthodes.

Code Source 2.27 – Ajustement de distribution sur données empiriques

```

1 > x <- rgamma(1000,40,3)
2 > optim(c(10,1),function(par) -sum(dgamma(x,par[1],par[2],log =
   TRUE)))
3 $par
4 [1] 37.216936 2.785522
5
6 $value
7 [1] 2193.865
8
9 $counts
10 function gradient
11 69 NA
12
13 $convergence
14 [1] 0
15
16 $message
17 NULL
18
19 > #install.packages("MASS")
20 > library(MASS)
21 > fitdistr(x,"gamma")
22      shape      rate
23 37.1275898 2.7787518
24 ( 1.6529478) ( 0.1245497)

```

Les mordus de statistiques parmi vous auront constaté que les distributions reconnues par `fitdistr` ne nécessitent pas toujours le même nombre de paramètres. Il s'agit là d'une complexité algorithmique de bonne taille. Dans le cadre de l'étude de cas, nous avons cru bon de créer une réplique de cette fonction afin d'expliquer comment nous pouvons nous y prendre pour créer des fonctions aussi flexibles. Voici pour commencer le code source de cette fameuse fonction :

Code Source 2.28 – Réplicat maison de la fonction `fitdistr`

```

1 #' Generic function for statistical distribution adjustment
2 #'
3 #' @param data A vector of value over which we want to fit the
   distribution
4 #' @param dist The distribution name
5 #' @param ... The initial values to be given to the optimisation
   function
6 #' @return A list containing :
7 #' the optimized parameters,
8 #' the error value,
9 #' the deviance measure,
10 #' the convergence indicator and
11 #' the number of iterations necessited
12 #' @examples

```

```

13 #' x <- rnorm(1000,10,2)
14 #' distFit(x,"Normal", 1, 1)
15 #' x <- rgamma(1000,5,1)
16 #' distFit(x,"Gamma", 1, 1)
17 #'
18 distFit <- function(data,dist,...)
19 {
20   dist = tolower(dist)
21   args = list(...)
22   if(dist == "normal")
23   {
24     law = "norm"
25     nbparam = 2
26   }
27   else if(dist == "exp")
28   {
29     law = "exp"
30     nbparam = 1
31     lower = 0
32     upper = 100/mean(data)
33   }
34   else if(dist == "gamma")
35   {
36     law = "gamma"
37     nbparam = 2
38   }
39   else if(dist == "lognormal")
40   {
41     law = "lnorm"
42     nbparam = 2
43   }
44   else if(dist == "weibull")
45   {
46     law = "weibull"
47     nbparam = 2
48   }
49   else if(dist == "pareto")
50   {
51     law = "pareto"
52     nbparam = 2
53   }
54   else if(dist == "invgaussian")
55   {
56     law = "invgauss"
57     nbparam = 2
58   }
59   else if(dist == "student")
60   {
61     law = "t"
62     nbparam = 1
63     lower = 0
64     upper = 100/mean(data)
65   }
66   else if(dist == "burr")
67   {
68     law = "burr"
69     nbparam = 3

```

```

70 }
71 else
72 {
73   message <- "The only distributions available are:
74   Normal, Exp, Gamma, LogNormal, Weibull, Pareto, Student, Burr
75   and InvGaussian.
76   (This case will be ignored)"
77   stop(message)
78 }
79 if(nbparam != length(args))
80 {
81   message <- paste("There is a mismatch between the number of
82   arguments passed to the
83   function and the number of arguments needed to
84   the distribution.",
85   "The",dist,"distribution is taking",nbparam,"
86   parameters and",
87   length(args),"parameters were given.")
88   stop(message)
89 }
90 # Treament
91 if(nbparam == 1)
92 {
93   param <- optim(par = args, function(par)
94     -sum(do.call(eval(parse(text = paste("d",law,sep=""))),c(list
95       (data),par,log = TRUE)))),
96     method = "Brent", lower = lower, upper = upper)
97 }
98 else{
99   param <- optim(par = args, function(par)
100     -sum(do.call(eval(parse(text = paste("d",law,sep=""))),c(list
101       (data),par,log = TRUE))))
102 }
103 # Deviance value
104 devValue <- sum((empPDF(x <- seq(0,max(data),0.1))-do.call(eval(
105   parse(text = paste("d",law,sep=""))),c(list(x),param$par)))*
106   2)
107 # Return List
108 distFitList <- list()
109 distFitList$param <- param$par
110 distFitList$errorValue <- param$value
111 distFitList$devValue <- devValue
112 distFitList$convergence <- param$convergence
113 distFitList$nbiter <- param$counts[1]
114 distFitList
115 }

```

Le [Code Source 2.28](#) peut sembler impressionnant à première vue, mais environ 90% de son corps ne sert qu'à faire de la gestion d'erreurs. Comme indiqué par les commentaires internes, les lignes de commandes renfermant le secret de ce type de fonction sont les suivantes :

```

param <- optim(par = args, function(par) -sum(do.call(eval(parse(text =
paste("d",law,sep=""))),c(list(data),par,log = TRUE))))

```

Sans trop rentrer dans les détails, la fonction `parse` permettra de créer des expressions non évaluée. Il existe plusieurs manières de générer ces expressions. Celle employée dans le cas présent se fera à partir d'un vecteur de caractères qui nous permettra de concaténer le "d" de la fonction de densité à l'identifiant R de la distribution choisie (Voir [Tableau 2.4](#) $\langle ID_R \rangle$). Une fois cette expression construite, nous pourrons la faire évaluer par R grâce à la fonction `eval` qui transformera la ligne de code en un objet (étant ici la fonction de densité de la distribution choisie). À cet objet, nous pourrons désormais lui fournir des paramètres au même titre que nous le ferions avec la fonction de densité correspondante. Alors pourquoi avons-nous senti le besoin d'utiliser `do.call`? La fonction `do.call` permet la possibilité de faire l'appel d'une fonction en s'occupant de lui fournir une liste de paramètres de taille quelconque pour autant que la fonction réceptrice accepte autant d'arguments que fournis et de type correspondant. Étant donné que le nombre de paramètres de nos distributions peut varier, nous n'aurions pas pu envisager de créer un traitement particulier pour tous les cas possibles.

Code Source 2.29 – Exemple d'utilisation de la fonction `distFit`

```

1 > x <- rexp(10000,4)
2 > distFit(x, "Exp",1)$param
3 [1] 4.102991
4 > x <- rt(10000,5)
5 > distFit(x, "Student",1)$param
6 [1] 5.056508
7 > x <- rgamma(10000,4,2)
8 > distFit(x, "Gamma",1,1)$param
9 [1] 3.982845 1.981206
10 > x <- rburr(10000,1,10,0.01)
11 > distFit(x, "Burr",0.9,1,0.1)$param
12 [1] 0.95531981 10.09683646 0.01006351
13 Warning messages:
14 1: In (function (x, shape1, shape2, rate = 1, scale = 1/rate, log =
      FALSE) :
15   NaNs produced
16 2: In (function (x, shape1, shape2, rate = 1, scale = 1/rate, log =
      FALSE) :
17   NaNs produced
18 3: In (function (x, shape1, shape2, rate = 1, scale = 1/rate, log =
      FALSE) :
19   NaNs produced
20 4: In (function (x, shape1, shape2, rate = 1, scale = 1/rate, log =
      FALSE) :
21   NaNs produced

```

Comme nous venons de le voir, la combinaison de ces trois fonctions ouvre les portes à un autre niveau de flexibilité pour la définition de fonctions utilitaires. Grâce à cet exemple, nous comprenons désormais un peu mieux la mécanique sous-entendue par le passage de paramètres additionnels via l'argument (...).

2.6 Calcul stochastique en R

Quand bien même que la génération de nombres aléatoires ai déjà été abordée à la [Tableau 2.4](#), il serait incorrect de s'imaginer que les capacités statistiques de R

s'arrête là. R est un excellent langage pour faire du calcul stochastique. Qui dit calcul stochastique dit aussi estimation par simulation d'un grand nombre d'observations pour estimer le comportement d'un phénomène difficilement quantifiable de manière déterministe.

La première fonction à connaître lorsque nous abordons une analyse de ce genre est la fonction `sample` [43]. Cette dernière sera utile dans les cas où nous cherchons à faire une pige aléatoire de taille quelconque (`size`) sur un ensemble de valeurs fourni par un vecteur. Il sera possible de préciser si nous voulons faire une pige avec ou sans remise avec l'argument `replace` ainsi la probabilité que chaque valeur survienne grâce à `prob`. Un aspect fort intéressant de cette fonction et sa capacité de faire des piges sur des valeurs textuelles. Le [Code Source 2.30](#) fait une revue de l'utilisation de la fonction `sample`. Lors du deuxième appel de la fonction, nous remarquons la génération de valeurs beaucoup plus élevées par rapport au premier appel. Toutefois, la seule différence a été de modifier la valeur de l'argument `prob` pour y assigner le poids relatif de l'altitude sur l'ensemble des altitudes favorisant ainsi les valeurs extrêmes. Le troisième appel expose, quant à lui, la capacité de travailler avec un vecteur de valeurs textuelles.

Code Source 2.30 – Pige aléatoire sur support vectoriel

```

1 > sample(airportsCanada$altitude, size = 10, replace = TRUE)
2 [1] 1211 24 1215 2351 1873 882 256 2314 2968 728
3 > probs = airportsCanada$altitude/sum(airportsCanada$altitude)
4 > sample(airportsCanada$altitude, size = 10, replace = TRUE, prob =
      probs)
5 [1] 2525 2264 2680 1220 1087 2277 4296 1536 1712 2000
6 > sample(unique(as.character(paste(airportsCanada$name))), size =
      10, replace = FALSE)
7 [1] "Fort Severn Airport"
8 [2] "CFB Trenton"
9 [3] "Waterville / Kings County Municipal Airport"
10 [4] "Salluit Airport"
11 [5] "Forestville Airport"
12 [6] "Taloyoak Airport"
13 [7] "Sandspit Airport"
14 [8] "Mary's Harbour Airport"
15 [9] "Pukatawagan Airport"
16 [10] "Deer Lake Airport"

```

En inspectant le [A](#), nous constatons la structure fonctionnelle et imbriquée du processus entrepris. Il sera fortement conseillé de procéder ainsi pour différentes raisons :

- Augmenter la clarté du processus de simulation ;
- Faciliter le débogage lors du développement ;
- Possibilité de facilement ajouter et retirer des blocs au casse-tête de simulation ;
- Identification simplifiée des parties limitantes et coûteuses en temps de calcul pour des fins d'optimisation ;
- Permettre la production d'une nouvelle itération par l'appel d'une fonction mère ne possédant idéalement aucun argument.

Ce ne sera qu'en présence de cette structure que la fonction `replicate` prendra tout son sens. À l'aide de cette fonction, nous pourrions commodément contrôler le

nombre de répliques effectuées. Dans le [Code Source 2.31](#), nous avons justement pris cette fonctionnalité pour reproduire à 6 reprises la génération de nombres aléatoires suivant une loi $Norm(\mu := 3, \sigma := 4)$.

Code Source 2.31 – Replication d’une analyse stochastique

```
1 fsimul <- function() qnorm(runif(100),3,4)
2 results <- replicate(6,fsimul())
3 g <- rep(c("a", "b", "c", "d", "e", "f"), each = 100)
4 #install.packages("lattice")
5 library(lattice)
6 histogram(~ as.vector(results) | g, xlab = "Results", ylab = "
  Frequency")
```

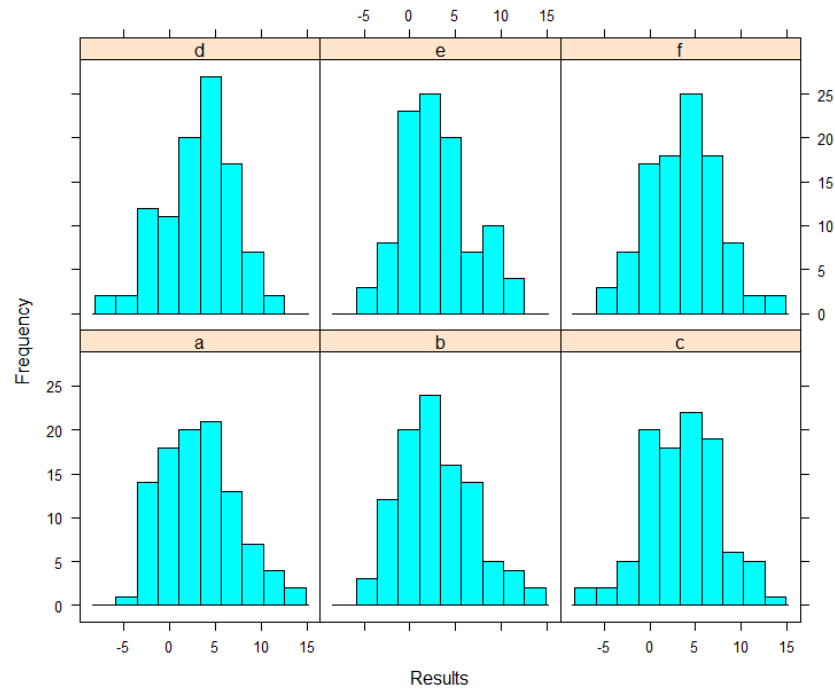


FIGURE 2.11 – Comparaison des résultats d’une analyse stochastique à 6 réplcats



Une "Poisson" dans une pisciculture...

La distribution Poisson sera souvent à la base des processus stochastiques en raison de ses propriétés particulières. Nous parlerons souvent du fait que cette loi ne possède pas de mémoire ce qui implique que le nombre de succès observés sur différents intervalles sera indépendant. Nous pouvons aussi mentionner que la somme des variables aléatoires suivant des lois Poisson indépendantes de paramètres λ_1 et λ_2 suivra à son tour une loi de Poisson de paramètre $\lambda_1 + \lambda_2$.

https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/course-notes/MIT6_262S11_chap02.pdf

https://fr.wikipedia.org/wiki/Loi_de_Poisson

Conclusion

Au terme de cette étude de cas, nous avons su intégrer différentes notions relatives à la programmation en R. Nous avons abordé des sujets aussi variés qu'actuels allant de l'importation des données jusqu'à la simulation stochastique.

Cette formation n'a jamais eu la prétention de pouvoir vous apprendre toutes les particularités du langage R ni même faire de vous des programmeurs parfaitement fonctionnels au terme de sa lecture. Par contre, nous croyons avoir bel et bien accompli l'objectif principal qui était de faire une revue des capacités de R tout en vous offrant un coffre d'outils qui facilitera grandement vos débuts avec ce langage de programmation. Il n'y a pas de secret pour apprendre à programmer, mais il existe certainement des moyens plus simples que d'autres. Selon nous, une connaissance adéquate de ce que l'on peut ou pas réaliser consiste un excellent point de départ. Par après, à un moment ou un autre, vous serez confronté à un problème qui vous semblera parfaitement adapté à l'utilisation d'un outil donné. Vous chercherez ensuite à accumuler les ressources nécessaires à sa résolution. Évidemment, rien ne vous empêche de vous créer des problèmes fictifs comme nous l'avons fait pour ensuite lire plusieurs centaines de pages de documentation pour finalement arriver à vos fins. La route sera souvent tortueuse, mais le résultat donc bien satisfaisant.

Dans une ère aussi axée sur le développement informatique et l'automatisation des tâches, il est de plus en plus important d'avoir de bonnes connaissances dans ces domaines. La connaissance du langage R est sans aucun doute une très bonne idée en raison de sa facilité d'accès, de la taille de sa communauté et sa simplicité. Comme nous pouvons le voir à la [Figure 2.12](#), R est toujours un langage d'actualité très prisé et utilisé qui en vaut le détour en se classant au 12^{ème} rang selon le classement *RedMonk* [17].

En raison du caractère libre du langage R, ce dernier a toujours été et restera en perpétuel développement. C'est la raison principale pourquoi nous parlons toujours de ce langage à l'heure actuelle, tandis que plusieurs autres sont tombés dans les oubliettes. Par contre, un des principes fondamentaux du développement libre implique la coopération de ses utilisateurs. Si nous profitons de ce que la communauté nous apporte, nous devrions aussi être en mesure de contribuer à la communauté lorsque nous pensons avoir réalisé une tâche qui pourra intéresser et être récupérées par d'autres utilisateurs. En ce qui nous concerne, sans l'accès aux données d'*OpenFlights*, la totalité de cette étude n'aurait pas pu être réalisée. En travaillant avec ces données, nous avons eu à faire un peu de reconstitution au niveau des fuseaux horaires. C'est pour cette raison qu'une contribution de notre part sera effectuée directement via le GitHub

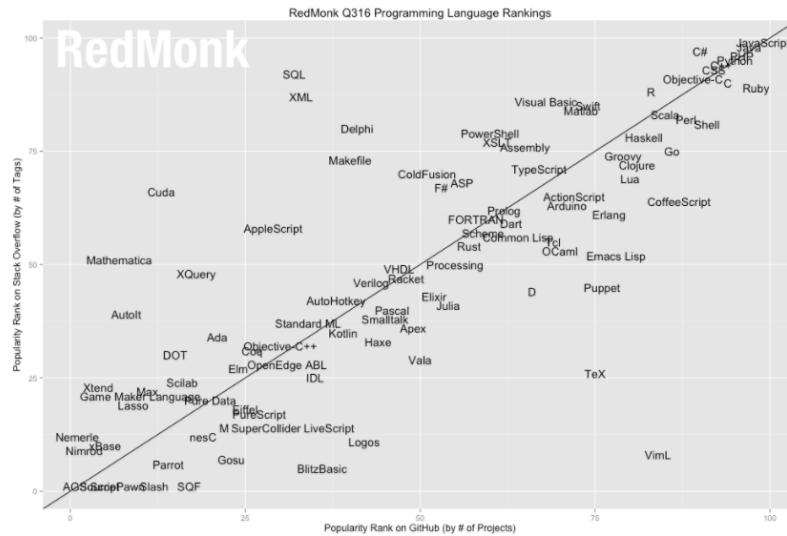


FIGURE 2.12 – Classement *RedMonk* des différents langages de programmation

du projet *OpenFlights* pour regarnir la variable `tzFormat`.

En guise de conclusion, je tiens à remercier David Beauchemin et Vincent Goulet pour leur support tout au long de l'écriture de ce document et sans qui je ne serais certainement pas parvenu à écrire le tout dans un si petit laps de temps. À vous cher, acolytes, en espérant retravailler dans un avenir rapproché!

Bibliographie

- [1] A quoi correspondent les extensions *.shp, *.dbf, *.prj, *.sbn, *.sbx et *.shx? <http://www.portailsig.org/content/quoi-correspondent-les-extensions-shp-dbf-prj-sbn-sbx-et-shx>.
- [2] CSV vs. Delimited Flat Files : How to Choose. <http://www.thoughtspot.com/blog/csv-vs-delimited-flat-files-how-choose>.
- [3] Doxygen. <http://www.stack.nl/~dimitri/doxygen/>.
- [4] Font Awesome - The iconic font and CSS toolkit. <http://fontawesome.io/>.
- [5] GitHub. <https://github.com/>.
- [6] Introduction à la programmation en R. https://cran.r-project.org/doc/contrib/Goulet_introduction_programmation_R.pdf.
- [7] Introduction à R - Atelier du colloque R à Québec 2017 (GitHub). <https://github.com/vigou3/raquebec-intro>.
- [8] OpenFlights. <https://openflights.org/data.html>.
- [9] Package 'leaflet'. <https://cran.r-project.org/web/packages/leaflet/leaflet.pdf>.
- [10] Probabilités et Statistique avec R. <http://ljk.imag.fr/membres/Bernard.Ycart/mel/dr/node7.html>.
- [11] Projection (Système de). <http://www.emse.fr/tice/uved/SIG/Glossaire/co/Projection.html>.
- [12] Quick Guide : Interpreting Simple Linear Model Output in R. <http://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R>.
- [13] R à Québec 2017. <http://raquebec.ulaval.ca/2017/programme-r-a-quebec-2017>.
- [14] roxygen2. <http://roxygen.org/>.
- [15] Structured Query Language (SQL). https://fr.wikipedia.org/wiki/Structured_Query_Language.
- [16] Tests statistiques avec R. <http://www.sthda.com/french/wiki/tests-statistiques-avec-r>.
- [17] Top Programming Languages to Learn in 2017. <https://www.codingame.com/blog/top-programming-languages-to-learn-in-2017/>.
- [18] Roger Bivand, Tim Keitt, and Barry Rowlingson. *Bindings for the Geospatial Data Abstraction Library*, 2017. <https://cran.r-project.org/web/packages/rgdal/index.html>.

- [19] Statistics Canada. Boundary Files, Reference Guide. <http://www.statcan.gc.ca/pub/92-160-g/92-160-g2011002-eng.htm>.
- [20] Vincent Goulet and Mathieu Pigeon. *Actuarial Functions and Heavy Tailed Distributions*, 2017. <https://cran.r-project.org/web/packages/actuar/index.html>.
- [21] G. Grothendieck. *Perform SQL Selects on R Data Frames*, 2014. <https://cran.r-project.org/web/packages/sqldf/index.html>.
- [22] David Kahle and Hadley Wickham. *Spatial Visualization with ggplot2*, 2016. <https://cran.r-project.org/web/packages/ggmap/index.html>.
- [23] Eric Muller. A shapefile of the TZ timezones of the world. <http://efe.le.net/maps/tz/world/>.
- [24] Jani Patokallio. Airport database. <https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat>.
- [25] Jani Patokallio. Route database. <https://raw.githubusercontent.com/jpatokal/openflights/master/data/routes.dat>.
- [26] Edzer Pebesma and Roger Bivand. *Classes and Methods for Spatial Data*, 2016. <https://cran.r-project.org/web/packages/sp/index.html>.
- [27] R documentation. *Add Straight Lines to a Plot*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/abline.html>.
- [28] R documentation. *Cross Tabulation and Table Creation*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/table.html>.
- [29] R documentation. *Data Input*. <https://stat.ethz.ch/R-manual/R-devel/library/utils/html/read.table.html>.
- [30] R documentation. *Draw Function Plots*. <https://www.math.ucla.edu/~anderson/rw1001/library/base/html/curve.html>.
- [31] R documentation. *Fitting Linear Models*. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html>.
- [32] R documentation. *General-purpose Optimization*. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html>.
- [33] R documentation. *Generic X-Y Plotting*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html>.
- [34] R documentation. *Get or Set Working Directory*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/getwd.html>.
- [35] R documentation. *Histograms*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/hist.html>.
- [36] R documentation. *Invoke a Data Viewer*. <https://stat.ethz.ch/R-manual/R-devel/library/utils/html/View.html>.
- [37] R documentation. *Kernel Density Estimation*. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/density.html>.
- [38] R documentation. *Maximum-likelihood Fitting of Univariate Distributions*. <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html>.
- [39] R documentation. *Maximum-likelihood Fitting of Univariate Distributions*. <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html>.

- [40] R documentation. *Object Summaries*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/summary.html>.
- [41] R documentation. *Programmer en R/Optimiser une fonction*. https://fr.wikibooks.org/wiki/Programmer_en_R/Optimiser_une_fonction.
- [42] R documentation. *Random Number Generation*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/Random.html>.
- [43] R documentation. *Random Samples and Permutations*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/sample.html>.
- [44] R documentation. *Return the First or Last Part of an Object*. <https://stat.ethz.ch/R-manual/R-devel/library/utils/html/head.html>.
- [45] R documentation. *Set or Query Graphical Parameters*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/par.html>.
- [46] R documentation. *Subsetting Vectors, Matrices and Data Frames*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/subset.html>.
- [47] Hadley Wickham. *Tools for Splitting, Applying and Combining Data*, 2016. <https://cran.r-project.org/web/packages/plyr/index.html>.

Annexe A

Code source du projet

Cette annexe présente les codes sources constituant l'ensemble du projet. Ceux-ci se divisent sous la forme de 6 parties correspondant aux différents thèmes abordés dans le présent document.

Code Source A.1 – Benchmark.R

```
1 #### RStudio: -*- coding: utf-8 -*-
2 ##
3 ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
   Goulet
4 ##
5 ## This file is part of the project
6 ## "Introduction a R – Atelier du colloque R a Quebec 2017"
7 ## http://github.com/vigou3/raquebec-atelier-introduction-r
8 ##
9 ## The creation is made available according to the license
10 ## Attribution–Sharing in the same conditions 4.0
11 ## of Creative Commons International
12 ## http://creativecommons.org/licenses/by-sa/4.0/
13
14 # Source code for the creation of the benchmark.csv file
15
16 # Setting working directory properly
17 setwd('C:/Users/Samuel/Documents/ColloqueR/Dev')
18 getwd()
19 setwd("..")
20 (path <- getwd())
21
22 # Parameters of the simulation
23 n <- 100000
24 x <- matrix(c(runif(2*n)), ncol = 2, byrow = TRUE)
25
26 # Generate weights with a LogNormal distribution
27 mul <- log(3000)
28 sigma1 <- log(1.8)
29 exp(mul+sigma1**2/2)
30 exp(2*mul+4*sigma1**2/2)-exp(mul+sigma1**2/2)**2
31 weights <- round(qlnorm(x[,1], mul, sigma1)/1000, 3)
```

```

32 hist(weights,breaks = 100,freq=FALSE)
33 mean(weights)
34 max(weights)
35
36 # Generate the errors on the weights
37 weightsTarifParam <- 0.7
38 weightsCost <- weights*weightsTarifParam
39 weightsError <- rnorm(n,mean(weightsCost),sd(weightsCost))
40 max(weightsError)
41 min(weightsError)
42 weightsFinalPrice <- weightsCost+weightsError
43 mean(weightsFinalPrice)
44 min(weightsFinalPrice)
45 var(weightsFinalPrice)
46
47 # Generate the distance with a LogNormal distribution
48 # routesCanada
49 # routesIATA <- cbind(paste(routesCanada$sourceAirport),paste(
50   routesCanada$destinationAirport))
51 # routesDistance <- apply(routesIATA, 1, function(x) airportsDist(x
52   [1],x[2])$value)
53 # max(routesDistance)
54 # mean(routesDistance)
55 mu2 <- log(650)
56 sigma2 <- log(1.4)
57 (distances <- round(qlnorm(x[,2],mu2,sigma2)))
58 hist(distances,breaks = 100,freq=FALSE)
59 mean(distances)
60 max(distances)
61
62 # Generate the errors on the distances
63 distancesTarifParam <- 0.0275
64 distancesCost <- distances*distancesTarifParam
65 distancesError <- rnorm(n,mean(distancesCost),sd(distancesCost))
66 distancesFinalPrice <- distancesCost+distancesError
67 mean(distancesFinalPrice)
68 var(distancesFinalPrice)
69 max(distancesFinalPrice)
70 min(distancesFinalPrice)
71
72 # Generate total price
73 baseCost <- 10
74 # taxRate <- sum(table(airportsCanada$province)*as.numeric(paste(
75   taxRates$taxRate)))/length(airportsCanada$province)
76 taxRate <- 1.082408
77 profitMargin <- 1.15
78 (totalCost <- round((baseCost + weightsFinalPrice +
79   distancesFinalPrice)*profitMargin*taxRate,2))
80 mean(totalCost)
81 var(totalCost)
82 max(totalCost)
83 min(totalCost)
84
85 # Export to csv format
86 (dataExport <- cbind(weights,distances,totalCost))
87 colnames(dataExport) <- c("Poids (Kg)","Distance (Km)","Prix (CAD $
88   )")

```

```

84
85 write.csv(dataExport, paste(path, "/Reference/benchmark.csv", sep=""),
      row.names = FALSE, fileEncoding = "UTF-8")

```

Code Source A.2 – CaseStudyDevQ1.R

```

1  ### RStudio: -*- coding: utf-8 -*-
2  ##
3  ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
   Goulet
4  ##
5  ## This file is part of the project
6  ## "Introduction a R – Atelier du colloque R a Quebec 2017"
7  ## http://github.com/vigou3/raquebec-atelier-introduction-r
8  ##
9  ## The creation is made available according to the license
10 ## Attribution–Sharing in the same conditions 4.0
11 ## of Creative Commons International
12 ## http://creativecommons.org/licenses/by-sa/4.0/
13
14 ##### Setting working directory properly #####
15 setwd("C:/Users/Samuel/Documents/ColloqueR/Dev")
16 getwd()
17 setwd("..")
18 (path <- getwd())
19 set.seed(31459)
20
21
22 # Extraction of airports.dat and routes.dat
23 airports <- read.csv("https://raw.githubusercontent.com/jpatokal/
   openflights/master/data/airports.dat",
24                     header = FALSE, na.strings=c("\\N",""))
25 routes <- read.csv("https://raw.githubusercontent.com/jpatokal/
   openflights/master/data/routes.dat",
26                    header = FALSE, na.strings=c("\\N",""))
27
28 # Columns names assignation based on the information available on
   the website
29 # https://openflights.org/data.html
30 colnames(airports) <- c("airportID", "name", "city", "country", "
   IATA", "ICAO",
31                        "latitude", "longitude", "altitude", "
   timezone", "DST",
32                        "tzFormat", "typeAirport", "Source")
33 colnames(routes) <- c("airline", "airlineID", "sourceAirport", "
   sourceAirportID",
34                      "destinationAirport", "destinationAirportID", "
   codeshare",
35                      "stops", "equipment")
36
37 # Filtering the observations relative to Canadian airports
38 airportsCanada <- subset(airports, country == "Canada")
39
40 # Extraction of general information about the variables contained
   in the dataset
41 View(airportsCanada)

```



```

42 summary(airportsCanada)
43 nbAirportCity <- table(airportsCanada$city)
44 (nbAirportCity <- head(sort(nbAirportCity, decreasing=TRUE)))
45
46 # Variable selection
47 # We will not use the typeAirport and Source variables since we
   only want to analyse air transport market
48 # We can also discard the country variable because we already
   filtered on Canadian airports
49 airportsCanada <- subset(airportsCanada, select = -c(country,
   typeAirport, Source))
50
51 # As seen in the summary, we do not have the IATA code of 27
   airports
52 subset(airportsCanada, is.na(IATA), select = c("airportID", "name", "
   IATA", "ICAO"))
53
54 # 82% of the time, the IATA code corresponds to the last three
   characters of the ICAO code
55 # We will use this relationship to assign default value for missing
   IATA codes
56 IATA <- as.character(airportsCanada$IATA)
57 ICAO <- as.character(airportsCanada$ICAO)
58 i <- is.na(IATA)
59 sum(IATA == substr(ICAO, 2, 4), na.rm = TRUE) / sum(!i)
60 IATA[i] <- substr(ICAO[i], 2, 4)
61 airportsCanada$IATA <- as.factor(IATA)
62 summary(airportsCanada)
63 # We will not need the ICAO code anymore
64 airportsCanada <- subset(airportsCanada, select = - ICAO)
65
66 # Finally, we are missing more than fifty time zone.
67 missingTZ <- airportsCanada[is.na(airportsCanada$timezone),]
68
69 # Since the TZ depend only on the geographical position we will
   locate the real time zone
70 # by mapping tools.
71
72 # install.packages("sp")
73 # install.packages("rgdal")
74 library(sp)
75 library(rgdal)
76 tz_world.shape <- readOGR(dsn=paste(path, "/Reference/tz_world", sep=
   ""), layer="tz_world")
77 unknown_tz <- airportsCanada[is.na(airportsCanada$timezone), c("
   airportID", "name", "longitude", "latitude")]
78 sppts <- SpatialPoints(unknown_tz[, c("longitude", "latitude")])
79 proj4string(sppts) <- CRS("+proj=longlat")
80 sppts <- spTransform(sppts, proj4string(tz_world.shape))
81 merged_tz <- cbind(unknown_tz, over(sppts, tz_world.shape))
82
83 # To retrieved the province of each airport, we will used the same
   mapping tools
84 prov_terr.shape <- readOGR(dsn=paste(path, "/Reference/prov_terr",
   sep=""), layer="gpr_000b11a_e")
85 unknown_prov <- airportsCanada[, c("airportID", "city", "longitude", "
   latitude")]

```

```

86 sppts <- SpatialPoints(unknown_prov[,c("longitude", "latitude")])
87 proj4string(sppts) <- CRS("+proj=longlat")
88 sppts <- spTransform(sppts, proj4string(proj_terr.shape))
89 merged_prov <- cbind(airportsCanada, over(sppts, proj_terr.shape))
90
91 # install.packages("sqldf")
92 library(sqldf)
93 airportsCanada <- sqldf("
94   select
95     a.*,
96     coalesce(a.tzFormat, b.TZID) as tzMerged,
97     c.PRENAME as provMerged
98   from airportsCanada a
99   left join merged_tz b
100   on a.airportID = b.airportID
101   left join merged_prov c
102   on a.airportID = c.airportID
103   order by a.airportID")
104 airportsCanada <- data.frame(as.matrix(airportsCanada))
105
106 # Since the timezone, DST and city are now useless, we remove them
107 # from the dataset.
108 # Since we have a complete data for the tz, we delete the tzFormat
109 # and will replace it with
110 # tzmerge
111 airportsCanada <- subset(airportsCanada, select = -c(timezone, DST,
112   tzFormat, city ))
113 summary(airportsCanada)
114
115 # install.packages("plyr")
116 library(plyr)
117 airportsCanada <- rename(airportsCanada, c("tzMerged"="tzFormat", "
118   provMerged"="province"))
119 summary(airportsCanada)
120 routesCanada <- sqldf("
121   select *
122   from routes
123   where sourceAirportID in (select distinct airportID
124     from airportsCanada)
125     and destinationAirportID in (select distinct airportID
126     from airportsCanada)")
127 routesCanada <- data.frame(as.matrix(routesCanada ))
128
129 # This code will give the same result :
130 # x <- routesCanada[!is.na(match(routesCanada$sourceAirportID,
131   airportsCanada$airportID)) &
132   !is.na(match(routesCanada$destinationAirportID,
133     airportsCanada$airportID)),]
134 # routesCanada <- routesCanada[!is.na(match(routesCanada$
135   sourceAirport, airportsCanada$IATA)) &
136   !is.na(match(routesCanada$destinationAirport,
137     airportsCanada$IATA)),]
138
139 summary(routesCanada)
140 unique(routesCanada$airline)
141 unique(routesCanada[,c("airline", "airlineID")])
142 unique(routesCanada$airlineID)

```

```

135 routesCanada[is.na(routesCanada$airlineID),]
136 unique(routesCanada$airlineID)
137 unique(routesCanada[is.na(routesCanada$airlineID),]$airline)
138 summary(routesCanada$stops)
139
140 # Since almost all routes are straight routes, we dont need the
    codeshare, and stops variables.
141 routesCanada <- subset(routesCanada, select = -c(codeshare, stops))
142 summary(routesCanada)
143
144 # Create a map showing the different airports
145 # install.packages("ggmap")
146 library(ggmap)
147 map <- get_map(location = "Canada", zoom = 3)
148 lon <- as.numeric(paste(airportsCanada$longitude))
149 lat <- as.numeric(paste(airportsCanada$latitude))
150 airportsCoord <- as.data.frame(lon, lat)
151 (mapPoints <- ggmap(map) + geom_point(data=airportsCoord, aes(lon,
    lat), alpha=0.5))
152
153 # Create a second map showing all possible routes between these
    different airports.
154 summary(routesCanada)
155 summary(airportsCanada)
156 routesCoord <- sqldf("
157     select
158         a.sourceAirport,
159         a.destinationAirport,
160         b.longitude as sourceLon,
161         b.latitude as sourceLat,
162         c.longitude as destLon,
163         c.latitude as destLat
164     from routesCanada a
165     left join airportsCanada b
166         on a.sourceAirport = b.IATA
167     left join airportsCanada c
168         on a.destinationAirport = c.IATA")
169 lonBeg <- as.numeric(paste(routesCoord$sourceLon))
170 latBeg <- as.numeric(paste(routesCoord$sourceLat))
171 lonEnd <- as.numeric(paste(routesCoord$destLon))
172 latEnd <- as.numeric(paste(routesCoord$destLat))
173 routesCoord <- as.data.frame(cbind(lonBeg, latBeg, lonEnd, latEnd))
174 (mapRoutes <- mapPoints + geom_segment(data=routesCoord, aes(x=
    lonBeg, y=latBeg, xend=lonEnd, yend=latEnd), alpha=0.5))
175
176 # Calculate an airport ridership index based on the number of
    incoming routes.
177 arrivalFlights <- table(routesCanada$destinationAirport)
178 departureFlights <- table(routesCanada$sourceAirport)
179 totalFlights <- arrivalFlights + departureFlights
180 max(totalFlights)
181 mean(totalFlights)
182 var(totalFlights)
183 sd(totalFlights)
184 head(sort(totalFlights, decreasing = TRUE), n = 30)
185 totalFlightsCDF <- ecdf(totalFlights)
186 IATA <- names(totalFlights)

```

```

187
188 # Index drawing
189 curve(totalFlightsCDF(x-1),from = 0,to = 60,n = 100,
190       xlab = "Nombre de routes par aeroport",
191       ylab = "CDF")
192
193 # Calculate a combined index from the index.
194 combinedIndex <- round(totalFlights/max(totalFlights),3)
195 combinedIndexTable <- data.frame(IATA,
196                                  as.numeric(paste(totalFlights)),
197                                  as.numeric(paste(combinedIndex)))
198 rownames(combinedIndexTable) <- NULL
199 colnames(combinedIndexTable) <- c("IATA","totalFlights","
    combinedIndex")
200 combinedIndexTable
201 airportsCanada <- sqldf("
202   select
203     a.*,
204     coalesce(b.combinedIndex,0) as combinedIndex
205   from airportsCanada a
206   left join combinedIndexTable b
207   on a.IATA = b.IATA")
208 airportsCanada <- data.frame(as.matrix(airportsCanada ))
209
210 # Create maps to visualize these indices using a bubble graph.
211 TrafficData <- subset(airportsCanada ,as.numeric(paste(combinedIndex)
    ) > 0.05)
212 lon <- as.numeric(paste(TrafficData$longitude))
213 lat <- as.numeric(paste(TrafficData$latitude))
214 size <- as.numeric(paste(TrafficData$combinedIndex))
215 airportsCoord <- as.data.frame(cbind(lon , lat , size))
216 mapPoints <-
217   ggmap(map) +
218   geom_point(data=TrafficData ,aes(x=lon ,y=lat , size=size) ,alpha=0.5,
    shape=16)
219 (mapTraffic <-
220   mapPoints +
221   scale_size_continuous(range = c(0, 20),name = "Traffic Index"))
222
223 # Map with markers of some airports
224 # The markers include the IATA, the airport name, the longitude and
    the latitude
225 # install.packages("leaflet")
226 library(leaflet)
227 url <- "http://hiking.waymarkedtrails.org/en/routebrowser/1225378/
    gpx"
228 download.file(url , destfile = paste(path,"/Reference/worldRoutes.
    gpx",sep=""), method = "wget")
229 worldRoutes <- readOGR(paste(path,"/Reference/worldRoutes.gpx",sep=
    ""), layer = "tracks")
230 markersData <- subset(airportsCanada ,IATA %in% c("YUL","YVR","YYZ",
    "YQB"))
231 markersWeb <- c("https://www.aeroportdequebec.com/fr/pages/accueil"
    ,
    "http://www.admtl.com/",
    "http://www.yvr.ca/en/passengers",
    "https://www.torontopearson.com/")
232
233
234

```

```

235
236 # Defining the description text to be displayed by the markers
237 descriptions <-paste("<b><FONT COLOR=#31B404> Airport Details</FONT  

    ></b> <br>",
238                      "<b>IATA: <a href=",markersWeb,">",markersData$IATA,"</a></b><br>",
239                      "<b>Name:</b>",markersData$name,"<br>",
240                      "<b>Coord.</b>: (",markersData$longitude,"",
                        markersData$latitude,") <br>",
241                      "<b>Traffic Index:</b>",markersData$combinedIndex)
242
243 # Defining the icon to be add on the markers from fontawesome
    library
244 icons <- awesomeIcons(icon = "paper-plane",
245                        iconColor = "black",
246                        library = "fa")
247
248 # Combinaison of the different components in order to create a
    standalone map
249 (mapTraffic <- leaflet(worldRoutes) %>%
250   addTiles(urlTemplate = "http://{s}.basemaps.cartocdn.com/light_
    all/{z}/{x}/{y}.png") %>%
251   addCircleMarkers(stroke = FALSE,data = TrafficData ,
252                   ~as.numeric(paste(longitude)), ~as.numeric(
    paste(latitude)),
253                   color = "black", fillColor = "green",
254                   radius = ~as.numeric(paste(combinedIndex))*30,
    opacity = 0.5) %>%
255   addAwesomeMarkers(data = markersData , ~as.numeric(paste(
    longitude)),
256                   ~as.numeric(paste(latitude)), popup =
    descriptions ,icon=icons))
257
258 # Resizing of the map
259 mapTraffic$width <- 874
260 mapTraffic$height <- 700
261
262 # Export of the map into html format
263 # install.packages("htmlwidgets")
264 library(htmlwidgets)
265 saveWidget(mapTraffic , paste(path,"/Reference/leafletTraffic.html" ,
    sep = ""), selfcontained = TRUE)
266
267 # addMarkers(data = subset(airportsCanada,IATA %in% c("YUL","YVR","
    YYZ","YQB")),
268 # ~as.numeric(paste(longitude)), ~as.numeric(paste(latitude)),
    popup = ~IATA) %>%

```

Code Source A.3 – CaseStudyDevQ2.R

```

1 #### RStudio: -*- coding: utf-8 -*-
2 ##
3 ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
    Goulet
4 ##

```

```

5 ## This file is part of the project
6 ## "Introduction a R – Atelier du colloque R a Quebec 2017"
7 ## http://github.com/vigou3/raquebec-atelier-introduction-r
8 ##
9 ## The creation is made available according to the license
10 ## Attribution–Sharing in the same conditions 4.0
11 ## of Creative Commons International
12 ## http://creativecommons.org/licenses/by-sa/4.0/
13
14 # Functions creation
15
16 #' Distance calculation function between two airports.
17 #'
18 #' @param sourceIATA The IATA of the departure airport.
19 #' @param destIATA The IATA of the arrival airport.
20 #' @return A list of the distance in Km between sourceIATA and
21 #'         destIATA, the index of the airports and the unit.
22 #' @examples
23 #' airportsDist("YUL","YQB")
24 #' airportsDist("YUL","YVR")
25
26 # install.packages("geosphere")
27 library(geosphere)
28 airportsDist <- function(sourceIATA,destIATA)
29 {
30   # Verification of the sourceIATA and destIATA
31   sourceFindIndex <- match(sourceIATA,airportsCanada$IATA)
32   if(is.na(sourceFindIndex))
33   {
34     stop(paste("sourceIATA :",sourceIATA,"is not a valid IATA code"
35               ))
36   }
37   destFindIndex <- match(destIATA,airportsCanada$IATA)
38   if(is.na(destFindIndex))
39   {
40     stop(paste("destIATA :",destIATA,"is not a valid IATA code"))
41   }
42   sourceLon <- as.numeric(paste(airportsCanada$longitude)[
43     sourceFindIndex])
44   sourceLat <- as.numeric(paste(airportsCanada$latitude)[
45     sourceFindIndex])
46   sourceCoord <- c(sourceLon,sourceLat)
47   destLon <- as.numeric(paste(airportsCanada$longitude)[
48     destFindIndex])
49   destLat <- as.numeric(paste(airportsCanada$latitude)[
50     destFindIndex])
51   destCoord <- c(destLon,destLat)
52   airportDistList <- list()
53   airportDistList$source <- sourceIATA
54   airportDistList$dest <- destIATA
55   airportDistList$value <- round(distGeo(sourceCoord,destCoord)/
56     1000)
57   airportDistList$metric <- "Km"
58   airportDistList$xy_dist <- sqrt((sourceLon - destLon)**2 + (
59     sourceLat - destLat)**2)
60   airportDistList$sourceIndex <- sourceFindIndex

```

```

54   airportDistList$destIndex <- destFindIndex
55   airportDistList
56 }
57 airportsDist("AAA", "YQB")
58 airportsDist("YUL", "AAA")
59 airportsDist("YPA", "YQB")
60 airportsDist("YUL", "YQB")
61 airportsDist("YUL", "YQB")$value
62
63 #' Function to establish the estimated time of arrival
64 #'
65 #' @param sourceIATA The IATA of the departure airport.
66 #' @param destIATA The IATA of the arrival airport.
67 #' @return A list of the arrival time at the destIATA airport, and
68 #'         the information relative to it.
69 #' @examples
70 #' arrivalTime("YUL", "YQB")
71 #' arrivalTime("YUL", "YVR")
72 #'
73 # install.packages("lubridate")
74 library(lubridate)
75 arrivalTime <- function(sourceIATA, destIATA)
76 {
77   topSpeed <- 850
78   adjustFactor <- list()
79   adjustFactor$a <- 0.0001007194 # found by regression (not
80     included)
81   adjustFactor$b <- 0.4273381 # found by regression (not included)
82   arrivalTimeList <- list()
83   arrivalTimeList$source <- sourceIATA
84   arrivalTimeList$dest <- destIATA
85   arrivalTimeList$departureTime <- Sys.time()
86   distance <- airportsDist(sourceIATA, destIATA)
87   cruiseSpeed <- (distance$value * adjustFactor$a + adjustFactor$b) *
88     topSpeed
89   arrivalTimeList$avgCruiseSpeed <- cruiseSpeed
90   arrivalTimeList$flightTime <- ms(round(distance$value / cruiseSpeed
91     * 60, digits = 1))
92   arrivalTimeList$departureTZ <- paste(airportsCanada[distance$
93     sourceIndex, "tzFormat"])
94   arrivalTimeList$arrivalTZ <- paste(airportsCanada[distance$
95     destIndex, "tzFormat"])
96   arrivalTimeList$value <- with_tz(arrivalTimeList$departureTime +
97     arrivalTimeList$flightTime,
98     tzone = arrivalTimeList$
99       arrivalTZ)
100   arrivalTimeList
101 }
102 arrivalTime("AAA", "YYZ")
103 arrivalTime("YUL", "AAA")
104 arrivalTime("YUL", "YYZ")
105 arrivalTime("YUL", "YVR")
106 arrivalTime("YUL", "YYZ")$value
107 difftime(arrivalTime("YUL", "YVR")$value, Sys.time())
108 difftime(arrivalTime("YUL", "YYZ")$value, Sys.time())

```

```

103
104 # Import tax rates by province directly from the web
105 #install.packages("XML")
106 #install.packages("RCurl")
107 #install.packages("rlist")
108 library(XML)
109 library(RCurl)
110 library(rlist)
111 theurl <- getURL("http://www.calculconversion.com/sales-tax-
    calculator-hst-gst.html",.opts = list(ssl.verifypeer = FALSE))
112 tables <- readHTMLTable(theurl)
113 provinceName <- as.character(sort(unique(airportsCanada$province)))
114 taxRates <- as.data.frame(cbind(provinceName,as.numeric(sub("%","",
    tables$'NULL'[-13,5]))/100+1))
115 colnames(taxRates) <- c("province","taxRate")
116 taxRates
117
118 #' Shipping cost calculation function
119 #'
120 #' @param sourceIATA The IATA of the departure airport.
121 #' @param destIATA The IATA of the arrival airport.
122 #' @param weight The weight of the shipping.
123 #' @param percentCredit A double with a default value of 0.
124 #' @param dollarCredit A double with a default value of 0.
125 #' @return A list of the information for a shipping between the
    sourceIATA airport to the destIATA airport.
126 #' @examples
127 #' shippingCost("YUL","YQB")
128 #' shippingCost("YUL","YVR")
129 #'
130
131 shippingCost <- function(sourceIATA, destIATA, weight,
132     percentCredit = 0, dollarCredit = 0)
133 {
134
135     # Verification of the existence of the route between sourceIATA
    and destIATA
136     routeConcat <- as.character(paste(routesCanada$sourceAirport,
    routesCanada$destinationAirport))
137     if(is.na(match(paste(sourceIATA,destIATA),routeConcat)))
138     {
139         stop(paste("the combination of sourceIATA and destIATA (",
    sourceIATA,"-",destIATA,") do not corresponds to existing
    route"))
140     }
141
142     if(weight < 0 || weight > 30)
143     {
144         stop("The weight must be between 0 and 30 Kg")
145     }
146
147     if(percentCredit < 0 || percentCredit > 1)
148     {
149         stop("The percentage of credit must be between 0 % and 100 %")
150     }
151
152     if(dollarCredit < 0)

```



```

153 {
154   stop("The dollar credit must be superior to 0 $")
155 }
156
157 minimalDist = 100
158 distance <- airportsDist(sourceIATA, destIATA)
159 if (distance$value < minimalDist)
160 {
161   # We verify if the distance of shipping is further than the
162     minimal requirement
163   stop(paste("The shipping distance is under the minimal
164     requirement of", minDist, "Km"))
165 }
166
167 # Pricing variables
168 distanceFactor <- 0.03
169 weightFactor <- 0.8
170 fixedCost <- 3.75
171 profitMargin <- 1.12
172
173 # Traffic Index
174 trafficIndexSource <- as.numeric(paste(airportsCanada[distance$
175   sourceIndex, "combinedIndex"]))
176 trafficIndexDest <- as.numeric(paste(airportsCanada[distance$
177   destIndex, "combinedIndex"]))
178
179 # Calculation of the base cost
180 baseCost <- fixedCost + (distance$value*distanceFactor + weight*
181   weightFactor)/(trafficIndexSource*trafficIndexDest)
182
183 # Additional automated credits
184 automatedCredit <- 1
185 # Lightweight
186 automatedCredit <- automatedCredit * ifelse(weight < 2, 0.5, 1)
187 # Gold Member
188 automatedCredit <- automatedCredit * ifelse(baseCost > 100, 0.9,
189   1)
190 # Partnership
191 automatedCredit <- automatedCredit * switch(sourceIATA,
192   "YUL" = 0.85,
193   "YHU" = 0.95,
194   "YMX" = 0.95,
195   "YYZ" = 0.9,
196   "YKZ" = 0.975,
197   "YTZ" = 0.975,
198   "YZD" = 0.975)
199
200 # The Migrator
201 if(distance$value > 3000)
202 {
203   automatedCredit <- automatedCredit * 0.9
204 }
205 else if(distance$value <= 3000 & distance$value > 2500)
206 {
207   automatedCredit <- automatedCredit * 0.8775
208 }
209 else if(distance$value <= 2500 & distance$value > 2000)
210 {

```

```

204     automatedCredit <- automatedCredit * 0.85
205   }
206
207   # Calculation of tax rate and control of text
208   taxRate <- as.numeric(paste(taxRates[match(airportsCanada[
209     distance$sourceIndex, "province"], taxRates$province), "taxRate"
210     ]))
211   price <- round(pmax(fixedCost*profitMargin*automatedCredit*
212     taxRate, (baseCost*automatedCredit*profitMargin - dollarCredit
213     )*(1 - percentCredit)*taxRate), 2)
214
215   # Return List
216   shippingCostList <- list()
217   shippingCostList$distance <- distance
218   shippingCostList$weight <- weight
219   shippingCostList$distanceFactor <- distanceFactor
220   shippingCostList$weightFactor <- weightFactor
221   shippingCostList$fixedCost <- fixedCost
222   shippingCostList$profitMargin <- profitMargin
223   shippingCostList$percentCredit <- percentCredit
224   shippingCostList$dollarCredit <- dollarCredit
225   shippingCostList$minimalDist <- minimalDist
226   shippingCostList$trafficIndex <- list(trafficIndexSource,
227     trafficIndexDest)
228   shippingCostList$baseCost <- baseCost
229   shippingCostList$automatedCredit <- 1-automatedCredit
230   shippingCostList$taxRate <- taxRate
231   shippingCostList$price <- price
232   shippingCostList
233 }
234 shippingCost("YUL", "YVR", 1)
235 shippingCost("YUL", "YQB", 1)
236 shippingCost("YUL", "YVR", 30)
237 shippingCost("YUL", "YQB", 30)

```

Code Source A.4 – CaseStudyDevQ3.R

```

1  ### RStudio: -*- coding: utf-8 -*-
2  ##
3  ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
4  ## Goulet
5  ##
6  ## This file is part of the project
7  ## "Introduction a R – Atelier du colloque R a Quebec 2017"
8  ## http://github.com/vigou3/raquebec-atelier-introduction-r
9  ##
10 ## The creation is made available according to the license
11 ## Attribution-Sharing in the same conditions 4.0
12 ## of Creative Commons International
13 ## http://creativecommons.org/licenses/by-sa/4.0/
14 ##
15 ##### Question 3 #####
16 # We visualize the impact of a changes in the weight from a
17 # starting at the YUL airport.
18 curve(shippingCost("YUL", "YQB", x)$price, 0.01, 50, ylim=c(0, 200),

```

```

18     main="Shipping Price Variation with Weight",xlab="weight (Kg)"
19     ,
19     ylab="price (CND $)",lwd = 2)
20 curve(shippingCost("YUL","YVR",x)$price,0.01,50,xlab="weight (Kg)",
21        ylab="price (CND $)",add=TRUE, col = "red", lwd = 2)
22 curve(shippingCost("YUL","YYZ",x)$price,0.01,50,xlab="weight (Kg)",
23        ylab="price (CND $)",add=TRUE, col = "blue", lwd = 2)
24 curve(shippingCost("YUL","YYC",x)$price,0.01,50,xlab="weight (Kg)",
25        ylab="price (CND $)",add=TRUE, col = "purple", lwd = 2)
26 text(x=c(25,25,25,25),y=c(50,90,140,175),c("YUL-YYZ","YUL-YQB","YUL
    -YVR","YUL-YYC"),adj = 0.5,
27       cex = 0.75,font = 2,col = c("blue","black","red","purple"))

```

Code Source A.5 – CaseStudyDevQ4.R

```

1  #### RStudio: -*- coding: utf-8 -*-
2  ##
3  ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
    Goulet
4  ##
5  ## This file is part of the project
6  ## "Introduction a R – Atelier du colloque R a Quebec 2017"
7  ## http://github.com/vigou3/raquebec-atelier-introduction-r
8  ##
9  ## The creation is made available according to the license
10 ## Attribution-Sharing in the same conditions 4.0
11 ## of Creative Commons International
12 ## http://creativecommons.org/licenses/by-sa/4.0/
13
14 ##### Question 4 #####
15 # Import data of the competition
16 compData <- read.csv(paste(path,"/Reference/benchmark.csv",sep=""))
17 View(compData)
18 colnames(compData) <- c("weight","distance","price")
19 summary(compData)
20
21 # Weight visualisation
22 hist(compData$weight, freq = TRUE, main = "Repartition according to
    the weight",
23       xlab = "weight (Kg)", col = "cadetblue",breaks = 50)
24 weightCDF <- ecdf(compData$weight)
25 curve(weightCDF(x),0,15,ylim = c(0,1),lwd = 2,
26        xlab = "weight (Kg)",
27        ylab = "Cumulative Distribution Function")
28
29 # Distance visualisation
30 hist(compData$distance, freq = TRUE, main = "Repartition according
    to the distance",
31       xlab = "distance (Km)", col = "cadetblue",breaks = 50)
32 distanceCDF <- ecdf(compData$distance)
33 curve(distanceCDF(x),0,2500,ylim = c(0,1),lwd = 2,
34        xlab = "distance (Km)",
35        ylab = "Cumulative Distribution Function")
36
37 # Price according to weight
38 plot(compData$weight,compData$price,main = "Price according to the
    weight",

```

```

39     xlab = "weight (Kg)", ylab = "Price (CAD $)")
40
41 # Price according to distance
42 plot(compData$distance, compData$price, main = "Price according to
    the distance",
43       xlab = "distance (Km)", ylab = "Price (CAD $)")
44
45 # Price according to weight and distance
46 # install.packages("rgl")
47 library(rgl)
48 plot3d(compData$weight, compData$distance, compData$price)
49
50 # Chi's Square Test of Independency between the two variables
51 weightsBinded <- as.numeric(cut(compData$weight, 25))
52 distancesBinded <- as.numeric(cut(compData$distance, 25))
53 contingencyTable <- table(weightsBinded, distancesBinded)
54 chisq.test(contingencyTable)
55 contingencyTable <- rbind(contingencyTable[1:4,], colSums(
    contingencyTable[5:18,]))
56 (contingencyTable <- cbind(contingencyTable[, 1:14], rowSums(
    contingencyTable[, 15:24])))
57 independencyTest <- chisq.test(contingencyTable)
58 head(independencyTest$expected)
59 head(independencyTest$observed)
60 head(independencyTest$stdres)
61 independencyTest
62 cor.test(compData$weight, compData$distance, method = "pearson")
63
64 # Linear model
65 # we assume the same profit margin to simplify the situation
66 # We keep an intercept since we have a fixed cost
67 profitMargin <- 1.12
68 avgTaxRate <- sum(table(airportsCanada$province)*as.numeric(paste(
    taxRates$taxRate)))/length(airportsCanada$province)
69 compModel <- lm(price/(profitMargin*avgTaxRate) ~ distance + weight
    , compData)
70 summary(compModel)
71
72 # We plot the model
73 par(mfrow=c(2,2))
74 plot(compModel)

```

Code Source A.6 – CaseStudyDevQ5.R

```

1 ##### RStudio: -*- coding: utf-8 -*-
2 ###
3 ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
    Goulet
4 ##
5 ## This file is part of the project
6 ## "Introduction a R – Atelier du colloque R a Quebec 2017"
7 ## http://github.com/vigou3/raquebec-atelier-introduction-r
8 ##
9 ## The creation is made available according to the license
10 ## Attribution-Sharing in the same conditions 4.0
11 ## of Creative Commons International

```

```

12 ## http://creativecommons.org/licenses/by-sa/4.0/
13
14 ##### Question 5 #####
15 # install.packages("actuar")
16 library("actuar")
17
18 distName <- c("Normal", "Gamma", "LogNormal", "Weibull", "Pareto", "
  InvGaussian")
19 empCDF <- ecdf(compData$weight)
20 empPDF <- function(x, delta=0.01)
21 {
22   (empCDF(x+delta/2)-empCDF(x-delta/2))/delta
23 }
24
25 #' Generic function for statistical distribution adjustment
26 #'
27 #' @param data A vector of value over which we want to fit the
  distribution
28 #' @param dist The distribution name
29 #' @param ... The initial values to be given to the optimisation
  function
30 #' @return A list containing :
31 #' the optimized parameters,
32 #' the error value,
33 #' the deviance measure,
34 #' the convergence indicator and
35 #' the number of iterations necessited
36 #' @examples
37 #' x <- rnorm(1000,10,2)
38 #' distFit(x,"Normal", 1, 1)
39 #' x <- rgamma(1000,5,1)
40 #' distFit(x,"Gamma", 1, 1)
41 #'
42 distFit <- function(data, dist, ...)
43 {
44   dist = tolower(dist)
45   args = list(...)
46   if(dist == "normal")
47   {
48     law = "norm"
49     nbparam = 2
50   }
51   else if(dist == "exp")
52   {
53     law = "exp"
54     nbparam = 1
55     lower = 0
56     upper = 100/mean(data)
57   }
58   else if(dist == "gamma")
59   {
60     law = "gamma"
61     nbparam = 2
62   }
63   else if(dist == "lognormal")
64   {
65     law = "lnorm"

```

```

66     nbparam = 2
67   }
68   else if(dist == "weibull")
69   {
70     law = "weibull"
71     nbparam = 2
72   }
73   else if(dist == "pareto")
74   {
75     law = "pareto"
76     nbparam = 2
77   }
78   else if(dist == "invgaussian")
79   {
80     law = "invgauss"
81     nbparam = 2
82   }
83   else if(dist == "student")
84   {
85     law = "t"
86     nbparam = 1
87     lower = 0
88     upper = 100/mean(data)
89   }
90   else if(dist == "burr")
91   {
92     law = "burr"
93     nbparam = 3
94   }
95   else
96   {
97     message <- "The only distributions available are:
98     Normal, Exp, Gamma, LogNormal, Weibull, Pareto, Student, Burr
99     and InvGaussian.
100    (This case will be ignored)"
101    stop(message)
102  }
103  if(nbparam != length(args))
104  {
105    message <- paste("There is a mismatch between the number of
106    arguments passed to the
107    function and the number of arguments needed to
108    the distribution.",
109    "The",dist,"distribution is taking",nbparam,"
110    parameters and",
111    length(args),"parameters were given.")
112    stop(message)
113  }
114  # Treament
115  if(nbparam == 1)
116  {
117    param <- optim(par = args, function(par)
118      -sum(do.call(eval(parse(text = paste("d",law,sep=""))),c(list
119        (data),par,log = TRUE))),
120      method = "Brent", lower = lower, upper = upper)
121  }

```

```

118   else{
119     param <- optim(par = args, function(par)
120       -sum(do.call(eval(parse(text = paste("d",law,sep="")),c(list
121         (data),par,log = TRUE))))
122   }
123   # Deviance value
124   devValue <- sum((empPDF(x <- seq(0,max(data),0.1))-do.call(eval(
125     parse(text = paste("d",law,sep="")),c(list(x),param$par)))*
126     2)
127
128   # Return List
129   distFitList <- list()
130   distFitList$param <- param$par
131   distFitList$errorValue <- param$value
132   distFitList$devValue <- devValue
133   distFitList$convergence <- param$convergence
134   distFitList$nbiter <- param$counts[1]
135   distFitList
136 }
137
138 (resultDistFitting <- sapply(distName,function(x) unlist(distFit(
139   compData$weight,x,1,1))))
140
141 law <- c("norm","gamma","lnorm","weibull","pareto","invgauss")
142 col <- c("red", "yellow", "purple", "green", "cyan", "blue")
143 x <- seq(0,30,0.1)
144
145 # Visualization of the fitting distribution
146 par(mfrow = c(1,2),font = 2)
147 plot(function(x) empCDF(x), xlim = c(0,15), main = "", xlab = "
148   weight (Kg)", ylab = "CDF(x)")
149 invisible(sapply(1:length(law),function(i) curve(do.call(eval(parse
150   (text = paste("p",law[i],sep = "))),
151     c(list(x),
152       as.
153       vector
154       (
155         resultDistFitting
156         [c
157           (1:2),
158           i]))),
159     add = TRUE, lwd =
160       3, col = col[i
161         ])))
162
163 hist(compData$weight, xlim = c(0,15), main = "", xlab = "weight (Kg
164   )", breaks = 300,freq = FALSE)
165 invisible(sapply(1:length(law),function(i) curve(do.call(eval(parse
166   (text = paste("d",law[i],sep = "))),
167     c(list(x),
168       as.
169       vector
170       (
171         resultDistFitting
172         [c
173           (1:2),
174           i]))),

```

```

151                                     add = TRUE, lwd =
                                           3, col = col[i
                                           ])))
152 legend(x="right", y = "center",distName, inset = 0.1, col = col,
      pch = 20, pt.cex = 2, cex = 1,
153       ncol = 1, bty = "n", text.width = 2, title = "Distribution")
154 mtext("Ajustement sur distribution empirique", side = 3, line = -2,
      outer = TRUE)
155
156 # We thus choose the LogNormal distribution which possesses the
      smallest deviance and the best fit.
157 distChoice <- "LogNormal"
158 (paramAdjust <- resultDistFitting[c(1:2),match(distChoice,distName)
      ])
159
160 # It is also possible to do the equivalent with fitdistr of the
      library MASS,
161 # but with less option for the distribution.
162 library("MASS")
163 (fit.normal <- fitdistr(compData$weight,"normal"))
164 (fit.gamma <- fitdistr(compData$weight, "gamma"))
165 (fit.lognormal <- fitdistr(compData$weight, "lognormal"))
166 (fit.weibull <- fitdistr(compData$weight, "weibull"))

```

Code Source A.7 – CaseStudyDevQ6.R

```

1 #### RStudio: -*- coding: utf-8 -*-
2 ##
3 ## Copyright (C) 2017 David Beauchemin, Samuel Cabral Cruz, Vincent
      Goulet
4 ##
5 ## This file is part of the project
6 ## "Introduction a R – Atelier du colloque R a Quebec 2017"
7 ## http://github.com/vigou3/raquebec-atelier-introduction-r
8 ##
9 ## The creation is made available according to the license
10 ## Attribution-Sharing in the same conditions 4.0
11 ## of Creative Commons International
12 ## http://creativecommons.org/licenses/by-sa/4.0/
13
14 ##### Question 6 #####
15 theurl <- getURL(paste("file://",path,"/Statement/Markdown/
      CaseStudyStatement.html",sep=""),.opts =
16                  list(ssl.verifypeer = FALSE))
17 tables <- readHTMLTable(theurl)
18 lambdaTable <- as.data.frame(tables$"NULL")
19 colnames(lambdaTable) <- c("Month","Avg3yrs")
20 lambdaTable
21
22 # The possible routes are filtered from the starting point 'YUL'
23 # and a distribution is created according to the destination index.
24 simAirportsDests <- as.character(paste(routesCanada[routesCanada$
      sourceAirport == "YUL",
25                                     "
                                           destinationAirport
                                           "]))

```



```

26 simCombinedIndex <- combinedIndex[names(combinedIndex) %in%
    simAirportsDests]
27 airportsDensity <- simCombinedIndex/sum(simCombinedIndex)
28
29 # Function for the simulation of the shipment prices.
30 simulShipmentPrice <- function(Arrival, Weight)
31 {
32   ownPrice <- ifelse(is(testSim <- try(shippingCost("YUL", Arrival,
    Weight)$price, silent = TRUE),
33     "try-error"), NA, testSim)
34   distance <- airportsDist("YUL", Arrival)$value
35   nd <- as.data.frame(cbind(distance, Weight))
36   colnames(nd) <- c("distance", "weight")
37   compPrice <- predict(compModel, newdata = nd)
38   customerChoice <- ifelse(is.na(ownPrice), 0, ifelse(ownPrice <
    compPrice, 1, 0))
39   rbind(Arrival, distance, Weight, ownPrice, compPrice, customerChoice)
40 }
41
42 # Function for the simulation of the shipment parameters, weights
    and destinations.
43 simulShipment <- function(simNbShipments)
44 {
45   # Weights are then generated for each of the packages.
46   simWeights <- eval(parse(text = paste("r", law[match(distChoice,
    distName)], sep = " "))(simNbShipments,
47
48   # We finally generate a destination for each package (the
    departure will always be from 'YUL').
49   simArrivals <- sample(size = simNbShipments, names(airportsDensity)
    ), prob = airportsDensity,
50     replace = TRUE)
51   sapply(seq(1, simNbShipments), function(x) simulShipmentPrice(
    simArrivals[x], simWeights[x]))
52 }
53
54 # Function for overall simulation
55 simulOverall <- function()
56 {
57   # We generate n observations of the Poisson distribution with
    param = sum(lambda).
58   # We know from probability notion that the sum of independent
    Poisson distribution follows
59   # a Poisson distribution with param = sum(lambda).
60   simNbShipments <- rpois(1, lambda = sum(as.numeric(paste(
    lambdaTable$Avg3yrs))))
61   # We simulate each shipment
62   simulShipment(simNbShipments)
63 }
64
65 nsim <- 1
66 simulResults <- replicate(nsim, simulOverall(), simplify = FALSE)
67 (marketShareSales <- sapply(1:nsim, function(x) sum(as.numeric(

```

```

paramA
[1
pa
[2

```

```

      simlResults[[x]][6,]))/length(simlResults[[x]][6,]))
68 (ownRevenus <- sum(as.numeric(simlResults[[1]][4,])*as.numeric(
      simlResults[[1]][6,]),na.rm = TRUE))
69 (compRevenus <- sum(as.numeric(simlResults[[1]][5,])*(1-as.numeric(
      (simlResults[[1]][6,]),na.rm = TRUE))
70 (marketShareRevenus <- ownRevenus/(ownRevenus+compRevenus))
71
72 arrivalSales <- as.character(simlResults[[1]][1, simlResults
      [[1]][6,]==1])
73 distanceSales <- as.numeric(simlResults[[1]][2, simlResults
      [[1]][6,]==1])
74 weightSales <- as.numeric(simlResults[[1]][3, simlResults
      [[1]][6,]==1])
75
76 arrivalComp <- as.character(simlResults[[1]][1, simlResults
      [[1]][6,]==0])
77 distanceComp <- as.numeric(simlResults[[1]][2, simlResults
      [[1]][6,]==0])
78 weightComp <- as.numeric(simlResults[[1]][3, simlResults
      [[1]][6,]==0])
79
80 # Representation of the result
81 table(arrivalSales)
82 mean(distanceSales)
83 table(arrivalComp)
84 mean(distanceComp)
85 par(mfrow = c(1,1))
86 hist(weightSales, freq = FALSE, breaks = 100, xlim = c(0,15), main =
87       "Sales vs Theoretical Weights Distribution", xlab = "
      weight (Kg)")
88 curve(do.call(eval(parse(text = paste("d",law[match(distChoice,
      distName)] , sep = " "))),
89       c(list(x),as.vector(paramAdjust))),add = TRUE, lwd =
      2)
90 abline(v = v <- exp(paramAdjust[1]+paramAdjust[2]**2/2), lwd = 2)
91 text(v+0.75,0.3,as.character(round(v,2)))
92 abline(v = v <- mean(weightSales), col = "red", lwd = 2)
93 text(v - 0.75,0.3,round(v,2), col = "red")
94
95 sample(airportsCanada$altitude, size = 10, replace = TRUE)
96 probs = airportsCanada$altitude/sum(airportsCanada$altitude)
97 sample(airportsCanada$altitude, size = 10, replace = TRUE, prob =
      probs)
98 sample(unique(as.character(paste(airportsCanada$name))), size = 10,
      replace = FALSE)

```
