

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	21
2.2 Création de fonctions utilitaires	27
2.3 Conception de graphiques en R	34
2.4 Outils d'analyse statistique en R	40
2.5 Ajustement de distributions statistiques sur données empiriques	47
2.6 Analyse par simulation en R	56
Conclusion	60
A Code source du projet	64
B Contribution au projet <i>OpenFlights</i>	85
C Premiers pas avec R	90
C.1 Installation de R	90
C.1.1 <i>Windows</i>	90
C.1.2 <i>(MAC) OS X</i>	90
C.1.3 <i>Linux or Unix-Like</i>	91
C.2 Installation de R Studio	91
C.3 Installation de paquetages	91
C.4 Mise à jour de R et ses paquetages	92
C.5 Installation de Git	92

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	17
2.3	Exemple de carte géographique produite avec <code>ggmap</code>	23
2.4	Exemple de carte géographique produite avec <code>leaflet</code>	24
2.5	Densité de la population canadienne	27
2.6	Passage de paramètres graphiques à la commande <code>plot</code>	36
2.7	Tracer une courbe avec la commande <code>plot</code>	37
2.8	Tracer une courbe avec la commande <code>curve</code>	38
2.9	Distribution des altitudes des aéroports canadiens	39
2.10	Représentation graphique de la fonction <code>f3</code>	50
2.11	Visualisation de l'ajustement sur la distribution empirique des poids (Cumulative à gauche et Densité à droite)	55
2.12	Comparaison des résultats de simulation obtenus avec 6 réplicats	58
2.13	Classement <i>RedMonk</i> des différents langages de programmation	60

Liste des codes sources

1.1	Environnement de travail	10
2.1	Extraction des données	13
2.2	Filtrer les données	15
2.3	Traitement standard de données géospatiales en R	18
2.4	Exemple de requête SQL	19
2.5	Fonctionnalités avancées de SQL	20
2.6	Fonctionnalités avancées de SQL	20
2.7	Fonctions de visualisation de données	22
2.8	Générer une carte du trafic aérien avec <code>ggmap</code>	22
2.9	Générer une carte du trafic aérien avec <code>leaflet</code>	23
2.10	Structure pour la définition d'une fonction	27
2.11	L'instruction <code>return</code> et le retour standard d'une fonction R	28
2.12	Définir des valeurs par défauts dans les fonctions utilitaires	28
2.13	Passage d'arguments à une fonction	30
2.14	L'assignation et les valeurs par défaut	31
2.15	Retour multiple par l'entremise d'une liste	31
2.16	Gestion des erreurs sous R	32
2.17	Utilisation de la commande <code>plot</code>	35
2.18	Utilisation de la commande <code>curve</code>	35
2.19	<code>hist</code> , <code>density</code> , <code>lines</code> , <code>abline</code> , <code>legend</code> et <code>mtext</code>	39
2.20	Fonctions relatives à la distribution Normale	41
2.21	Génération de nombres aléatoires	42
2.22	Fonctions de densité et de répartition empiriques	43
2.23	Tests d'indépendance et de corrélation entre distributions	44
2.24	Régression linéaire sur données empiriques	45
2.25	Optimisation générique avec R	48
2.26	Maximisation d'une fonction avec <code>optim</code>	49
2.27	Ajustement de distribution sur données empiriques	50
2.28	Réplicat maison de la fonction <code>fitdistr</code>	51
2.29	Exemple d'utilisation de la fonction <code>distFit</code>	54
2.30	Pige aléatoire sur support vectoriel	56
2.31	Replication d'une analyse par simulation	57
A.1	<code>benchmark.R</code>	64
A.2	<code>caseStudy1.R</code>	65
A.3	<code>caseStudy2.R</code>	70
A.4	<code>caseStudy3.R</code>	76
A.5	<code>caseStudy4.R</code>	76
A.6	<code>caseStudy5.R</code>	78

A.7	caseStudy6.R	82
B.1	tzFormatRefill.R	86

Liste des tableaux

2.1	Comparaison entre les informations d' <i>OpenFlights</i> et les résultats des fonctions <code>airportsDist</code> ainsi que <code>arrivalTime</code>	33
2.2	Liste des distributions statistiques disponibles en R	41
2.3	Comparaison entre les coefficients réels et estimés par régression linéaire	46
2.4	Ajustement de distribution sur poids observés selon la fonction utilitaire <code>distFit</code> et la fonction <code>fitdistr</code> du paquetage <code>MASS</code>	55
2.5	Comparaison des paramètres de la distribution ajustée et théorique des poids de colis (Kg)	56
B.1	Étude comparative de concordance entre les différentes sources de fuseaux horaires	86

Préface

Dans le cadre du colloque "R à Québec" qui s'est tenu le 25 et 26 mai 2017 sur le campus de l'Université Laval, une séance d'introduction au langage de programmation R fut offerte aux participants. Cette séance visait principalement la maîtrise des rudiments de cet environnement de programmation tout en prenant conscience des capacités de ce langage. [17] Elle 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 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 son aboutissement ultime qui consiste à faire une analyse de compétitivité.

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. Il est aussi important de 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 ou avancé, mais toujours apportées de manière simplifiée et accessible à n'importe qui n'ayant 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'événement. Nous croyons sincèrement que R est un langage d'actualité qui se révélera 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. Nous fonderons notre analyse sur le jeu de données d'*OpenFlights*. [10]



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 [32]
routes.dat	Données relatives à tous les trajets possibles dans le monde [31]
airlines.dat	Données relatives aux compagnies aériennes [30]

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, nous serons amenés à analyser les prix de la concurrence pour extrapoler leur tarification. Nous pourrons ainsi s'assurer que la nouvelle tarification sera efficiente et profitable. Finalement, nous comparerons ces

deux tarifications pour déterminer la compétitivité de notre nouvelle structure de prix en procédant à une analyse par simulation.



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. [7] L'information disponible y est étonnamment très complète et facile d'approche. Ces caractéristiques rendent 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'ordre 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 des dossiers parents dans laquelle il se trouve. Pour ce faire, nous devons utiliser le principe de liens relatifs plutôt qu'absolus. En R, deux fonctions bien spécifiques nous fournissent les outils rendant cette tâche possible. Il s'agit de `getwd` et `setwd` [42]. 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 `src` [9] et que nous pourrions vouloir avoir accès à d'autres parties du répertoire au sein du code, le [Code Source 1.1](#) nous permettra de placer notre racine de projet à un niveau supérieur dans l'arborescence des dossiers 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.²

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 l'appellation "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 nombres pseudo-aléatoires (GNPA) ce qui aura comme impact de toujours produire les mêmes résultats

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

2. Il faut comprendre que les chemins relatifs n'enlèvent pas toutes les dépendances, mais seulement celles qui sont externes au dépôt du projet.

pour autant que le GNPA utilisé soit le même. L'instruction `set.seed` [50] dans le [Code Source 1.1](#) se chargera de fournir une valeur de départ aux calculs du GNPA.

Code Source 1.1 – Environnement de travail

```
1 ##### Setting working directory properly #####
2 # Recommended :
3 path <- paste(getwd(), "..", sep = "/")
4 # Alternatively :
5 # setwd('.')
6 # path <- getwd()
7
8 # Root Pseudo Random Number Generator (PRNG)
9 set.seed(31459)
```



Code source du projet

Les codes sources du projet dans son intégrité sont en annexe à ce document. N'hésitez pas à vous y référer au besoin.
Voir l'[Appendice A](#).

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 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` [37]. Bien que le nom de la fonction indique qu'elle permet de lire un fichier présenté dans un format *comma-separated values* `.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 autre symbole.[4] 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 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 considérer les titres comme étant des entrées en soi.¹ Bien qu'il serait possible de travailler avec des données sans en-tête, il s'agit ici d'une très mauvaise pratique. Pour remédier à la situation, nous assignerons donc des noms aux colonnes grâce à la méthode `colnames` de la classe `data.frame` en lui passant un vecteur contenant les noms convoités.

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



Pourquoi ne pas avoir choisi la pillule rouge ? !

Vous vous demandez probablement pourquoi R utilise un `data.frame` plutôt qu'un `array` ou une `matrix` pour contenir les données. Tout d'abord, le conteneur le plus simple de R est le `vector`. De cette classe, nous aurons ensuite le `array` étant en fait une spécialisation de `vector` qui possèdera un attribut `dim`. La classe `matrix` sera à son tour une spécialisation de la classe `array` qui ne pourra avoir que deux dimensions (lignes et colonnes). Comme nous pouvons le voir tous ces conteneurs possèdent des caractéristiques intéressantes d'un point de vue mathématique, mais ils souffriront tous de la même lacune ; Ils ne peuvent contenir que des éléments possédant le même mode. La classe `data.frame` contournera ce problème en héritant plutôt de la classe `list`. Elle conservera tout de même les propriétés mathématiques des 3 autres conteneurs précédents en utilisant la technique de la composition. En d'autres mots, un `data.frame` n'est rien d'autre qu'une `list` de `vector`.

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

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

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

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

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 et l'affichage des variables ne sera pas impacté. Ceci s'explique par le fait que R créera des formats d'affichage qui associeront à chaque facteur une valeur unique correspondante. Le seul impact réel réside plutôt 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` [48] 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 par un facteur 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
- Alimenter les valeurs manquantes avec des sources de données externes (si possible) ou appliquer un traitement approximatif justifiable en documentant les impacts possibles sur le reste 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. L'accès à ces attributs se fera grâce à des fonctions définies à cet effet. Comme il sera possible de le voir plus loin, l'opérateur `$` servira aussi à l'accès aux attributs dans le cas particulier où l'objet manipulé sera de mode `list`. Vous vous demandez probablement : Comment savoir si nous sommes en présence d'un objet ? C'est simple, tout dans R est un objet !



`$` et le `data.frame`

Nous ne serons pas étonné d'apprendre que l'opérateur `$` nous permettra d'extraire des variables d'un `data.frame`. Comme nous l'avons précisé à la [sous-section 2.1.1](#), un `data.frame` n'est rien d'autre qu'une liste de vecteurs.

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 se retrouver en argument ou en valeur de retour d'une autre fonction. De plus, il est possible de définir des fonctions dites anonymes qui se révéleront très pratiques. À ce sujet, les personnes habituées au paradigme procédural présent dans les langages comme *SAS* ou *VBA* devront s'habituer à l'évaluation d'une expression de son point central vers l'extérieur au lieu du chemin traditionnel allant du début vers la fin.

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. [8] 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

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

nommés qui nous intéressent.

Avec ces outils en mains, il devient très facile de filtrer les aéroports canadiens à l'aide de la variable que nous avons nommée *country* du `data.frame` `airports`. Par un raisonnement connexe, la fonction `subset` [54] 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 de fonctions 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 cette tâche ? 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 fonctionnalités de base plus aisée et compréhensible... 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 que vous puissiez vous construire 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 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 consistait à simplement prendre le code *ICAO*, mais le code *IATA* nous semblait plus universel.³

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 la création de plusieurs *Application Programming Interface* (API) disponibles gratuitement sur le web. 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 qu'il soit 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

3. Il s'agit du code communément utiliser pour le transport des particuliers.

nous avons pris le fichier `.shp` qui définit les provinces et territoires du Canada. [24] En ce qui concerne les fuseaux horaires, nous avons initialement trouvé ceux-ci sur un site [29] 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 2016.⁴ En approfondissant nos recherches, nous sommes tombés sur un projet *GitHub* visant à créer un outil pour extraire l'information d'*Open Street Map* (OSM) pour construire une *ShapeFile* des fuseaux horaire mondiaux qui inclus les eaux territoriales. [25]⁵



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 du premier logiciel 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 vus comme un standard pour contenir ce type d'information.

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

Pour être en mesure de 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`.

4. Les fuseaux horaires n'ont pas tendance à changer souvent dans les pays industrialisés comme le Canada, ceci ne consistait donc pas en un enjeu majeur.

5. Les impacts de ce changement seront mineurs dans le cadre de notre étude se concentrant sur les aéroports canadiens. Vous réferez à l'Appendice B pour plus d'informations sur les raisons de ce changement.

- .prj** Contient les informations sur la projection associée (le modèle mathématique permettant d'interpréter les informations du **.shp** [14]).
- .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 paquetages **rgdal** [23] et **sp**[33]. 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* que nous cherchons à combiner. La nécessité que nos points soit sous la même projection que celle du *ShapeFile* vient 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 des points géographiques sur une couche donnée. Le retour de la fonction **over** sera finalement un **data.frame** de même dimension équivalente au nombre de points fournis en argument que nous pourrions facilement combiner avec 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 layer
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 pour se débarrasser de tout ce qui ne sera plus utile. Bien que les fonctionnalités de base de R vous permettent d'accomplir la tâche, nous profitons de cette étape du processus pour vous présenter les paquetages `sqldf` [27] et `dplyr` [55].

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 d'interpréter ces instructions qu'il y ait ou non des fonctionnalités permettant de répliquer leur comportement.⁶ [19] Nous conseillons fortement à tous les analystes de données de s'appropriier 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 du SQL, qui se rapproche énormément d'une phrase complète et structurée, rend sa compréhension presque immédiate, et ce, même à des personnes ignorant 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")
```

La requête ci-dessus pourrait être transformée de manière textuelle sous la forme suivante :

1. Sélectionne les variables `name`, `IATA`, `altitude` et `province`
 2. Dans la relation `airportsCanada`
 3. Dont la province est *New Brunswick*
 4. En triant le tout par *name*
-
6. Minimale, offrir un paquetage permettant leur interprétation.

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 feront parties de notre quotidien, nous retrouvons `*` qui lorsque placé dans l'instruction `select` permettra d'extraire l'ensemble des variables de la relation sans avoir à les écrire une à une. 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 mère⁷ même s'il n'y a 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.

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.^{8 9}

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)")
```

7. La relation se situant à la gauche dans le merge.

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

9. L'utilisation de la case dans les exemples ne sert qu'à 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 possibilités 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 `sqldf` 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 pour prendre en compte le fonctionnement intrinsèque de R, soit le calcul vectoriel. 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 temps d'exécution irraisonnables. [16]



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.

<code>View</code>	Permet d'ouvrir un <code>data.frame</code> 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. [44]
<code>head</code>	Renvoie en console un extrait des premières observations d'une relation (par défaut, 10 observations sont renvoyées). [52]
<code>summary</code>	Compilation de statistiques pertinentes au sujet des différentes variables contenues dans une table. 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. [36]

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 peu conviviale et pratique. De plus, il sera facile de se perdre dans le contenu présenté plus la relation possèdera de variables. Pour contrer ces problèmes, 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 combiner la visualisation de données géospatiales sur support statique disponible en ligne, tels que ceux de *Google Maps*, avec la puissance du paquetage **ggplot2**. [28]

En jetant un coup d’œil au [Code Source 2.8](#), nous voyons qu’il est possible de produire des cartes très rapidement avec seulement quelques lignes de code. Malgré la facilité d’utilisation de **ggmap**, nous ressentons vite 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))
8 size <- as.numeric(paste(TrafficData$combinedIndex))
```

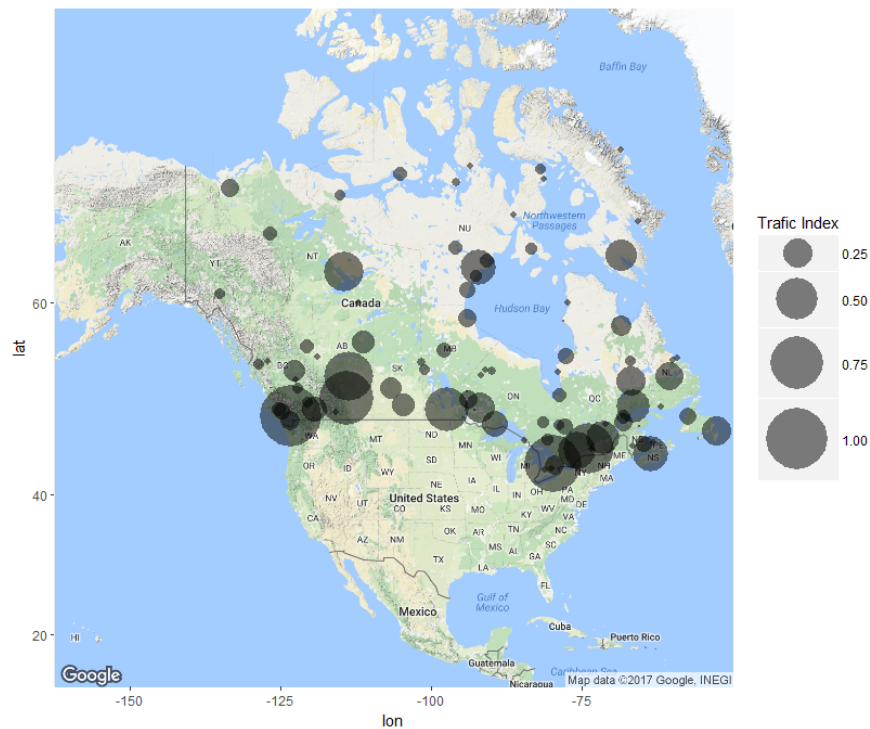


FIGURE 2.3 – Exemple de carte géographique produite avec `ggmap`

```

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

```

Pour ce faire, le paquetage `leaflet` [11] viendra à notre secours avec un faible coût en complexité compte tenu de la flexibilité impressionnante rajoutée. Cette paquetage n'est rien d'autre qu'une enveloppe permettant de faire appel à la librairie *JavaScript*. [3] Le [Code Source 2.9](#) est à l'origine de la [Figure 2.4](#) provenant en fait d'une carte interactive.

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/"

```

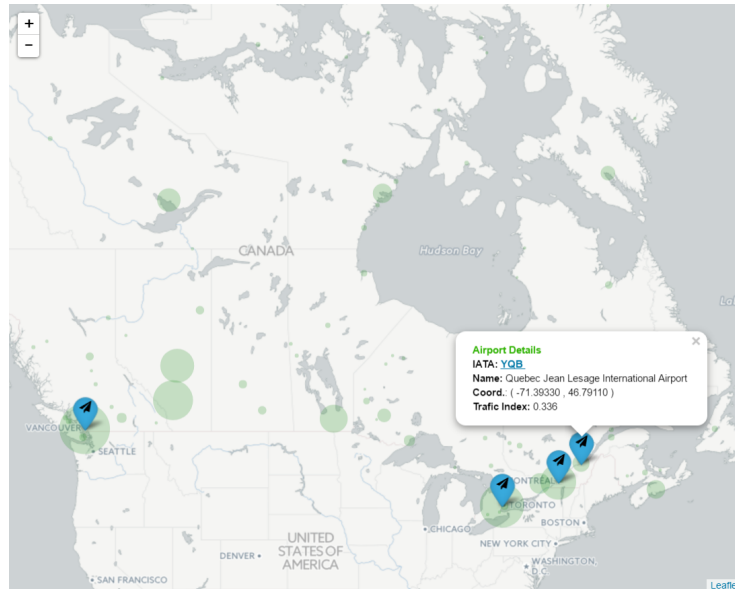



FIGURE 2.4 – Exemple de carte géographique produite avec leaflet

```

gpx"
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',
  iconColor = 'black',
  library = 'fa')
21
22

```

```

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. 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* [6] qui est célèbre auprès des utilisateurs \LaTeX 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 stocker beaucoup d'information au sein du même objet sans alourdir indûment 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 assigné. 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. [11]



Est-ce tout ce que peut accomplir leaflet ?

Bien entendu, les exemples présentés dans ce document font l'éloge que de deux applications grossières de ces deux paquets. Vous serez en mesure de 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 :

<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 appréciable (Voir [Figure 2.5](#)).

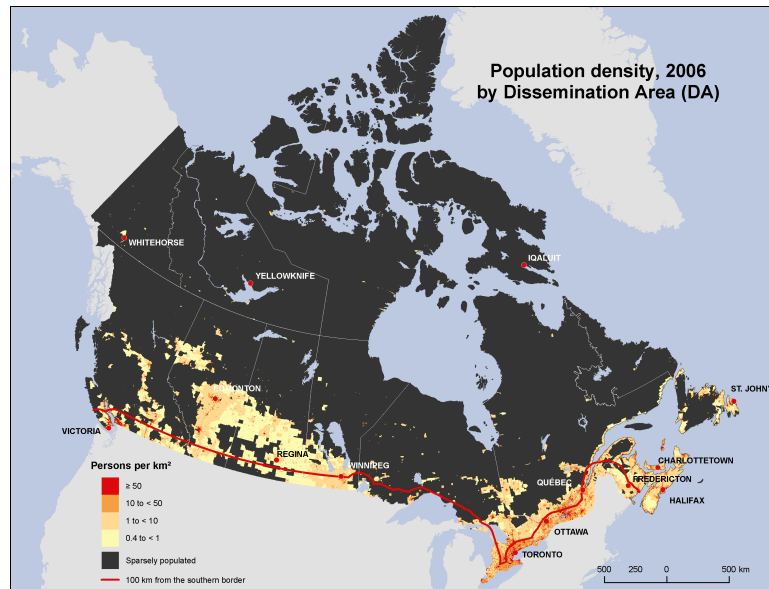


FIGURE 2.5 – Densité de la population canadienne

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 code. Dans le cadre du projet, nous avons pris l'initiative de construire les trois fonctions suivantes :

- airportsDist** Calcule la distance en Km entre deux aéroports
- arrivalTime** Calcule l'heure d'arrivée d'un colis posté au moment du calcul
- shippingCost** Calcule 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 à éviter l'utilisation du `return` ou d'autres modificateurs de flux du même genre. Nous préférons plutôt n'avoir 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 pour toutes les autres fonctions que nous avons déjà utilisées dans la section précédente. 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 les fonctions `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 puisse sembler anodine pour un nouveau programmeur, vous apprendrez rapidement 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 lequel un utilisateur donné ne pourrait avoir une influence sans directement modifier le code source. Malgré le fait que vous soyez convaincus que votre programme ne sera jamais utile dans un autre dessin que celui qui vous a initialement amené à le créer, ce genre de pratique, en plus d'être inefficace, 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 retirer que du positif d'adopter de 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 bénéficierez de votre rigueur même si cette dernière vous aura fait perdre du temps précieux au cours 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 pour le 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** [5] 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** [18]. 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 utilisons 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`, `.tex` (L^AT_EX), 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, L^AT_EX 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 seront tentés de créer un paquetage et de le publier sur *Comprehensive R Archive Network (CRAN)*.
<https://cran.r-project.org/doc/manuals/R-exprs.html#Marking-text>

En reprenant les fonctions `fctest1`, `fctest2` et `fctest3`, 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 > fctest1(3)
2 Error in fctest1(3) : argument "b" is missing, with no default
3 > fctest2(3)
4 [1] 3
5 > fctest2(b=5)
6 Error in fctest2(b = 5) : argument "a" is missing, with no default
7 > fctest3(3)
8 [1] 6
9 > fctest3(3,5)
10 [1] 8
11 > fctest3(b=5)
12 [1] 7
13 > fctest3(b=5,3)
14 [1] 8
15 > fctest3(3,5,4)
16 Error in fctest3(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 `fctest2(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'ait 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 `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 `ftest2` qu'une valeur à `b`.

L'appel `ftest3(b=5,3)` expose la flexibilité tout aussi incroyable que dangereuse des procédés d'assignation de valeurs lors des appels de fonction en R. Cette flexibilité de pouvoir alterner 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 et les valeurs restantes seront attribuées de manière positionnelle sur les arguments n'ayant toujours pas reçu de valeur. Il faut toutefois faire attention puisqu'aucune discrimination ne sera effectuée par rapport aux paramètres ayant des valeurs par défaut ([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 distincts. 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 pour ensuite 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 à aborder au sujet des fonctions est 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 en présence de ces cas particuliers. Dans le cas contraire, 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 composé 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.

```

IATA		Distance (km)		Temps (hh :mm)	
Source	dest.	<i>OpenFlights</i>	airportsDist	<i>OpenFlights</i>	arrivalTime
YUL	YQB	232	233	0 :47	0 :36
YUL	YVR	3679	3693	5 :04	5 :26
YUL	YYZ	505	508	1 :07	1 :14

TABLE 2.1 – Comparaison entre les informations d’*OpenFlights* et les résultats des fonctions `airportsDist` ainsi que `arrivalTime`



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>

À des fins de validation, nous avons comparé les valeurs retournées par nos fonctions avec les valeurs disponibles directement sur l’outil d’*OpenFlights*. Les résultats sont tout de même très satisfaisant compte tenu de la méthode utilisée. (Tableau 2.1)

Les différences entre les temps de vol d’*OpenFlights* et le retour de la fonction `arrivalTime` sont facilement explicables. Si nous examinons le code source de cette fonction, nous remarquons la définition d’une vitesse moyenne de croisière. Cette dernière est déterminée par interpolation linéaire à partir de la vitesse de croisière optimale utilisée pour les vols commerciaux. [12] Cela a pour effet de nous faire sous-estimer le temps pour des vols de courte durée tout en sur-estimant celle des vols de longue durée. En effet, plus la proportion du temps de vol destinée à faire décoller et ratterir un avion est important, moins sa vitesse de croisière moyenne s’approchera de notre estimation. En ce qui concerne les vols de moyenne durée, nos calculs sont très représentatifs de la réalité.

La divergence sur la distance entre YUL et YVR est toutefois un peu plus délicate à justifier puisque la vraie valeur devrait plutôt se situer quelque part entre les deux distances présentées. [2] En considérant la valeur mentionnée sur le site de *Air Miles Calculator*, nous voyons que notre divergence est beaucoup plus acceptable et cohérente avec la légère sur-estimation que nous retrouvons pour les deux autres destinations.

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 survoler les 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` [41] est de loin la fonction graphique la plus générique en R. Cette fonction ne possède que trois arguments : `x`, `y` et `...`. Naturellement, nous devrons 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 nuage de points. 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 de son utilisation la plus commune, soit 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 dans le 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` [53] (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 et explicite 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 précisions.

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 semble justifiée, elle pourra jouer de mauvais tours aux utilisateurs non-avertis. Comme le montre la [Figure 2.7](#), dépendamment de l'espacement entre les points calculés, nous pouvons 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` [38] 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. Nous pourrions aussi spécifier 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

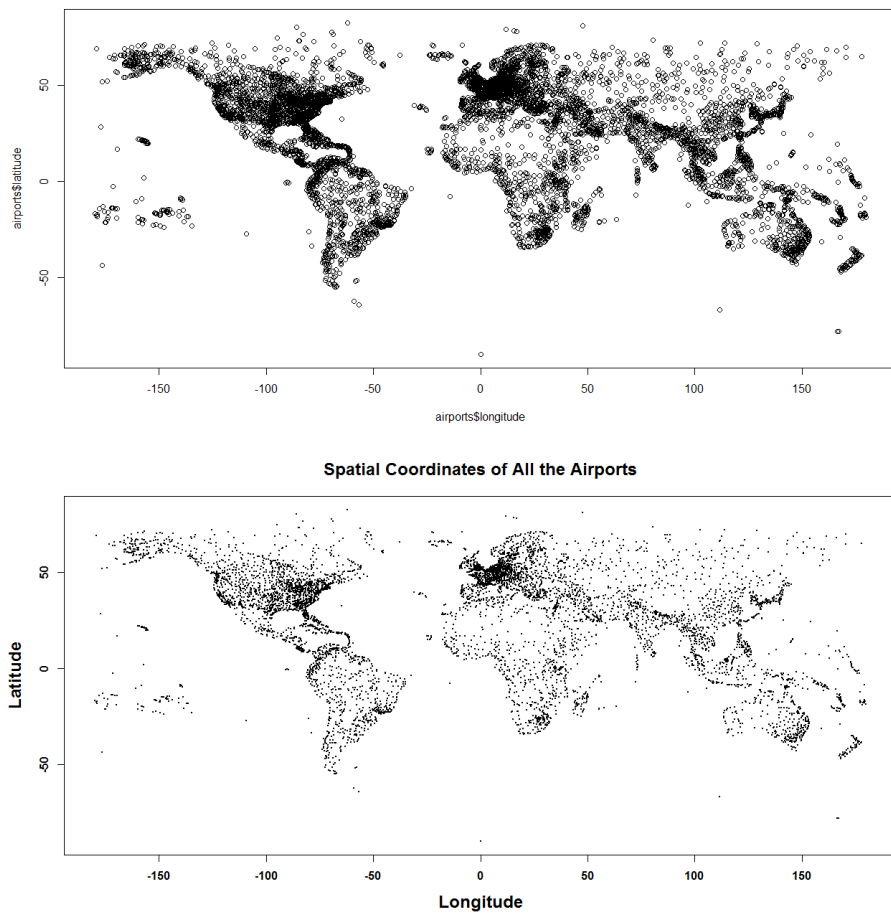


FIGURE 2.6 – Passage de paramètres graphiques à la commande `plot`

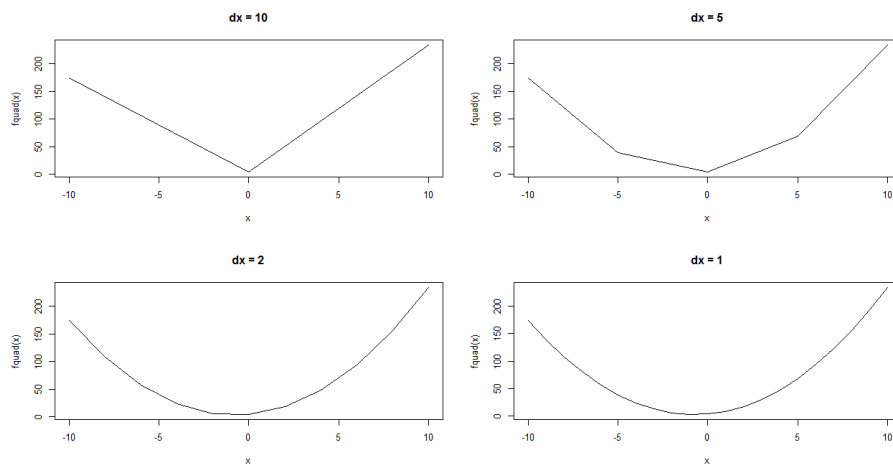


FIGURE 2.7 – Tracer une courbe avec la commande `plot`

commande `hist` [43] est de loin le plus important puisqu'il permettra d'obtenir un visuel 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 de largeur 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). Autrement, ce sera la probabilité empirique d'observer une valeur dans chacun des regroupements qui sera exhibée.



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` [45] est aussi très intéressante d'un côté pratique pour estimer la fonction de densité empirique sous-jacente. Cette fonction possède un argument `adjust` 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

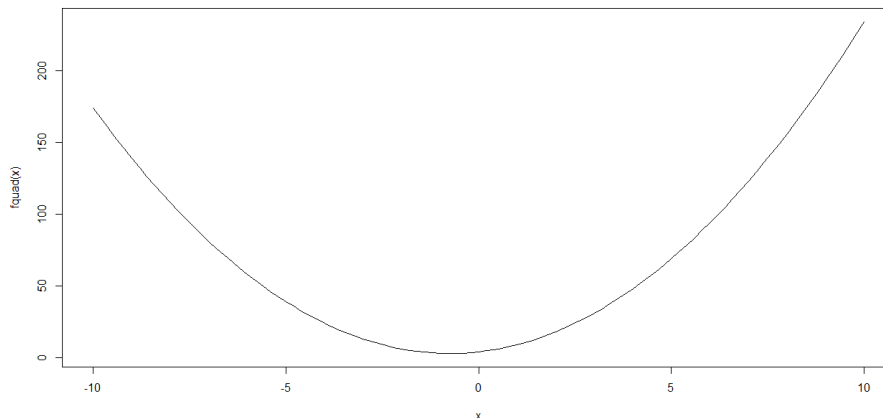


FIGURE 2.8 – Tracer une courbe avec la commande `curve`

distribution discrète. Inversement, une valeur supérieure à 1 aura pour effet de lisser davantage 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`. À titre d'exemple, nous serons en mesure de superposer la fonction de densité renvoyée par `density` grâce à la commande `lines`.

La commande `abline` [34] 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 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 vos analyses.

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, générera 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 ex-

plications 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 = gray(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)

```

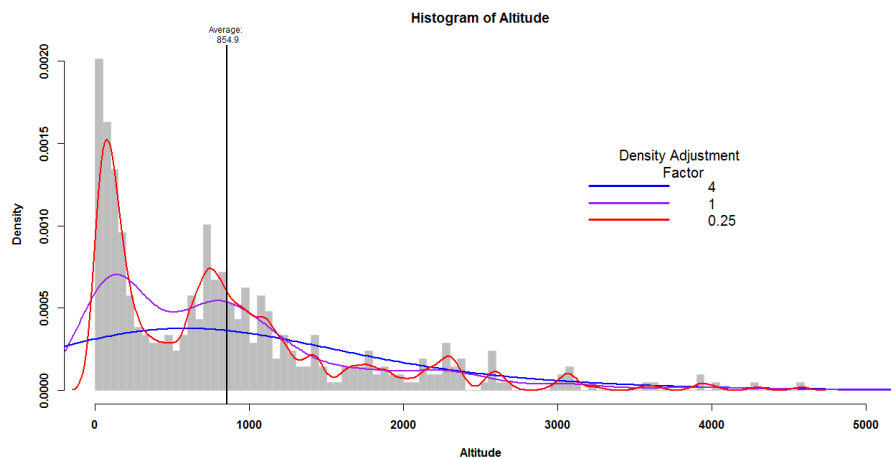


FIGURE 2.9 – Distribution des altitudes des aéroports canadiens



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 standards 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 graphiques 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 paquetages intègrent des modules de visualisation optimisée pour les objets qui leur sont propres. Il serait 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. ^{a b}

^a. Rien ne vous empêche de vous-mêmes surdéfinir vos propres fonctions d'affichage pour augmenter votre productivité.

^b. Il ne faudra pas oublier de préciser vos options d'affichage avant de partager vos programmes avec des personnes qui n'auraient pas accès à ces dites fonctions.

2.4 Outils d'analyse statistique en R

Une des caractéristiques 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. [13]

D'autres distributions deviendront aussi disponibles via des paquetages dédiés à cette fin. Le paquetage `actuar` [26] donne accès à plusieurs distributions supplémentaires communément utilisées en actuariat. La distribution *Pareto* en est un bon exemple.

Un aspect particulièrement intéressant de ces fonctions (qu'elles soient disponibles par défaut en R ou via l'importation d'un paquetage) est la constance dans leur implémentation. Pour chacune des distributions, nous retrouverons minimalement les quatre 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 $\langle ID_R \rangle$.

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 2.2 – Liste des distributions statistiques disponibles en R

$r\langle ID_R \rangle$ Permet de générer des valeurs aléatoires suivant 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`.¹⁰ 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

```

10. 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.

```

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)
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 qui consiste à générer des valeurs aléatoires suivant une loi uniforme de paramètre $a := 0$ et $b := 1$ pour ensuite trouver les quantiles 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

```

i**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 ([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. [20]

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 obligés 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 <- cut(compData$weight,25)
2 > distancesBinded <- cut(compData$distance,25)
3 > contingencyTable <- table(weightsBinded,distancesBinded)
4 > rownames(contingencyTable) <- NULL
5 > colnames(contingencyTable) <- NULL
6 > chisq.test(contingencyTable)
7
8   Pearson's Chi-squared test
9
10 data:  contingencyTable
11 X-squared = NaN, df = 576, p-value = NA
12
13 Warning message:
14 In chisq.test(contingencyTable) :
15   Chi-squared approximation may be incorrect

```

```

16 > contingencyTable <- rbind(contingencyTable[1:6,], colSums(
    contingencyTable[7:25,]))
17 > contingencyTable <- cbind(contingencyTable[,1:12], rowSums(
    contingencyTable[,13:25]))
18 > (independencyTest <- chisq.test(contingencyTable))
19
20 Pearson's Chi-squared test
21
22 data:  contingencyTable
23 X-squared = 52.312, df = 72, p-value = 0.961
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 = 0.89049, df = 99998, p-value = 0.3732
31 alternative hypothesis: true correlation is not equal to 0
32 95 percent confidence interval:
33  -0.003382045  0.009013785
34 sample estimates:
35      cor
36 0.002815978

```

Il est pertinent d'évoquer que le test du Khi carré possède des limitations importantes dans le cas de distributions devenant un peu trop clairessemées dans les extrêmes. C'est pour cette même raison que nous combinons les dernières lignes et colonnes de la table de contingence. [35] La *p-value* d'environ 96% nous indique que l'hypothèse nulle est juste. 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 37% 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)))/
3 + length(airportsCanada$province)
4 > compModel <- lm(price/(profitMargin*avgTaxRate) ~ distance +
    weight, compData)
5 > summary(compModel)
6
7 Call:
8 lm(formula = price/(profitMargin * avgTaxRate) ~ distance + weight,
9     data = compData)
10
11 Residuals:
12      Min       1Q   Median       3Q      Max

```

```

13 -30.0086   -4.6571    0.0157    4.6462   30.4167
14
15 Coefficients:
16             Estimate Std. Error t value Pr(>|t|)
17 (Intercept) 3.228e+01  7.510e-02  429.83  <2e-16 ***
18 distance    2.819e-02  9.202e-05  306.28  <2e-16 ***
19 weight      7.254e-01  9.607e-03   75.51  <2e-16 ***
20 ---
21 Signif. codes:
22 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
23
24 Residual standard error: 6.932 on 99997 degrees of freedom
25 Multiple R-squared:  0.4991, Adjusted R-squared:  0.4991
26 F-statistic: 4.982e+04 on 2 and 99997 DF, p-value: < 2.2e-16

```

L'appel de la fonction `lm` [39] est assez élémentaire. 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 [Tableau 2.2](#). Nous voyons aussi que pour chaque coefficient un test de Student a été effectué pour déterminer à quel point les estimations sont significativement différentes 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. [15]

Lorsque l'on compare les valeurs réellement utilisées dans le `??` et les coefficients estimés, nous voyons que ces derniers sont très proches les uns des autres. La [Tableau 2.3](#) fait la revue de ces valeurs.

Variable	Valeur réelle	Valeur estimée
distance	0.0275	0.02819
poids	0.7	0.7254

TABLE 2.3 – Comparaison entre les coefficients réels et estimés par régression linéaire



Lire des tables directement sur le web

Pour récupérer les niveaux de taxe pour chacune des provinces canadiennes, 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^a chaque fois que le script sera exécuté. 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 de la fonction `readHTMLTable`. Cette dernière ira lire le code source de la page web en question pour y détecter les occurrences des balises `<table>`, `<tr>`, `<th>` et `<td>`.

`<table><td>...</td></table>`

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

a. Nous serons toutefois dépendants de la structure de la page et tout changement majeur pourra compromettre la suite du programme. Des procédés de validation devront être mis en place si le script est mis en production.

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` [40] et `fitdistr` [46]. 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 ne pourra que minimiser la fonction étudiée et l'obligation de lui fournir des valeurs de départ. [49]

Parmi les arguments de la fonction `optim`, nous devons minimalement désigner les valeurs de départ de nos paramètres avec `par`¹¹ et fournir à `fn` la fonction à minimiser. 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

11. À ne pas confondre avec la fonction permettant de contrôler l'affichage graphique `par`.

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 aux résultats. 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   $par
6   [1] 0.4532121 0.2968149
7
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 employée 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 étudié. 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   )
3 $par
4 [1] 1.5001064 0.9999031
5
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 par laquelle 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 empiriques. Évidemment, nous pourrions très bien passer par `optim` pour réaliser le même travail moyennant un certain coût de complexité. Or, ce mal sera parfois nécessaire puisque la fonction `fitdistr` n’est définie que pour les distributions suivantes : [\[47\]](#)

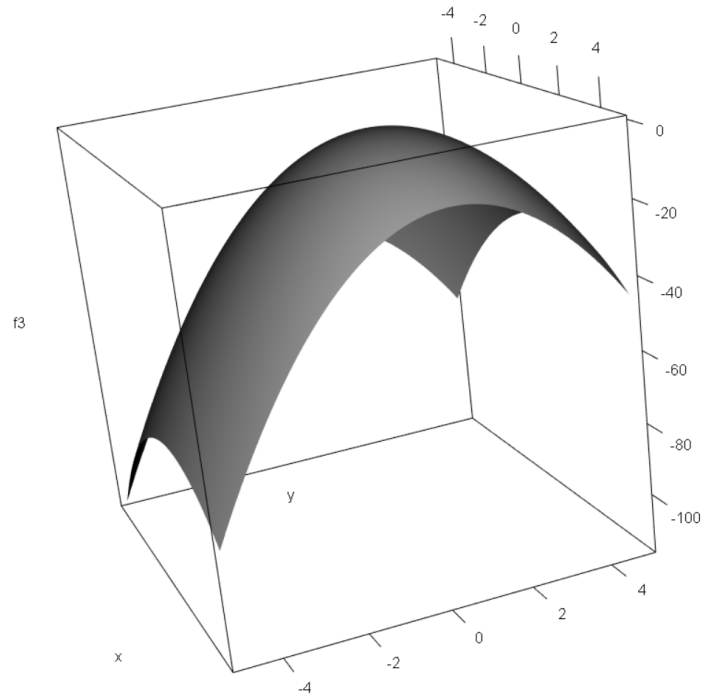


FIGURE 2.10 – Représentation graphique de la fonction **f3**

- | | |
|----------------------|---------------|
| ► 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 octroyer un comportement aussi flexible. 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

```

```

    and InvGaussian.
    (This case will be ignored)"
  stop(message)
}
if(nbparam != length(args))
{
  message <- paste("There is a mismatch between the number of
    arguments passed to the
    function and the number of arguments needed to
    the distribution.",
    "The",dist,"distribution is taking",nbparam,"
    parameters and",
    length(args),"parameters were given.")
  stop(message)
}
}
# Treament
if(nbparam == 1)
{
  param <- optim(par = args, function(par)
    -sum(do.call(eval(parse(text = paste("d",law,sep="")),c(list
      (data),par,log = TRUE)))),
    method = "Brent", lower = lower, upper = upper)
}
else{
  param <- optim(par = args, function(par)
    -sum(do.call(eval(parse(text = paste("d",law,sep="")),c(list
      (data),par,log = TRUE))))
}
# Deviance value
devValue <- sum((empPDF(x <- seq(0,max(data),0.1))-do.call(eval(
  parse(text = paste("d",law,sep="")),c(list(x),param$par)))*
  2)
100
# Return List
101 distFitList <- list()
102 distFitList$param <- param$par
103 distFitList$errorValue <- param$value
104 distFitList$devValue <- devValue
105 distFitList$convergence <- param$convergence
106 distFitList$nbiter <- param$counts[1]
107 distFitList
108 }
109

```

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 de précisions, la fonction `parse` permettra de créer des expressions non-évaluées. 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 (ID_R) 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` rend possible l'appel d'une fonction ayant un nombre d'arguments quelconque. Elle s'occupera de fournir une liste de paramètres à la fonction considérée pour autant que la fonction réceptrice accepte autant d'arguments que fournis et que les types correspondent. 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 `(...)`.

En ce qui a trait à notre étude de cas, l'ajustement de distribution fut nécessaire pour déterminer selon quelle loi les poids des colis pouvaient être modélisés. Une fois cette modélisation réalisée, il sera beaucoup plus simple de procéder à l'analyse par simulation de la [Figure 2.6](#). La [Tableau 2.4](#) consiste en un sommaire des résultats d'ajustement obtenus selon la fonction utilitaire `distFit` et la fonction `fitdistr` du paquetage **MASS**.

Somme toute, les résultats de notre fonction utilitaire sont drôlement impressionnant compte tenu de la flexibilité additionnelle que nous lui avons fourni en compa-

Distribution	distFit		fitdistr	
	$Param_1$	$Param_2$	$Param_1$	$Param_2$
Gamma	3.055112	0.8589427	3.054802645	0.857914221
InvGaussian	3.560507	8.643489	NA	NA
LogNormal	1.097521	0.5874043	1.097461059	0.587425157
Normal	3.560237	2.282066	3.560736550	2.281963899
Pareto	13 974 470	49 759 350	NA	NA
Weibull	1.704325	4.020126	1.704313997	4.019855513

TABLE 2.4 – Ajustement de distribution sur poids observés selon la fonction utilitaire `distFit` et la fonction `fitdistr` du paquetage MASS

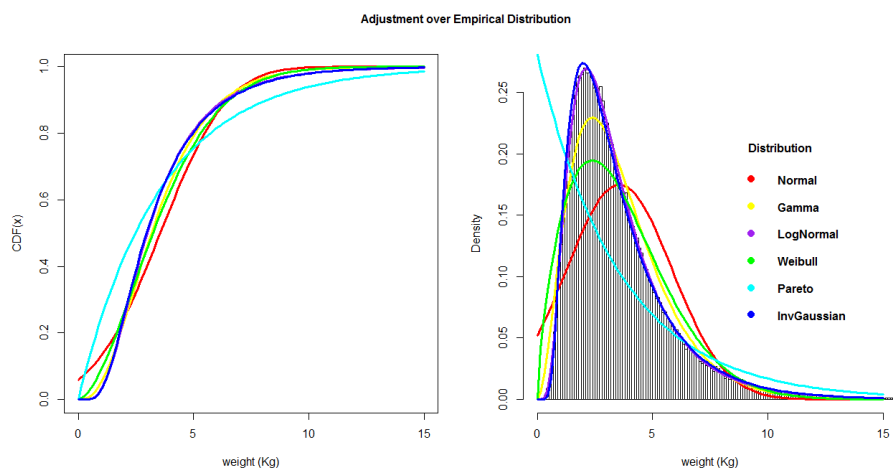


FIGURE 2.11 – Visualisation de l’ajustement sur la distribution empirique des poids (Cumulative à gauche et Densité à droite)

raison à son homologue. De l’autre côté de la médaille, notre fonction est loin d’être optimale et pourra s’avérer peu efficace sur des plus gros jeux de données.

De ces résultats, nous pouvons désormais conclure que les poids suivent une loi $LogNormal(\mu := 1.0975, \sigma := 0.587)$. Nous voyons toutefois qu’une loi $InvGaussian(\mu := 3.56, \sigma := 8.64)$ aurait pu être un bon choix ! Cela n’est pas une situation aussi rare qu’elle peut ne le paraître et le choix de la distribution devra s’appuyer sur d’autres considérations que coefficient d’ajustement. Nous pourrions préférer choisir une distribution avec un moins bon ajustement en raison de ses propriétés simplifiant ainsi le reste de l’analyse par exemple. En terminant, puisque nous sommes dans un contexte éducationnel et que les données ont été produites par le [Code Source A.1](#), il nous ai aussi possible de comparer la distribution sélectionnée avec celle qui est à l’origine de ces observations. Vous constaterez par vous même la précision de nos résultats.

Distribution Ajustée		Distribution Théorique	
<i>Param₁</i>	<i>Param₂</i>	<i>Param₁</i>	<i>Param₂</i>
1.0975	0.587	$\log 3 \approx 1.0986$	$\log 1.8 \approx 0.5878$

TABLE 2.5 – Comparaison des paramètres de la distribution ajustée et théorique des poids de colis (Kg)



Une variable plutôt invariante...

Il est possible de démontrer que si $X \sim \text{LogNormal}(\mu, \sigma)$ alors $aX \sim \text{LogNormal}(\mu + \log a, \sigma)$. C'est cette propriété qui nous permet de faire le pont entre le paramètre μ que l'on retrouve dans le [Code Source A.1](#) ($\log 3000$) et la valeur théorique que l'on affiche à la [Tableau 2.5](#). La différence provient du fait que nous avons créé notre fonction théorique en gramme pour ensuite la convertir en Kg en divisant par 1000. En appliquant le postulat, nous obtenons :

$$\log 3000 + \log \frac{1}{1000} = \log 3000 - \log 1000 = \log \frac{3000}{1000} = \log 3.$$

https://en.wikipedia.org/wiki/Log-normal_distribution#Related_distributions

2.6 Analyse par simulation en R

Quand bien même que la génération de nombres aléatoires ai déjà été abordée à la [Tableau 2.4](#), il serait injuste de s'imaginer que les capacités de R s'arrêtent là. R est un excellent langage pour faire des simulations complexes. L'estimation par simulation sera souvent un excellent moyen pour évaluer 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` [51]. Cette dernière sera utile dans les cas où nous cherchons à faire une pige aléatoire de taille quelconque (`size`) sur un ensemble de valeurs contenues dans un vecteur. Il sera possible de préciser si nous voulons faire une pige avec ou sans remise avec l'argument `replace` ainsi que la probabilité de survenance de chaque élément grâce à l'argument `prob`. Un aspect fort intéressant de cette fonction est sa capacité de faire des piges sur des valeurs textuelles. Le [Code Source 2.30](#) illustre un exemple simplifié 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 positives. 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 > altitude <- as.numeric(paste(airportsCanada$altitude))
2 > sample(altitude,size = 10, replace = TRUE)
3 [1] 1408 713 210 1912 703 602 3903 925 39 152
4 > probs <- pmax(0,altitude)/sum(pmax(0,altitude))
5 > sample(altitude,size = 10, replace = TRUE,prob = probs)
6 [1] 1023 3126 2364 1211 1000 1892 951 770 1653 2567
7 > sample(unique(as.character(paste(airportsCanada$name))),size =
      10,replace = FALSE)
8 [1] "Kangisujuaq (Wakeham Bay) Airport"
9 [2] "St Jean Airport"
10 [3] "Fort Frances Municipal Airport"
11 [4] "South Indian Lake Airport"
12 [5] "Prince George Airport"
13 [6] "Pembroke Airport"
14 [7] "Kugluktuk Airport"
15 [8] "Haines Junction Airport"
16 [9] "Edson Airport"
17 [10] "Eastmain River Airport"
```

En inspectant le [Code Source A.7](#), nous constatons la structure fonctionnelle et imbriquée du processus emprunté. 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 par simulation

```
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")
```

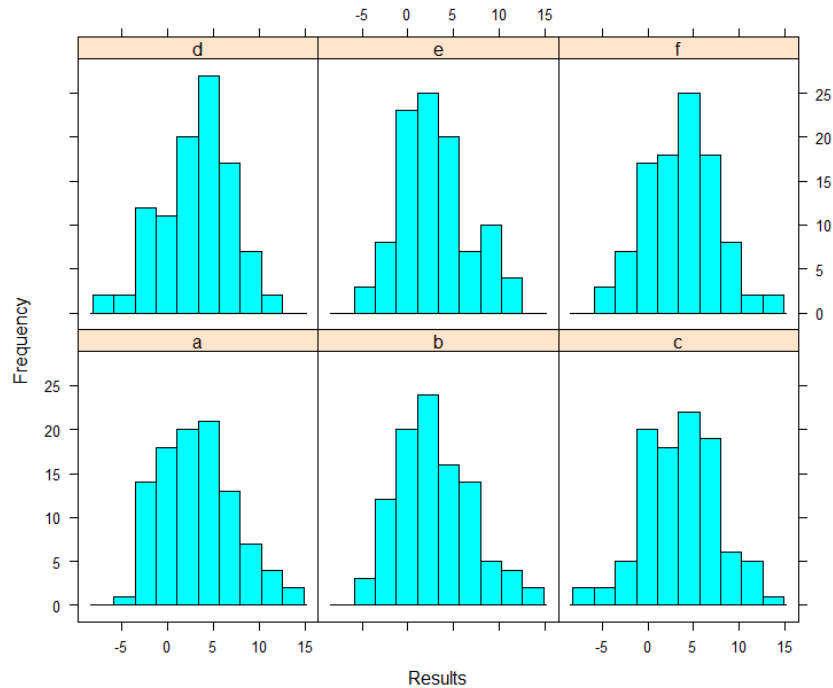


FIGURE 2.12 – Comparaison des résultats de simulation obtenus avec 6 réplicats



Une "Poisson" dans une pisciculture...

La distribution Poisson sera souvent à la base des processus de simulation 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 seront indépendants. 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 l’analyse par simulation.

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 d’étaler au grand jour les capacités de R tout en vous offrant un coffre d’outils qui facilitera grandement vos débuts avec ce langage. Il n’y a pas de secret pour apprendre à programmer, mais il existe certainement des moyens plus efficaces que d’autres. Selon nous, une connaissance adéquate de ce que l’on peut ou pas réaliser consiste en 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 de R. Vous chercherez ensuite à accumuler les ressources et connaissances nécessaires à sa résolution.¹² Ce n’est qu’en mettant en pratique vos connaissances que la maîtrise du langage sera atteignable. La route sera souvent tortueuse, mais le résultat donc bien satisfaisant.

Par ailleurs, dans une ère aussi axée sur le développement informatique et l’automatisation des tâches, il est de plus en plus important d’avoir de connaissances tangibles dans le domaine de la programmation. L’apprentissage 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 de sa simplicité. Comme nous pouvons le voir à la [Figure 2.13](#), R est toujours un langage d’actualité très prisé qui en vaut le détour en se plaçant au 12^{ème} rang selon le classement *RedMonk* [\[22\]](#).

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 récupérée 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,

12. Évidemment, rien ne vous empêche de vous créer des problèmes fictifs comme nous l’avons fait avec cette étude

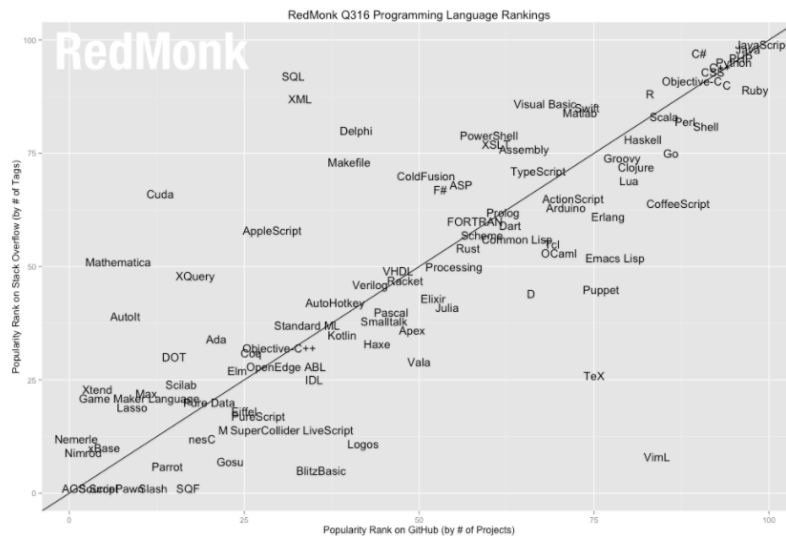


FIGURE 2.13 – Classement *RedMonk* des différents langages de programmation

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 du projet *OpenFlights* pour regarnir la variable `tzFormat`. (Voir [Appendice B](#))

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 composer le tout dans une si petite période de temps. À vous chers, acolytes, en espérant retravailler avec vous 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] Air Miles Calculator. <http://www.airmilescalculator.com/distance/yul-to-yvr/>.
- [3] An Open-Source JavaScript Library for Mobile-Friendly Interactive Maps. <http://leafletjs.com/>.
- [4] CSV vs. Delimited Flat Files : How to Choose. <http://www.thoughtspot.com/blog/csv-vs-delimited-flat-files-how-choose>.
- [5] Doxygen. <http://www.stack.nl/~dimitri/doxygen/>.
- [6] Font Awesome - The iconic font and CSS toolkit. <http://fontawesome.io/>.
- [7] GitHub. <https://github.com/>.
- [8] Introduction à la programmation en R. https://cran.r-project.org/doc/contrib/Goulet_introduction_programmation_R.pdf.
- [9] Introduction à R - Atelier du colloque R à Québec 2017 (GitHub). <https://github.com/vigou3/raquebec-intro>.
- [10] OpenFlights. <https://openflights.org/data.html>.
- [11] Package 'leaflet'. <https://cran.r-project.org/web/packages/leaflet/leaflet.pdf>.
- [12] Parc aérien d'Air Canada. <https://www.aircanada.com/ca/fr/aco/home/fly/onboard/fleet.html>.
- [13] Probabilités et Statistique avec R. <http://ljk.imag.fr/membres/Bernard.Ycart/mel/dr/node7.html>.
- [14] Projection (Système de). <http://www.emse.fr/tice/uved/SIG/Glossaire/co/Projection.html>.
- [15] Quick Guide : Interpreting Simple Linear Model Output in R. <http://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R>.
- [16] R Data Frames - dplyr vs sqldf. <http://www.starkingdom.co.uk/r-data-frames-dplyr-vs-sqldf/>.
- [17] R à Québec 2017. <http://raquebec.ulaval.ca/2017/programme-r-a-quebec-2017>.
- [18] roxygen2. <http://roxygen.org/>.
- [19] Structured Query Language (SQL). https://fr.wikipedia.org/wiki/Structured_Query_Language.

- [20] Tests statistiques avec R. <http://www.sthda.com/french/wiki/tests-statistiques-avec-r>.
- [21] Time Zone Map. <https://www.timeanddate.com/time/map/#!cities=574>.
- [22] Top Programming Languages to Learn in 2017. <https://www.codingame.com/blog/top-programming-languages-to-learn-in-2017/>.
- [23] 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>.
- [24] Statistics Canada. Boundary Files, Reference Guide. <http://www.statcan.gc.ca/pub/92-160-g/92-160-g2011002-eng.htm>.
- [25] Evan Siroky. timezone-boundary-builder. <https://github.com/evansiroky/timezone-boundary-builder>.
- [26] Vincent Goulet and Mathieu Pigeon. *Actuarial Functions and Heavy Tailed Distributions*, 2017. <https://cran.r-project.org/web/packages/actuar/index.html>.
- [27] G. Grothendieck. *Perform SQL Selects on R Data Frames*, 2014. <https://cran.r-project.org/web/packages/sqldf/index.html>.
- [28] David Kahle and Hadley Wickham. *Spatial Visualization with ggplot2*, 2016. <https://cran.r-project.org/web/packages/ggmap/index.html>.
- [29] Eric Muller. A shapefile of the TZ timezones of the world. <http://efele.net/maps/tz/world/>.
- [30] Jani Patokallio. Airline database. <https://raw.githubusercontent.com/jpatokal/openflights/master/data/airlines.dat>.
- [31] Jani Patokallio. Airport database. <https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat>.
- [32] Jani Patokallio. Route database. <https://raw.githubusercontent.com/jpatokal/openflights/master/data/routes.dat>.
- [33] Edzer Pebesma and Roger Bivand. *Classes and Methods for Spatial Data*, 2016. <https://cran.r-project.org/web/packages/sp/index.html>.
- [34] R documentation. *Add Straight Lines to a Plot*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/abline.html>.
- [35] R documentation. *Chi-squared Test of Independence*. <http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence>.
- [36] R documentation. *Cross Tabulation and Table Creation*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/table.html>.
- [37] R documentation. *Data Input*. <https://stat.ethz.ch/R-manual/R-devel/library/utils/html/read.table.html>.
- [38] R documentation. *Draw Function Plots*. <https://www.math.ucla.edu/~anderson/rw1001/library/base/html/curve.html>.
- [39] R documentation. *Fitting Linear Models*. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html>.
- [40] R documentation. *General-purpose Optimization*. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html>.
- [41] R documentation. *Generic X-Y Plotting*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html>.

- [42] R documentation. *Get or Set Working Directory*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/getwd.html>.
- [43] R documentation. *Histograms*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/hist.html>.
- [44] R documentation. *Invoke a Data Viewer*. <https://stat.ethz.ch/R-manual/R-devel/library/utils/html/View.html>.
- [45] R documentation. *Kernel Density Estimation*. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/density.html>.
- [46] R documentation. *Maximum-likelihood Fitting of Univariate Distributions*. <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html>.
- [47] R documentation. *Maximum-likelihood Fitting of Univariate Distributions*. <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html>.
- [48] R documentation. *Object Summaries*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/summary.html>.
- [49] R documentation. *Programmer en R/Optimiser une fonction*. https://fr.wikibooks.org/wiki/Programmer_en_R/Optimiser_une_fonction.
- [50] R documentation. *Random Number Generation*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/Random.html>.
- [51] R documentation. *Random Samples and Permutations*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/sample.html>.
- [52] R documentation. *Return the First or Last Part of an Object*. <https://stat.ethz.ch/R-manual/R-devel/library/utils/html/head.html>.
- [53] R documentation. *Set or Query Graphical Parameters*. <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/par.html>.
- [54] R documentation. *Subsetting Vectors, Matrices and Data Frames*. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/subset.html>.
- [55] 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 (path <- paste(getwd(), "..", sep="/"))
18
19 # Parameters of the simulation
20 n <- 100000
21 x <- matrix(c(runif(2*n)), ncol = 2, byrow = TRUE)
22
23 # Generate weights with a LogNormal distribution
24 mul <- log(3000)
25 sigma1 <- log(1.8)
26 exp(mul+sigma1**2/2)
27 exp(2*mul+4*sigma1**2/2)-exp(mul+sigma1**2/2)**2
28 weights <- round(qlnorm(x[,1], mul, sigma1)/1000,3)
29 hist(weights, breaks = 100, freq=FALSE)
30 mean(weights)
31 max(weights)
```

```

32
33 # Generate the errors on the weights
34 weightsTarifParam <- 0.7
35 weightsCost <- weights*weightsTarifParam
36 weightsError <- rnorm(n,mean(weightsCost),sd(weightsCost))
37 max(weightsError)
38 min(weightsError)
39 weightsFinalPrice <- weightsCost+weightsError
40 mean(weightsFinalPrice)
41 min(weightsFinalPrice)
42 var(weightsFinalPrice)
43
44 # Generate the distance with a LogNormal distribution
45 mu2 <- log(650)
46 sigma2 <- log(1.4)
47 distances <- round(qlnorm(x[,2],mu2,sigma2))
48 hist(distances,breaks = 100,freq=FALSE)
49 mean(distances)
50 max(distances)
51
52 # Generate the errors on the distances
53 distancesTarifParam <- 0.0275
54 distancesCost <- distances*distancesTarifParam
55 distancesError <- rnorm(n,mean(distancesCost),sd(distancesCost))
56 distancesFinalPrice <- distancesCost+distancesError
57 mean(distancesFinalPrice)
58 var(distancesFinalPrice)
59 max(distancesFinalPrice)
60 min(distancesFinalPrice)
61
62 # Generate total price
63 baseCost <- 10
64 taxRate <- 1.082408
65 profitMargin <- 1.15
66 totalCost <- round((baseCost + weightsFinalPrice +
67   distancesFinalPrice)*profitMargin*taxRate,2)
68 mean(totalCost)
69 var(totalCost)
70 max(totalCost)
71 min(totalCost)
72
73 # Export to csv format
74 dataExport <- cbind(weights,distances,totalCost)
75 colnames(dataExport) <- c("Poids (Kg)","Distance (Km)","Prix (CAD $
76   )")
77 write.csv(dataExport,
78   paste(path,"/ref/benchmark.csv",sep="")
79   ,row.names = FALSE, fileEncoding = "UTF-8")

```

Code Source A.2 – caseStudy1.R

```

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

```

```

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 (path <- paste(getwd(), "..", sep = "/"))
16 set.seed(31459)
17
18 # Extraction of airports.dat and routes.dat
19 airports <- read.csv("https://raw.githubusercontent.com/jpatokal/
    openflights/master/data/airports.dat",
20                     header = FALSE, na.strings=c("\\N",""))
21 routes <- read.csv("https://raw.githubusercontent.com/jpatokal/
    openflights/master/data/routes.dat",
22                   header = FALSE, na.strings=c("\\N",""))
23
24 # Columns names assignation based on the information available on
    the website
25 # https://openflights.org/data.html
26 colnames(airports) <- c("airportID", "name", "city", "country", "
    IATA", "ICAO",
27                        "latitude", "longitude", "altitude", "
    timezone", "DST",
28                        "tzFormat", "typeAirport", "Source")
29 colnames(routes) <- c("airline", "airlineID", "sourceAirport", "
    sourceAirportID",
30                      "destinationAirport", "destinationAirportID", "
    codeshare",
31                      "stops", "equipment")
32
33 # Filtering the observations relative to Canadian airports
34 airportsCanada <- subset(airports, country == "Canada")
35
36 # Extraction of general information about the variables contained
    in the dataset
37 View(airportsCanada)
38 summary(airportsCanada)
39 nbAirportCity <- table(airportsCanada$city)
40 (nbAirportCity <- head(sort(nbAirportCity, decreasing=TRUE)))
41
42 # Variable selection
43 # We will not use the typeAirport and Source variables since we
    only want to analyse air transport market
44 # We can also discard the country variable because we already
    filtered on Canadian airports
45 airportsCanada <- subset(airportsCanada, select = -c(country,
    typeAirport, Source))
46
47 # As seen in the summary, we do not have the IATA code of 27
    airports
48 subset(airportsCanada, is.na(IATA), select = c("airportID", "name", "
    IATA", "ICAO"))

```

```

49 # 82% of the time, the IATA code corresponds to the last three
    characters of the ICAO code
50 # We will use this relationship to assign default value for missing
    IATA codes
51 IATA <- as.character(airportsCanada$IATA)
52 ICAO <- as.character(airportsCanada$ICAO)
53 i <- is.na(IATA)
54 sum(IATA == substr(ICA0,2,4),na.rm = TRUE)/sum(!i)
55 IATA[i] <- substr(ICA0[i],2,4)
56 airportsCanada$IATA <- as.factor(IATA)
57 summary(airportsCanada)
58 # We do not need the ICAO code anymore
59 airportsCanada <- subset(airportsCanada, select = - ICAO)
60
61 # There are more than fifty airports having missing time zone
62 missingTZ <- subset(airportsCanada, is.na(timezone))
63 # Time zones only depend on the geographical position of the
    airports
64 # We will determine the real time zone using mapping tools
65 # install.packages("sp")
66 # install.packages("rgdal")
67 library(sp)
68 library(rgdal)
69 tz_world.shape <- readOGR(dsn=paste(path, "/ref/tz_world_2", sep=""),
    layer="combined_shapefile")
70 unknown_tz <- subset(airportsCanada, is.na(tzFormat), c("airportID", "
    name", "longitude", "latitude"))
71 sppts <- SpatialPoints(subset(unknown_tz, select = c("longitude", "
    latitude")))
72 proj4string(sppts) <- CRS("+proj=longlat")
73 sppts <- spTransform(sppts, proj4string(tz_world.shape))
74 merged_tz <- cbind(unknown_tz, over(sppts, tz_world.shape))
75
76 # To retrieved the province of each airport, we will use the same
    technique
77 prov_terr.shape <- readOGR(dsn=paste(path, "/ref/prov_terr", sep=""),
    layer="gpr_000b11a_e")
78 unknown_prov <- subset(airportsCanada, select = c("airportID", "city", "
    longitude", "latitude"))
79 sppts <- SpatialPoints(subset(unknown_prov, select = c("longitude", "
    latitude")))
80 proj4string(sppts) <- CRS("+proj=longlat")
81 sppts <- spTransform(sppts, proj4string(prov_terr.shape))
82 merged_prov <- cbind(airportsCanada, over(sppts, prov_terr.shape))
83
84 # We merge the new information with the main database
85 # install.packages("sqldf")
86 # install.packages("tcltk")
87 library(sqldf)
88 library(tcltk)
89 airportsCanada <- sqldf("
90     select
91     a.*,
92     coalesce(a.tzFormat,b.TZID) as tzMerged,
93     c.PRENAME as provMerged
94 from airportsCanada a
95 left join merged_tz b

```

```

96   on a.airportID = b.airportID
97   left join merged_prov c
98   on a.airportID = c.airportID
99   order by a.airportID")
100 airportsCanada <- data.frame(as.matrix(airportsCanada))
101
102 # We do not need timezone, DST, city and tzFormat variables anymore
103 airportsCanada <- subset(airportsCanada, select = -c(timezone, DST,
104   tzFormat, city ))
105 summary(airportsCanada)
106
107 # Rename of the merged variables
108 # install.packages("plyr")
109 library(plyr)
110 airportsCanada <- rename(airportsCanada, c("tzMerged"="tzFormat", "
111   provMerged"="province"))
112 summary(airportsCanada)
113
114 # Extraction of canadian internal routes
115 routesCanada <- sqldf("
116   select *
117   from routes
118   where sourceAirportID in (select distinct airportID
119     from airportsCanada)
120     and destinationAirportID in (select distinct airportID
121     from airportsCanada)")
122 routesCanada <- data.frame(as.matrix(routesCanada ))
123 summary(routesCanada)
124
125 # Since almost all routes are direct, we dont need the codeshare
126 # and stops variables
127 summary(routesCanada$stops)
128 routesCanada <- subset(routesCanada, select = -c(codeshare, stops))
129 summary(routesCanada)
130
131 # Create a map showing the different airports
132 # install.packages("ggmap")
133 library(ggmap)
134 map <- get_map(location = "Canada", zoom = 3)
135 lon <- as.numeric(paste(airportsCanada$longitude))
136 lat <- as.numeric(paste(airportsCanada$latitude))
137 airportsCoord <- as.data.frame(cbind(lon, lat))
138 (mapPoints <- ggmap(map) + geom_point(data=airportsCoord, aes(lon,
139   lat), alpha = 0.5))
140
141 # Create a second map showing all possible routes between the
142 # canadian airports.
143 routesCoord <- sqldf("
144   select
145     a.sourceAirport,
146     a.destinationAirport,
147     b.longitude as sourceLon,
148     b.latitude as sourceLat,
149     c.longitude as destLon,
150     c.latitude as destLat
151   from routesCanada a
152   left join airportsCanada b

```

```

148   on a.sourceAirport = b.IATA
149   left join airportsCanada c
150   on a.destinationAirport = c.IATA")
151 lonBeg <- as.numeric(paste(routesCoord$sourceLon))
152 latBeg <- as.numeric(paste(routesCoord$sourceLat))
153 lonEnd <- as.numeric(paste(routesCoord$destLon))
154 latEnd <- as.numeric(paste(routesCoord$destLat))
155 routesCoord <- as.data.frame(cbind(lonBeg, latBeg, lonEnd, latEnd))
156 (mapRoutes <- mapPoints +
157   geom_segment(data=routesCoord, aes(x=lonBeg, y=latBeg, xend=lonEnd
158     , yend=latEnd), alpha = 0.5))
159 # Calculation of a standard traffic index based on the number
160   arrivals and departures
161 arrivalFlights <- table(routesCanada$destinationAirport)
162 departureFlights <- table(routesCanada$sourceAirport)
163 totalFlights <- arrivalFlights + departureFlights
164 max(totalFlights)
165 mean(totalFlights)
166 var(totalFlights)
167 sd(totalFlights)
168 head(sort(totalFlights, decreasing = TRUE), n = 30)
169 IATA <- names(totalFlights)
170 combinedIndex <- round(totalFlights/max(totalFlights), 3)
171 combinedIndexTable <- data.frame(IATA, as.vector(totalFlights), as.
172   vector(combinedIndex),
173   row.names = NULL)
174 colnames(combinedIndexTable) <- c("IATA", "totalFlights", "
175   combinedIndex")
176 airportsCanada <- sqldf("
177   select
178   a.*,
179   coalesce(b.combinedIndex, 0) as combinedIndex
180   from airportsCanada a
181   left join combinedIndexTable b
182   on a.IATA = b.IATA")
183 airportsCanada <- data.frame(as.matrix(airportsCanada ))
184 # Create a map to visualize the traffic index using it as the size
185   of the circle
186 TrafficData <- subset(airportsCanada, as.numeric(paste(combinedIndex
187   ) > 0.05))
188 lon <- as.numeric(paste(TrafficData$longitude))
189 lat <- as.numeric(paste(TrafficData$latitude))
190 size <- as.numeric(paste(TrafficData$combinedIndex))
191 airportsCoord <- as.data.frame(cbind(lon, lat, size))
192 mapPoints <-
193   ggmap(map) +
194   geom_point(data=TrafficData, aes(x=lon, y=lat, size=size), alpha=0.5,
195     shape=16)
196 (mapTraffic <-
197   mapPoints +
198   scale_size_continuous(range = c(0, 20), name = "Traffic Index"))
199 # Interactive map with markers of some principal airports
200 # install.packages("leaflet")
201 library(leaflet)

```

```

198 url <- "http://hiking.waymarkedtrails.org/en/routebrowser/1225378/
      gpx"
199 download.file(url, destfile = paste(path, "/ref/worldRoutes.gpx", sep
    =""), method = "wget")
200 worldRoutes <- readOGR(paste(path, "/ref/worldRoutes.gpx", sep=""),
    layer = "tracks")
201
202 # Defining the description text to be displayed by the markers
203 markersData <- subset(airportsCanada, IATA %in% c("YUL", "YVR", "YYZ",
    "YQB"))
204 markersWeb <- c("https://www.aeroportoquebec.com/fr/pages/accueil"
    ,
205                 "http://www.admtl.com/",
206                 "http://www.yvr.ca/en/passengers",
207                 "https://www.torontopearson.com/")
208 descriptions <- paste("<b><FONT COLOR=#31B404> Airport Details</FONT
    ></b> <br>",
209                      "<b>IATA: <a href=", markersWeb, ">", markersData$I
    ATA, "</a></b><br>",
210                      "<b>Name:</b>", markersData$name, "<br>",
211                      "<b>Coord.</b>: (", markersData$longitude, ",",
    markersData$latitude, ") <br>",
212                      "<b>Traffic Index:</b>", markersData$
    combinedIndex)
213
214 # Defining the icon to be add on the markers from fontawesome
    library
215 icons <- awesomeIcons(icon = "paper-plane",
216                       iconColor = "black",
217                       library = "fa")
218
219 # Combinaison of the different components in order to create a
    standalone map
220 (mapTraffic <- leaflet(worldRoutes) %>%
221   addTiles(urlTemplate = "http://{s}.basemaps.cartocdn.com/light_
    all/{z}/{x}/{y}.png") %>%
222   addCircleMarkers(stroke = FALSE, data = TrafficData,
223                   ~as.numeric(paste(longitude)), ~as.numeric(
    paste(latitude)),
224                   color = "black", fillColor = "green",
225                   radius = ~as.numeric(paste(combinedIndex))*30,
    opacity = 0.5) %>%
226   addAwesomeMarkers(data = markersData, ~as.numeric(paste(
    longitude)),
227                   ~as.numeric(paste(latitude)), popup =
    descriptions, icon=icons))
228
229 # Resizing of the map
230 mapTraffic$width <- 875
231 mapTraffic$height <- 700
232
233 # Export of the map into html format
234 # install.packages("htmlwidgets")
235 library(htmlwidgets)
236 saveWidget(mapTraffic, paste(path, "/ref/leafletTraffic.html", sep = "
    "), selfcontained = TRUE)

```

Code Source A.3 – caseStudy2.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 # install.packages("geosphere")
15 library(geosphere)
16
17 #' Distance between two airports
18 #'
19 #' @param sourceIATA The IATA of the departure airport
20 #' @param destIATA The IATA of the arrival airport
21 #' @return A list containing:
22 #'   source ,
23 #'   dest ,
24 #'   value ,
25 #'   metric ,
26 #'   xy_dist ,
27 #'   sourceIndex and
28 #'   destIndex
29 #' @examples
30 #' airportsDist("YUL","YQB")
31 #' airportsDist("YUL","YVR")
32 #'
33 airportsDist <- function(sourceIATA,destIATA)
34 {
35   sourceFindIndex <- match(sourceIATA,airportsCanada$IATA)
36   if(is.na(sourceFindIndex))
37   {
38     stop(paste("sourceIATA :",sourceIATA,"is not a valid IATA code"
39               ))
40   }
41   destFindIndex <- match(destIATA,airportsCanada$IATA)
42   if(is.na(destFindIndex))
43   {
44     stop(paste("destIATA :",destIATA,"is not a valid IATA code"))
45   }
46   sourceLon <- as.numeric(paste(airportsCanada$longitude)[
47     sourceFindIndex])
48   sourceLat <- as.numeric(paste(airportsCanada$latitude)[
49     sourceFindIndex])
50   sourceCoord <- c(sourceLon,sourceLat)
51   destLon <- as.numeric(paste(airportsCanada$longitude)[
52     destFindIndex])
53   destLat <- as.numeric(paste(airportsCanada$latitude)[
54     destFindIndex])

```



```

50 destCoord <- c(destLon, destLat)
51 airportDistList <- list()
52 airportDistList$source <- sourceIATA
53 airportDistList$dest <- destIATA
54 airportDistList$value <- round(distGeo(sourceCoord, destCoord)/
    1000)
55 airportDistList$metric <- "Km"
56 airportDistList$xy_dist <- sqrt((sourceLon - destLon)**2 + (
    sourceLat - destLat)**2)
57 airportDistList$sourceIndex <- sourceFindIndex
58 airportDistList$destIndex <- destFindIndex
59 airportDistList
60 }
61 airportsDist("AAA", "YQB")
62 airportsDist("YUL", "AAA")
63 airportsDist("YPA", "YQB")
64 airportsDist("YUL", "YQB")
65 airportsDist("YUL", "YQB")$value
66 airportsDist("YUL", "YVR")$value
67 airportsDist("YUL", "YYZ")$value
68
69
70 # install.packages("lubridate")
71 library(lubridate)
72
73 #' Establish the time of arrival
74 #'
75 #' @param sourceIATA The IATA of the departure airport
76 #' @param destIATA The IATA of the arrival airport
77 #' @return A list containing:
78 #'   sourceIATA,
79 #'   destIATA,
80 #'   departureTime,
81 #'   avgCruiseSpeed,
82 #'   flightTime,
83 #'   departureTZ,
84 #'   arrivalTZ and
85 #'   value
86 #' @examples
87 #' arrivalTime("YUL", "YQB")
88 #' arrivalTime("YUL", "YVR")
89 #'
90 arrivalTime <- function(sourceIATA, destIATA)
91 {
92   topSpeed <- 850
93   adjustFactor <- list()
94   adjustFactor$a <- 0.0001007194 # found by interpolation (not
    included)
95   adjustFactor$b <- 0.4273381 # found by interpolation (not
    included)
96   arrivalTimeList <- list()
97   arrivalTimeList$source <- sourceIATA
98   arrivalTimeList$dest <- destIATA
99   arrivalTimeList$departureTime <- Sys.time()
100   distance <- airportsDist(sourceIATA, destIATA)
101   cruiseSpeed <- (distance$value*adjustFactor$a + adjustFactor$b)*
    topSpeed

```

```

102 arrivalTimeList$avgCruiseSpeed <- cruiseSpeed
103 arrivalTimeList$flightTime <- ms(round(distance$value/cruiseSpeed
    *60, digits = 1))
104 arrivalTimeList$departureTZ <- paste(airportsCanada[distance$
    sourceIndex, "tzFormat"])
105 arrivalTimeList$arrivalTZ <- paste(airportsCanada[distance$
    destIndex, "tzFormat"])
106 arrivalTimeList$value <- with_tz(arrivalTimeList$departureTime +
    arrivalTimeList$flightTime,
107                                 tzone = arrivalTimeList$
                                    arrivalTZ)
108 arrivalTimeList
109 }
110 arrivalTime("AAA", "YYZ")
111 arrivalTime("YUL", "AAA")
112 arrivalTime("YUL", "YYZ")
113 arrivalTime("YUL", "YVR")
114 arrivalTime("YUL", "YYZ")$value
115 difftime(arrivalTime("YUL", "YQB")$value, Sys.time())
116 difftime(arrivalTime("YUL", "YVR")$value, Sys.time())
117 difftime(arrivalTime("YUL", "YYZ")$value, Sys.time())
118
119
120 # Import tax rates by province directly from the web
121 #install.packages("XML")
122 #install.packages("RCurl")
123 #install.packages("rlist")
124 library(XML)
125 library(RCurl)
126 library(rlist)
127 theurl <- getURL("http://www.calculconversion.com/sales-tax-
    calculator-hst-gst.html",
128                 .opts = list(ssl.verifypeer = FALSE))
129 tables <- readHTMLTable(theurl)
130 provinceName <- as.character(sort(unique(airportsCanada$province)))
131 taxRates <- as.data.frame(cbind(provinceName, as.numeric(sub("%", "",
    tables$'NULL'[-13,5]))/100+1))
132 colnames(taxRates) <- c("province", "taxRate")
133 taxRates
134
135
136 #' Shipping cost calculation
137 #'
138 #' @param sourceIATA The IATA of the departure airport.
139 #' @param destIATA The IATA of the arrival airport.
140 #' @param weight The weight of the shipping.
141 #' @param percentCredit A double with a default value of 0.
142 #' @param dollarCredit A double with a default value of 0.
143 #' @return A list containing:
144 #'   distance ,
145 #'   weight ,
146 #'   distanceFactor ,
147 #'   weightFactor ,
148 #'   fixedCost ,
149 #'   profitMargin ,
150 #'   percentCredit ,
151 #'   dollarCredit ,

```

```

152 #' minimalDist ,
153 #' trafficIndex ,
154 #' baseCost ,
155 #' automatedCredit ,
156 #' taxRate and
157 #' price
158 #' @examples
159 #' shippingCost("YUL","YQB")
160 #' shippingCost("YUL","YVR")
161 #'
162 shippingCost <- function(sourceIATA, destIATA, weight ,
163                           percentCredit = 0, dollarCredit = 0)
164 {
165   routeConcat <- as.character(paste(routesCanada$sourceAirport ,
166                                     routesCanada$destinationAirport))
167   if (is.na(match(paste(sourceIATA,destIATA),routeConcat)))
168   {
169     stop(paste("the combination of sourceIATA and destIATA (",
170               sourceIATA,"-",destIATA,")
171               do not corresponds to existing route"))
172   }
173   if(weight < 0 || weight > 30)
174   {
175     stop("The weight must be between 0 and 30 Kg")
176   }
177   if(percentCredit < 0 || percentCredit > 1)
178   {
179     stop("The percentage of credit must be between 0 % and 100 %")
180   }
181   if(dollarCredit < 0)
182   {
183     stop("The dollar credit must be superior to 0 $")
184   }
185   minimalDist = 100
186   distance <- airportsDist(sourceIATA, destIATA)
187   if (distance$value < minimalDist)
188   {
189     stop(paste("The shipping distance is under the minimal
190               requirement of",minDist,"Km"))
191   }
192   # Pricing variables
193   distanceFactor <- 0.03
194   weightFactor <- 0.8
195   fixedCost <- 3.75
196   profitMargin <- 1.12
197   # Traffic Index
198   trafficIndexSource <- as.numeric(paste(airportsCanada[distance$
199                                     sourceIndex,"combinedIndex"]))
200   trafficIndexDest <- as.numeric(paste(airportsCanada[distance$
201                                     destIndex,"combinedIndex"]))
202
203

```

```

204 # Calculation of the base cost
205 baseCost <- fixedCost + (distance$value*distanceFactor + weight*
    weightFactor)/
206     (trafficIndexSource*trafficIndexDest)
207
208 # Additional automated credits
209 automatedCredit <- 1
210 # Lightweight
211 automatedCredit <- automatedCredit * ifelse(weight < 2, 0.5, 1)
212 # Gold Member
213 automatedCredit <- automatedCredit * ifelse(baseCost > 100, 0.9,
    1)
214 # Partnership
215 automatedCredit <- automatedCredit * switch(sourceIATA,
216     "YUL" = 0.85,
217     "YHU" = 0.95,
218     "YMX" = 0.95,
219     "YYZ" = 0.9,
220     "YKZ" = 0.975,
221     "YTZ" = 0.975,
222     "YZD" = 0.975)
223 # The Migrator
224 if(distance$value > 3000)
225 {
226     automatedCredit <- automatedCredit * 0.9
227 }
228 else if(distance$value <= 3000 & distance$value > 2500)
229 {
230     automatedCredit <- automatedCredit * 0.8775
231 }
232 else if(distance$value <= 2500 & distance$value > 2000)
233 {
234     automatedCredit <- automatedCredit * 0.85
235 }
236
237 taxRate <- as.numeric(paste(taxRates[match(airportsCanada[
    distance$sourceIndex, "province"],
238     taxRates$province), "
    taxRate"]))
239 price <- round(pmax(fixedCost*profitMargin*automatedCredit*
    taxRate,
240     (baseCost*automatedCredit*profitMargin -
    dollarCredit)
241     *(1 - percentCredit)*taxRate), 2)
242
243 # Return List
244 shippingCostList <- list()
245 shippingCostList$distance <- distance
246 shippingCostList$weight <- weight
247 shippingCostList$distanceFactor <- distanceFactor
248 shippingCostList$weightFactor <- weightFactor
249 shippingCostList$fixedCost <- fixedCost
250 shippingCostList$profitMargin <- profitMargin
251 shippingCostList$percentCredit <- percentCredit
252 shippingCostList$dollarCredit <- dollarCredit
253 shippingCostList$minimalDist <- minimalDist
254 shippingCostList$trafficIndex <- list(trafficIndexSource,

```

```

        trafficIndexDest)
255 shippingCostList$baseCost <- baseCost
256 shippingCostList$automatedCredit <- 1-automatedCredit
257 shippingCostList$taxRate <- taxRate
258 shippingCostList$price <- price
259 shippingCostList
260 }
261 shippingCost("YUL", "YVR", 1)
262 shippingCost("YUL", "YQB", 1)
263 shippingCost("YUL", "YVR", 30)
264 shippingCost("YUL", "YQB", 30)

```

Code Source A.4 – caseStudy3.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
15 ##### Question 3 #####
16 # We create a visual support plotting the relationship between the
  weight and the price
17 # Source airport : YUL
18 # Destination airport : YQB, YVR, YYZ and YYC
19 curve(shippingCost("YUL", "YQB", x)$price, 0.01, 50, ylim=c(0, 200),
20       main="Shipping Price Variation according to the Weight", xlab=
  "weight (Kg)",
21       ylab="price (CND $)", lwd = 2)
22 curve(shippingCost("YUL", "YVR", x)$price, 0.01, 50, xlab="weight (Kg)",
23       ylab="price (CND $)", add=TRUE, col = "red", lwd = 2)
24 curve(shippingCost("YUL", "YYZ", x)$price, 0.01, 50, xlab="weight (Kg)",
25       ylab="price (CND $)", add=TRUE, col = "blue", lwd = 2)
26 curve(shippingCost("YUL", "YYC", x)$price, 0.01, 50, xlab="weight (Kg)",
27       ylab="price (CND $)", add=TRUE, col = "purple", lwd = 2)
28 text(x=c(25, 25, 25, 25), y=c(50, 90, 140, 175), c("YUL-YYZ", "YUL-YQB", "YUL
  -YVR", "YUL-YYC"),
29       adj = 0.5, cex = 0.75, font = 2, col = c("blue", "black", "red", "
  purple"))

```

Code Source A.5 – caseStudy4.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, "/ref/benchmark.csv", sep=""))
17 View(compData)
18 colnames(compData) <- c("weight", "distance", "price")
19 summary(compData)
20
21 # Weight distribution
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 distribution
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 <- cut(compData$weight, 25)
52 distancesBinded <- cut(compData$distance, 25)
53 contingencyTable <- table(weightsBinded, distancesBinded)
54 rownames(contingencyTable) <- NULL
55 colnames(contingencyTable) <- NULL
56 chisq.test(contingencyTable)

```

```

57 contingencyTable <- rbind(contingencyTable[1:6,], colSums(
    contingencyTable[7:25,]))
58 contingencyTable <- cbind(contingencyTable[,1:12], rowSums(
    contingencyTable[,13:25]))
59 independencyTest <- chisq.test(contingencyTable)
60 head(independencyTest$expected)
61 head(independencyTest$observed)
62 head(independencyTest$stdres)
63 independencyTest
64 cor.test(compData$weight, compData$distance, method = "pearson")
65
66 # Linear model
67 # we assume the same profit margin to simplify the situation
68 # We let an intercept because shipping pricing always have fixed
    cost component
69 profitMargin <- 1.12
70 avgTaxRate <- sum(table(airportsCanada$province)*as.numeric(paste(
    taxRates$taxRate)))/
71 length(airportsCanada$province)
72 compModel <- lm(price/(profitMargin*avgTaxRate) ~ distance + weight
    , compData)
73 summary(compModel)
74
75 # We plot the model
76 par(mfrow=c(2,2))
77 plot(compModel)

```

Code Source A.6 – caseStudy5.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 #' 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))),c(list
120        (data),par,log = TRUE)))
121    method = "Brent", lower = lower, upper = upper)
122  }
123  else{
124    param <- optim(par = args, function(par)
125      -sum(do.call(eval(parse(text = paste("d",law,sep=""))),c(list
126        (data),par,log = TRUE))))
127  }
128  # Deviance value
129  devValue <- sum((empPDF(x <- seq(0,max(data),0.1))-
130    do.call(eval(parse(text = paste("d",law,sep="")
131      )),c(list(x),param$par)))*2)
132
133  # Return List
134  distFitList <- list()
135  distFitList$param <- param$par
136  distFitList$errorValue <- param$value
137  distFitList$devValue <- devValue

```

```

131 distFitList$convergence <- param$convergence
132 distFitList$nbiter <- param$counts[1]
133 distFitList
134 }
135
136 (resultDistFitting <- sapply(distName, function(x) unlist(distFit(
    compData$weight, x, 1, 1))))
137
138 law <- c("norm", "gamma", "lnorm", "weibull", "pareto", "invgauss")
139 col <- c("red", "yellow", "purple", "green", "cyan", "blue")
140 x <- seq(0, 30, 0.1)
141
142 # Visualization of the goodness of the fit
143 par(mfrow = c(1, 2), font = 2)
144 plot(function(x) empCDF(x), xlim = c(0, 15), main = "", xlab = "
    weight (Kg)", ylab = "CDF(x)")
145 invisible(sapply(1:length(law),
146                 function(i) curve(do.call(eval(parse(text = paste(
    "p", law[i], sep = ""))),
147                                   c(list(x), as.vector(
    resultDistFitting[c
    (1:2), i]))),
148                                   add = TRUE, lwd = 3, col = col[i
    ]))))
149 hist(compData$weight, xlim = c(0, 15), main = "", xlab = "weight (Kg
    )", breaks = 300, freq = FALSE)
150 invisible(sapply(1:length(law),
151                 function(i) curve(do.call(eval(parse(text = paste(
    "d", law[i], sep = ""))),
152                                   c(list(x), as.vector(
    resultDistFitting[c
    (1:2), i]))),
153                                   add = TRUE, lwd = 3, col = col[i
    ]))))
154 legend(x="right", y = "center", distName, inset = 0.1, col = col,
    pch = 20, pt.cex = 2, cex = 1,
155         ncol = 1, bty = "n", text.width = 2, title = "Distribution")
156 mtext("Adjustment over Empirical Distribution", side = 3, line =
    -2, outer = TRUE)
157
158 # We thus choose the LogNormal distribution which possesses the
    smallest deviance and the best fit
159 distChoice <- "LogNormal"
160 (paramAdjust <- resultDistFitting[c(1:2), match(distChoice, distName)
    ])
161
162 # It is also possible to do the equivalent with fitdistr of the
    MASS library
163 library("MASS")
164 (fit.normal <- fitdistr(compData$weight, "normal"))
165 (fit.gamma <- fitdistr(compData$weight, "gamma"))
166 (fit.lognormal <- fitdistr(compData$weight, "lognormal"))
167 (fit.weibull <- fitdistr(compData$weight, "weibull"))
168
169 altitude <- as.numeric(paste(airportsCanada$altitude))
170 sample(altitude, size = 10, replace = TRUE)
171 probs <- pmax(0, altitude)/sum(pmax(0, altitude))

```

```

172 sample(altitude, size = 10, replace = TRUE, prob = probs)
173 sample(unique(as.character(paste(airportsCanada$name))), size = 10,
         replace = FALSE)

```

Code Source A.7 – caseStudy6.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 = " "),
16                 .opts = list(ssl.verifypeer = FALSE))
17 tables <- readHTMLTable(theurl)
18 lambdaTable <- as.data.frame(tables$"NULL")
19 colnames(lambdaTable) <- c("Month", "Avg3yrs")
20 lambdaTable
21
22 # The possible routes are filtered as having a departure from 'YUL'
23 # and a distribution is created according to the destination index
24 simAirportsDests <- routesCanada$destinationAirport[routesCanada$
   sourceAirport == "YUL"]
25 simCombinedIndex <- combinedIndex[names(combinedIndex) %in%
   simAirportsDests]
26 airportsDensity <- simCombinedIndex/sum(simCombinedIndex)
27
28 # Function for the simulation of the shipment prices.
29 simulShipmentPrice <- function(Arrival, Weight)
30 {
31   ownPrice <- ifelse(is(testSim <- try(shippingCost("YUL", Arrival,
   Weight)$price, silent = TRUE)),
32                     "try-error", NA, testSim)
33   distance <- airportsDist("YUL", Arrival)$value
34   nd <- as.data.frame(cbind(distance, Weight))
35   colnames(nd) <- c("distance", "weight")
36   compPrice <- predict(compModel, newdata = nd)
37   customerChoice <- ifelse(is.na(ownPrice), 0, ifelse(ownPrice <
   compPrice, 1, 0))
38   rbind(Arrival, distance, Weight, ownPrice, compPrice, customerChoice)
39 }
40
41 # Function for the simulation of the shipment parameters, weights
   and destinations.
42 simulShipment <- function(simNbShipments)
43 {

```

```

44 # Weights are then generated for each of the packages.
45 simWeights <- eval(parse(
46   text = paste("r",law[match(distChoice,distName)],sep = "")))(
47   simNbShipments,
48   paramAdjust
49   [1],
50   paramAdjust
51   [2])
52
53 # We finally generate a destination for each package (the
54   departure will always be from 'YUL').
55 simArrivals <- sample(size = simNbShipments,names(airportsDensity
56   ),prob = airportsDensity,
57   replace = TRUE)
58 supply(seq(1,simNbShipments),function(x) simulShipmentPrice(
59   simArrivals[x],simWeights[x]))
60 }
61
62 # Function for overall simulation
63 simulOverall <-function()
64 {
65   # We generate n observations of the Poisson distribution with
66   param = sum (lambda).
67   # We know from probability notion that the sum of independent
68   Poisson distribution follows
69   # a Poisson distribution with param = sum (lambda).
70   simNbShipments <- rpois(1 ,lambda = sum(as.numeric(paste(
71     lambdaTable$Avg3yrs))))
72   # We simulate each shipment
73   simulShipment(simNbShipments)
74 }
75
76 nsim <- 1
77 simulResults <- replicate(nsim, simulOverall(),simplify = FALSE)
78 (marketShareSales <- apply(1:nsim,function(x)
79   sum(as.numeric(simulResults[[x]][6,]))/length(simulResults[[x
80     ]][6,])))
81 (ownRevenus <- sum(as.numeric(simulResults[[1]][4,])*
82   as.numeric(simulResults[[1]][6,]),na.rm = TRUE
83   ))
84 (compRevenus <- sum(as.numeric(simulResults[[1]][5,])*
85   (1-as.numeric(simulResults[[1]][6,])),na.rm =
86   TRUE))
87 (marketShareRevenus <- ownRevenus/(ownRevenus+compRevenus))
88
89 arrivalSales <- as.character(simulResults[[1]][1,simulResults
90   [[1]][6,]=1])
91 distanceSales <- as.numeric(simulResults[[1]][2,simulResults
92   [[1]][6,]=1])
93 weightSales <- as.numeric(simulResults[[1]][3,simulResults
94   [[1]][6,]=1])
95
96 arrivalComp <- as.character(simulResults[[1]][1,simulResults
97   [[1]][6,]=0])
98 distanceComp <- as.numeric(simulResults[[1]][2,simulResults
99   [[1]][6,]=0])

```

```

82 weightComp <- as.numeric(simulResults[[1]][3, simulResults
    [[1]][6,] == 0])
83
84 # Representation of the result
85 table(arrivalSales)
86 mean(distanceSales)
87 table(arrivalComp)
88 mean(distanceComp)
89 par(mfrow = c(1,1))
90 hist(weightSales, freq = FALSE, breaks = 100, xlim = c(0,15), main =
91       "Sales vs Theoretical Weights Distribution", xlab = "
           weight (Kg)")
92 curve(do.call(eval(parse(text = paste("d", law[match(distChoice,
           distName)], sep = " "))),
93        c(list(x), as.vector(paramAdjust))), add = TRUE, lwd =
           2)
94 abline(v = v <- exp(paramAdjust[1] + paramAdjust[2]**2/2), lwd = 2)
95 text(v+0.75, 0.3, as.character(round(v, 2)))
96 abline(v = v <- mean(weightSales), col = "red", lwd = 2)
97 text(v - 0.75, 0.3, round(v, 2), col = "red")

```

Annexe B

Contribution au projet *OpenFlights*

Cette annexe présente les développements qui ont menés à notre contribution au projet *OpenFlights*.

Comme mentionné à la [sous-section 2.1.2](#), un traitement a été nécessaire afin de repopuler la variable `tzFormat` contenant les informations sur le fuseau horaire des différents aéroports. Pour ce faire, deux sources externes ont été comparées avec les données actuelles pour déterminer la manière la plus précise de repopuler cette information pour parvenir à la publication d'un jeu de données corrigé. Nous nous attarderons pas sur le procédé utilisé, mais davantage sur les résultats obtenus. Vous pourrez au besoin vous référer au [Code Source B.1](#) contenant tous les traitements qui ont menés à la contribution.

Tout d'abord, les deux sources externes utilisées sont :

- ▶ [tz_world](#)
- ▶ [timezone-boundary-builder](#)

La principale différence entre ces dernières est que *timezone-boundary-builder* utilise *OpenStreetMap* afin d'inclure les eaux territoriales dans la définition des bornes de fuseaux horaires. De plus, il y a une mention sur le site de *tz_world* que l'information n'a pas été mise à jour depuis le 28 mai 2016.

Outre ces considérations, les deux sources sont construites sensiblement selon le même format. Il s'agit de deux *ShapeFile* à partir desquels nous irons extraire l'identifiant du fuseau horaire nommé TZID.

Initialement, le jeu de données d'*OpenFlights* contenait 593 aéroports sans information sur le fuseau horaire. Nous chercherons bien évidemment à réduire le plus possible cette proportion. Sur ce point, le *ShapeFile* de *timezone-boundary-builder* performera beaucoup mieux en réduisant le nombre de valeurs manquantes à 7 comparativement à *tz_world* qui en contiendra toujours 250.

Source 1	Source 2	% Concordance
<i>OpenFlights</i>	<i>tz_world</i>	87.5 %
<i>OpenFlights</i>	<i>timezone-boundary-builder</i>	87.6 %
<i>tz_world</i>	<i>timezone-boundary-builder</i>	99.7 %

TABLE B.1 – Étude comparative de concordance entre les différentes sources de fuseaux horaires

Par contre, la réduction du nombre de valeurs manquantes n'est pas un critère suffisant pour discriminer une source par rapport à l'autre. Encore faut-il que ces nouvelles données soit précises ! Pour ce faire, nous avons mener une étude comparative par rapport aux valeurs qui étaient déjà présentes dans le jeu de données d'*OpenFlights*. Le principe de cette étude consistait à regarder dans quelle proportion des cas, les valeurs connues ou extraites concordent si les deux valeurs ne sont pas manquantes. La [Tableau B.1](#) présente ces différentes proportions.

À partir de ces résultats, il sera possible de conclure que les sources *tz_world* et *timezone-boundary-builder* sont plus fiables que l'information actuellement contenue dans la variable **tzFormat**. Il ne reste plus qu'à savoir si *timezone-boundary-builder* est réellement une source plus fiable que *tz_world*. Pour ce faire, nous avons étudié tous les cas où le fuseau horaire identifié par ces deux sources n'était pas le même. Parmi ces 22 cas, seulement 2 cas indiquaient un résultat légèrement plus précis pour *tz_world*. D'autre part, nous devons admettre que la variable **tzFormat** s'est révélée plus exacte que nos deux sources seulement pour 1 cas.

Nous en sommes donc venus à conclure que nous devrions n'ont pas seulement compléter les champs manquants de la variable **tzFormat**, mais bien complètement remplacer cette dernière par les informations de *timezone-boundary-builder*.

Certaines considérations sont à prendre en compte :

- Le fuseau horaire *America/Punta_Arenas* n'est pas reconnu comme un format valide par le paquetage **lubridate**. Cependant, la variable **tzFormat** assignait ces aéroports au fuseau horaire *America/Santiago* ce qui pourrait introduire des erreurs dans le calcul du temps de vol. [21]
De plus, ce problème est mineur puisqu'il n'implique que 5 aéroports.
- Aucun travail n'a été fait auprès des autres variables relatives au fuseau horaire. Ces dernières devront faire l'objet d'une extrapolation à partir de la nouvelle information ou tout simplement être supprimée puisque le format contient déjà toute l'information nécessaire.

Tout le contenu de la contribution est disponible directement sur la page suivante : <https://github.com/jpatokal/openflights/pull/736>.

Code Source B.1 – tzFormatRefill.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 (path <- paste(getwd(),"../..",sep = "/"))
15
16 # Extraction of airports.dat
17 airports <- read.csv("https://raw.githubusercontent.com/jpatokal/
   openflights/master/data/airports.dat",
18                     header = FALSE, na.strings=c("\\N",""))
19 colnames(airports) <- c("airportID", "name", "city", "country", "
   IATA", "ICAO",
20                        "latitude", "longitude", "altitude", "
   timezone", "DST",
21                        "tzFormat", "typeAirport", "Source")
22
23
24 # Initial Proportion of missing tzFormat
25 length(airports$tzFormat[is.na(airports$tzFormat)])/length(airports
   $tzFormat)
26
27 # We use ShapeFile of TZ_world to fill the missing values
28 # install.packages("sp")
29 # install.packages("rgdal")
30 library(sp)
31 library(rgdal)
32 obs <- subset(airports, select = c("airportID", "name", "longitude", "
   latitude"))
33 sppts <- SpatialPoints(subset(obs, select = c("longitude", "latitude"
   )))
34 proj4string(sppts) <- CRS("+proj=longlat")
35
36 tz_world_1.shape <- readOGR(dsn=paste(path, "/ref/tz_world_1", sep="")
   ), layer="tz_world")
37 sppts <- spTransform(sppts, proj4string(tz_world_1.shape))
38 merged_tz_1 <- cbind(obs, over(sppts, tz_world_1.shape))
39 sum(merged_tz_1$TZID == "uninhabited", na.rm = TRUE)
40 subset(merged_tz_1, TZID == "uninhabited")
41 is.na(merged_tz_1) <- merged_tz_1 == "uninhabited"
42 sum(merged_tz_1$TZID == "uninhabited", na.rm = TRUE)
43
44 tz_world_2.shape <- readOGR(dsn=paste(path, "/ref/tz_world_2", sep="")
   ), layer="combined_shapefile")
45 sppts <- spTransform(sppts, proj4string(tz_world_2.shape))
46 merged_tz_2 <- cbind(obs, over(sppts, tz_world_2.shape))
47
48 # install.packages("sqldf")
49 library(sqldf)

```



```

50 airports <- sqldf("select
51     a.*,
52     b.TZID as tzMerged_1,
53     c.TZID as tzMerged_2
54 from airports a
55 left join merged_tz_1 b
56 on a.airportID = b.airportID
57 left join merged_tz_2 c
58 on a.airportID = c.airportID
59 order by a.airportID")
60 airports <- as.data.frame(as.matrix(airports))
61 summary(airports)
62 names(airports)
63
64 sum(is.na(airports$tzMerged_1))
65 sum(is.na(airports$tzMerged_2))
66
67 # Verification with available time zones
68 # Test 1
69 test1 <- subset(airports, !is.na(tzFormat) & !is.na(tzMerged_1))
70 sum(paste(test1$tzFormat) == paste(test1$tzMerged_1))/length(test1$
    tzFormat)
71 # Test 2
72 test2 <- subset(airports, !is.na(tzFormat) & !is.na(tzMerged_2))
73 sum(paste(test2$tzFormat) == paste(test2$tzMerged_2))/length(test2$
    tzFormat)
74 # Test 3
75 test3 <- subset(airports, !is.na(tzMerged_1) & !is.na(tzMerged_2))
76 sum(paste(test3$tzMerged_1) == paste(test3$tzMerged_2))/length(
    test3$tzFormat)
77
78 errors1 <- subset(airports, (paste(tzFormat) != paste(tzMerged_1) &
    !is.na(tzFormat) & !is.na(tzMerged_1)))
79 errors2 <- subset(airports, (paste(tzFormat) != paste(tzMerged_2) &
    !is.na(tzFormat) & !is.na(tzMerged_2)))
80 errorsTot <- subset(airports, (paste(tzFormat) != paste(tzMerged_1)
    & !is.na(tzFormat) & !is.na(tzMerged_1)) | (paste(tzFormat) !=
    paste(tzMerged_2) & !is.na(tzFormat) & !is.na(tzMerged_2)))
81
82 # Export of the errors into a report
83 # install.packages("knitr")
84 library(knitr)
85 mdErrorsTable <- kable(subset(errorsTot, select = c("airportID", "
    name", "IATA", "tzFormat", "tzMerged_1", "tzMerged_2")), format = "
    markdown")
86 knit("errors", text = mdErrorsTable, "../valid/errors.md")
87
88 # install.packages("lubridate")
89 library(lubridate)
90 x <- Sys.time()
91 mean(totaldiff1 <- sapply(1:length(errors1$tzFormat), function(i)
    difftime(force_tz(x, paste(errors1$tzMerged_1[i]), force_tz(x,
    paste(errors1$tzFormat[i])))))
92 mean(totaldiff2 <- sapply(1:length(errors2$tzFormat), function(i)
    difftime(force_tz(x, paste(errors2$tzMerged_2[i]), force_tz(x,
    paste(errors2$tzFormat[i])))))
93

```

```

94 couple <- unique(cbind(paste(errorsTot$tzFormat), paste(errorsTot$
    tzMerged_1), paste(errorsTot$tzMerged_2)))
95 mean(coupledifff1 <- sapply(1:nrow(couple), function(i) difftime(
    force_tz(x, couple[i,1]), force_tz(x, couple[i,2]))))
96 mean(coupledifff2 <- sapply(1:nrow(couple), function(i) difftime(
    force_tz(x, couple[i,1]), force_tz(x, couple[i,3]))))
97 couplediffTot <- cbind(coupledifff1, coupledifff2)
98
99 toValid <- subset(airports, paste(tzMerged_1) != paste(tzMerged_2) &
    !is.na(tzMerged_1) & !is.na(tzMerged_2))
100 mdValidTable <- kable(subset(toValid, select = c("airportID", "name",
    "IATA", "tzFormat", "tzMerged_1", "tzMerged_2")), format = "
    markdown")
101 knit("valid", text = mdValidTable, "../valid/valid.md")
102
103 # install.packages("rmarkdown")
104 # library(rmarkdown)
105 # render(input = mdErrorsTable, output_file = "file", output_dir =
    ".")
106 # install.packages("markdown")
107 library(markdown)
108 markdownToHTML("../valid/errors.md", "../valid/errors.html", encoding
    = "utf8")
109 markdownToHTML("../valid/valid.md", "../valid/valid.html", encoding =
    "utf8")
110
111 airports <- subset(airports, select = -c(tzFormat, tzMerged_1))
112 summary(airports)
113
114 # install.packages("dplyr")
115 library(dplyr)
116 airports <- plyr::rename(airports, c("tzMerged_2" = "tzFormat"))
117
118 # Final Proportion of missing tzFormat
119 length(airports$tzFormat[is.na(airports$tzFormat)]) / length(airports
    $tzFormat)
120
121 # Export final database
122 summary(airports)
123 write.table(airports, file = "../data/airports_Updated.dat", row.
    names = FALSE, col.names = FALSE)

```

Annexe C

Premiers pas avec R

Cette annexe est destinée à fournir des indications sur comment installer R sur votre poste de travail ainsi que les outils qui faciliteront vos débuts avec ce langage de programmation.

C.1 Installation de R

Le première étape consiste à l'installation de R. Pour ce faire, rendez-vous sur *Comprehensive R Archive Network* ([CRAN](#)). Cliquer ensuite sur le lien correspondant à votre système d'exploitation.¹ Il suffira ensuite de suivre les instructions fournies qui varieront selon votre système d'exploitation. Ces dernières sont résumées ci-dessous.

C.1.1 *Windows*

Pour les utilisateurs *Windows*, il suffira de cliquer sur *install R for the first time* et ensuite *Download R x.y.z for Windows*. Une fois l'*executable* (**.exe**) téléchargé, il vous suffira de l'exécuter et de suivre les instructions d'installation.

Nous vous recommandons de modifier le chemin d'installation par défaut (`C:\Program Files\R\R-x.y.z`) pour `C:\R\R-x.y.z`. Les raisons de cette modification sont :

- Éviter de devoir toujours exécuter R avec les privilèges administrateur afin d'installer des paquetages dans la librairie principale.
- Éviter les problèmes qui peuvent provenir de la présence d'espace dans les chemins d'installation.

C.1.2 *(MAC) OS X*

Si vous opérer un *MAC*, vous n'aurez qu'à télécharger et exécuter le *package* (`R-x.y.z.pkg`) de la version la plus récente.

1. Il est fortement recommandé de télécharger une distribution binaire précompilée. Cette annexe ne traitera que de cette manière de procéder.

C.1.3 *Linux or Unix-Like*

R est présentement disponible pour les distributions suivantes :

- Debian
- RedHat
- SUSE
- Ubuntu

Choisissez votre distribution et suivez les instructions indiquées. Si vous êtes sur une distribution qui n'est pas supportée, il sera nécessaire de compiler R à partir du code source. Cette procédure est décrite à la section 2.5.1 de la *Frequently Asked Question list* ([R FAQ](#)).

C.2 Installation de R Studio

Une fois le programme R installer, bien qu'il sera possible d'utiliser R via son interface standard, il sera préférable d'utiliser un *Integrated Development Environment* (IDE) pour créer des scripts qui pourront facilement être conservés et/ou partagés. Si vous êtes peu habitués à ce genre d'outils, il sera conseillé d'utiliser R Studio. R Studio est un logiciel libre et gratuit. Rendez-vous directement sur le site de R Studio pour procéder à son installation.

<https://www.rstudio.com/products/rstudio/download2/>

Prendre note que ceux qui seraient déjà familiers avec un autre éditeur, tel que Emacs, peuvent très bien conserver leurs habitudes et éviter l'installation d'un IDE spécifique pour chaque langage de programmation.

C.3 Installation de paquets

L'installation de paquets R supplémentaires à ceux offerts par défaut pourra se faire de plusieurs manières différentes. Les deux premières méthodes sont très manuelles et ne sont présentées qu'à titre informatif. Il sera beaucoup plus efficient d'utiliser la troisième méthode.

Lorsque nous sommes dans l'interface standard de R, il suffira de cliquer sur l'onglet *Packages* et d'ensuite sélectionner *Install package(s)*.... Une fenêtre vous demandera ensuite de choisir un *mirror*. Nous conseillons de prendre *0-Cloud [https]*. Une autre fenêtre s'ouvrira enfin et vous pourrez choisir le paquetage désiré en trouvant son nom dans la liste.

De manière similaire, sous R Studio, vous pourrez aller à l'onglet *Tools* et sélectionner *Install Packages*.... Une fenêtre apparaîtra et vous pourrez procéder à une recherche du paquetage désiré.

Finalement, vous pourrez toujours installer des paquets directement via la console avec la commande R `install.packages("<nom paquetage>")`.

Une fois les paquets installés, n’oubliez pas d’utiliser la commande `library(<nom paquetage>)` avant toute utilisation de fonctions rendues accessibles sous ces derniers.²

C.4 Mise à jour de R et ses paquets

paquetage `installr`

C.5 Installation de Git

Tout développement devrait être fait à l’aide d’un gestionnaire de versions. Ce sujet est très large et hors d’application du document. Nous nous contentons de donner des liens vers des références intéressantes à ce sujet.

compte privés version étudiante <https://github.com/> <https://git-scm.com/download>
<https://www.youtube.com/playlist?list=PL-osiE80TeTuRUfjRe54Eea17-YfnOOAx> tutoriels youtube très intéressants

2. L’installation d’un paquetage donné ne devra se faire qu’une seule fois lors de sa première utilisation.