

ACTULAB

Résolution de problématique de  
Co-Operators

Où sont les clients que nous ciblons ?

Présenter par  
David Beauchemin

© 2017 David Beauchemin



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 :

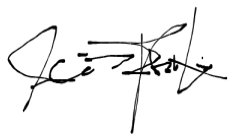


**Attribution** — Vous devez créditer l’œuvre, intégrer un lien vers le contrat et indiquer si des modifications ont été effectuées à l’œuvre. Vous devez indiquer ces informations par tous les moyens possibles, mais vous ne pouvez suggérer que l’offrant vous soutient ou soutient la façon dont vous avez 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 mêmes conditions, c’est-à-dire avec le même contrat avec lequel l’œuvre originale a été diffusée.

## Remerciements

A handwritten signature in black ink, appearing to be 'J. P. R.' with a stylized flourish at the end.

# Table des matières

# Rapport sommaire

## Détails techniques sommaires

### Analyse du mandat

Déterminer la distribution d'une ou plusieurs variables sur un territoire. Prédire la distribution des personas sur un territoire à l'aide de la distribution des variables.

### Collecte des données

Voici la liste des différentes données et rapports ayant été utilisés lors de la réalisation du rapport :

- Données libres du recensement canadien de [2016](#)
- Listes des études utilisées pour les hypothèses :
  - Bulletin d'information statistique du ministère de la famille du Québec : [Quelle famille ?](#) ;
  - Enquête nationale auprès des ménages [2011](#) ;
  - Statistiques de l'enseignement supérieur [2013](#).
- Polygones des régions de tri d'acheminement [nationaux](#)
- Listes des régions de tri [d'acheminement](#)

### Hypothèse utilisée

Hypothèses utilisées pour le projet :

1. Indépendance entre les variables
2. Distribution uniforme sur les intervalles d'âges
3. Distribution uniforme des travailleurs (provincial)
4. Distribution uniforme des étudiants (provincial)
5. Distribution uniforme des colocataires (provincial)
6. Retraite à 65 ans

## Application *Shiny*

Afin de visualiser la distribution des personas et de permettre une flexibilité d'analyse future, une application *Shiny* à été développée. L'application [personasIdentificateur](#) permet de :

- sélectionner les différentes variables à modéliser ;
- de modéliser visuellement la densité de la distribution ;
- d'afficher les territoires observés ;
- d'afficher, par région de tri d'acheminement, la population totale et la prédiction du nombre de personas identifier dans cette région.



### *Utilisation*

Pour utiliser l'application, il suffit d'accéder à la page web et de sélectionner les variables désirées. Pour afficher les informations de la région de tri d'acheminement, il suffit de glisser le curseur sur celle-ci.

Le délai d'exécution peut parfois prendre quelques secondes.

## Résultat sommaire

.....

## Code source

Il est possible de consulter l'ensemble du code source et les données utilisées pour le projet à partir de la [page web](#) du dépôt [GitHub](#).

# Rapport détaillé

## Détails de la problématique

Étant une coopérative d'assurance, Co-operators utilise un réseau de distribution à l'aide de différentes agences dispersées à l'échelle nationale. Celles-ci offrent un ensemble de produits d'assurances vie et d'assurances dommages.

Afin d'offrir un support amélioré aux différentes agences d'assurances de Co-operators, la définition de certaines caractéristiques de profil type est extraite des données clients. Afin d'identifier la modélisation des variables, l'utilisation des données libres est utilisée pour construire et représenter la distribution.

## Détails techniques

### Mandat

Le mandat de cette problématique Actulab correspond à déterminer la distribution des différentes variables observées des personas en plus de toutes autres variables jugées pertinente. Cette distribution doit pouvoir être représenter sur un territoire avec une échelle de région de tri d'acheminement (RTA).

Prédire la distribution des personas sur un territoire avec une échelle de RTA à l'aide de la distribution des variables du personas. La modélisation des personas et de leurs variables doit provenir de données en source libre et gratuite.

La modélisation des personas permettra d'offrir une personnalisation des produits, une meilleur approche des clients ainsi que de mieux cibler le style de communication entre Co-operators et les personas.

Le projet sera remis à Co-operators dans le cadre du projet Actulab de l'automne 2017.

## Collecte des données

Depuis l'essor du *Big Data* et du *deep learning*, de nombreuses bases de données sont devenues disponibles gratuitement en ligne. Ce mouvement de *démocratisation* des données a aussi été incorporé auprès de nombreuses instances gouvernementales. À cet effet, Statistique Canada, offre l'ensemble des jeux de données des recensements depuis 1991. De plus, de nombreux rapports et jeux de données provinciales sont aussi disponibles. Cet accès à un nombre important de données a permis de résoudre la problématique abordée.

Parmi les données utilisées on retrouve le recensement canadien [2016](#). Le découpage des données est déjà été segmenté en RTA par Statistique Canada. Cette base de données contient de nombreuses variables essentielles à l'analyse. On retrouve en outre, la population par RTA, le nombre de femmes et d'hommes par RTA, le nombre de personnes par tranches de salaire et beaucoup d'autres encore.

Il faut aussi noter que pour des fins de simplification, seulement la province du Québec a été étudiée. De plus, une liste réduite de villes a été sélectionnée pour représenter cartographiquement les variables sur le territoire. De plus, afin de modéliser la distribution des variables non segmentées par RTA, différentes hypothèses ont été retenues. Celles-ci seront discutées plus loin.

Pour soutenir la crédibilité des hypothèses, des publications et des études de différents départements nationaux ou provinciaux ont été étudiées pour en dégager les grandes tendances.

Tout d'abord, afin de mieux représenter la réalité familiale dans la province du Québec le bulletin d'information statistique du ministère de la famille a été utilisé. Le bulletin *Quelle famille ?* a permis d'établir le comportement général de la population à l'égard de la colocation. On remarque une plus forte proportion de colocation pour les personnes de plus de 70 ans. C'est la même proportion de la population habitant en colocation a été utilisée pour déterminer la probabilité d'un résident d'habiter en colocation. Il est à noter que ces données ne sont pas segmentées par RTA.

De plus, l'enquête nationale auprès des ménages de [2011](#) a permis d'établir les hypothèses sur les professions des Québécois. Ces différentes données sur l'emploi ont permis d'établir la probabilité d'un résident d'occuper une profession quelconque. Le jeu de données ne contient pas le nombre d'étudiant étant donné que les étudiants ne font pas partie de la population active. Afin de déterminer la proportion de la population québécoise étudiante, le rapport des statistiques de l'enseignement supérieur de [2013](#) a été utilisé. Celui-ci a permis d'établir la proportion d'étudiant dans chacun des différents établissements collégiaux et universitaires, autrement dit, la région métropolitaine de résidence. De plus, il a permis d'établir la probabilité d'être un étudiant. Il est à noter que ces données ne sont pas segmentées par RTA.



Finalement, afin de représenter cartographiquement les prédictions et les variables, les données sur les polygones des régions de tri d'acheminement **nationaux** a été utiliser. Il s'agit de la correspondance cartographique des RTA sous forme d'un polynome à multiple côté. Ainsi, il a été possible de représenter dans l'application la distribution. Afin de déterminer les codes postaux des villes étudier, la listes des région de tri **d'acheminement** a été utiliser. Il s'agit essentiellement des codes postaux par ville ou arrondissement.

## Traitement des données

Avant de débiter la manipulation des données, l'analyse des caractéristiques des personas a permis de ressortir les différents besoins en données ainsi que des possibles manipulations à effectuer.

### Caractéristiques des personas

On note les caractéristiques suivantes pour les différents profil type :

#### Persona 1

Âge : entre 16 et 26 ans - (Génération *millennials*) ;  
Salaire annuel : 20 000 \$  
Occupation : Étudiant  
Style de vie : Célibataire, en colocation et numérique

#### Persona 2

Âge : entre 32 et 39 ans - (fin de la génération *x* et début de la génération *millennials*) ;  
Occupation : Professionnel  
Salaire annuel : 51 000 \$  
Style de vie : conjoint de fait/marié, propriétaire et parents

#### Persona 3

Âge : entre 40 et 52 ans - (Génération *x*) ;  
Occupation : Travailleur autonome - propriétaire d'une PME  
Salaire annuel : 105 000 \$  
Style de vie : conjoint de fait/marié, propriétaire et parents

#### Persona 4

Âge : entre 53 et 65 ans - (Génération *baby-boom*) ;  
Occupation : Professionnel, semi-retraite

Salaire annuel : 50 000 \$  
Style de vie : marié, propriétaire et parents mature

### Persona 5

Âge : entre 66 et 76 an - (Génération *baby-boom*);  
Occupation : Professionnel retraité  
Salaire annuel : 43 000 \$  
Style de vie : marié, propriétaire ou locataire dans une résidence et grand-parents

Pour l'ensemble des personas, la caractéristique d'âge, de statut civil et de type d'occupation peuvent facilement être extraite du recensement. Par contre, pour les autres variables tel que des hypothèses et de la simplification a été nécessaire pour obtenir l'information.

En ayant les différentes caractéristiques nécessaire à l'analyse, la segmentation des données d'âges, de salaire, de profession et autres a été effectué. Afin de faciliter l'utilisation et de diminuer la charge sur l'application, les données découpées à partir des différents jeux de données a été enregistrer dans un nouveau jeu de donnée.

## Objectifs

L'objectifs de ce projet est de répondre aux mandats de Co-operators, soit d'offrir une analyse prédictive de la distribution des personas et d'offrir une application permettant d'observer cette distribution sur le territoire.

## Outils utilisés

Les différents outils numériques utilisées lors de la résolution de cette problématique sont l'utilisation de statistique R. Ayant une solide formation informatique avec R, l'utilisation de ce langage paraît plutôt évidente. Par contre, on note différent avantages notablse de R par rapport aux différentes alternatives tel que Excel/VBA ou Python. Il s'agit de la possibilité de créé une application *Shiny*. Une application *Shiny* correspond à une application écrite en R qui utilise un serveur pour effectuer les calculs, en étant disponible sur une page web. Cet avantage a permis d'écrire l'application d'analyse statistique disponible avec ce projet. De plus, la grande communauté R et la quantité importante de *package* à aussi permis de traiter les fichiers de polygone cartographique et l'affichage d'une carte dynamique.

## Méthodologie

Le principal enjeu de ce projet est la validité de la prédiction de la distribution des caractéristiques. Afin de bien représenter celles-ci sur le territoire certaines hypothèses ont été nécessaires afin de simplifier le modèle et de permettre la réalisation du mandat.

### Indépendance des variables

Tout d'abord, une hypothèse essentielle est l'indépendance entre les variables. Il s'agit d'une hypothèse courante en modélisation et celle-ci permet de simplifier le modèle. Grâce à cette hypothèse, l'ajout de variable dans le modèle a été beaucoup plus facile à réaliser. En effet, pour prédire le nombre de persona correspondant à une série de caractéristique, on applique le concept suivant :

$$f_{\mathbb{X}}(x) = f_{X_1}(x_1) \times \dots \times f_{X_n}(x_n)$$

Où  $\mathbb{X}$  est un vecteur de  $n$  variables et  $X_i$  est la variables caractéristiques  $i$ .

Ainsi, si on recherche le nombre d'étudiant de 16 à 26 ans habitant en colocation, on utilise

$$f_{\mathbb{X}}(x) = f_{X_{\text{Occupation}}}(x) \times f_{X_{\text{Age}}}(x) \times f_{X_{\text{Colocation}}}(x)$$

Il devient ainsi beaucoup plus simple de modéliser la fonction conjointe des différentes variables.

Étant donné la segmentation originale du recensement, il est possible de valider l'hypothèse. En effet, à l'aide du test de khi-deux, il est possible de vérifier l'hypothèse entre les variables. Par exemple, si on teste avec la population d'homme âgé entre 16 et 26 ans on obtient la *p-value* suivante :

```
chisqTest$p.value
## [1] 0.4189084
```

Autrement dit, le test est non significatif et ne permet pas de rejeter l'hypothèse d'indépendance entre les variables.

### Distribution uniforme sur les intervalles d'âges

De plus, une seconde hypothèse importante est la distribution sur les intervalles d'âges. Il a été supposé que sur un intervalle d'âge de  $x$  ans à  $x+c$  ans, la répartition est distribuée uniformément. Autrement dit, la probabilité d'avoir  $x$  ans est la même que d'avoir  $x+c$  ans. Cette hypothèse permet d'effectuer une interpolation sur la

segmentation du recensement. Les intervalles d'âge n'étant pas ceux recherchés, il a ainsi été possible de construire les intervalles d'âges des personnes. Cette hypothèse n'a pas été mathématiquement testée mais en considérant une certaine constante dans le taux de natalité, hormis le baby-boom d'après guerre, il est convenable de dire que la distribution est uniforme.

### **Distribution uniforme des travailleurs**

Les trois prochaines hypothèses ont été les plus difficiles à établir et à valider. Elles sont importantes dans l'analyse. N'ayant pu de données segmenter sur les travailleurs, il a été difficile de bien capter l'information. En effet, depuis la simplification du formulaire de recensement, le domaine d'occupation n'est pas inclus dans le recensement de 2011 et 2016. Par contre, l'enquête nationale auprès des ménages contient l'information à l'échelle provinciale. Étant donné que l'information a été segmentée en 11 secteurs d'occupation, l'hypothèse suivante semblait adéquate. Soit que la distribution des 11 catégories d'occupation est uniformément distribuée dans la province. Autrement dit, la probabilité de retrouver un gestionnaire dans la RTA X1X est la même que dans la RTA X2X. Cela a ainsi permis de capter une partie de l'information sur l'occupation professionnelle des personnes.

### **Distribution uniforme des étudiants**

Toujours dans l'occupation des personnes, la distribution de la caractéristique étudiante n'est pas incluse dans le recensement. En effet, les étudiants ne sont pas compris dans la population active, il devient donc difficile de segmenter l'information jusqu'à la RTA. Pour trouver la proportion des étudiants pour l'ensemble du Québec, la contribution du rapport sur les statistiques de l'enseignement supérieur fut nécessaire. Celui-ci a permis d'établir le nombre d'étudiant qui fréquente un établissement d'enseignement supérieur. En consultant les chiffres on constate qu'environ 50 % de la population étudiante fréquente un établissement collégial de Montréal et de 60 % pour un établissement universitaire. Il devient donc difficile de déterminer la distribution des étudiants en périphérie des établissements scolaires. En supposant que l'étalement urbain autour de cet établissement est exponentiellement décroissant selon la facilité d'accès via le transport en commun et le réseau routier normal. On peut supposer une certaine distribution uniforme des étudiants dans les régions avoisinantes. Il a été ainsi conclu que l'hypothèse de distribution uniforme des étudiants sur le territoire est plausible et sera utilisée pour le projet. L'analyse d'une piste de solution sur cette variable est abordée plus loin.

### **Distribution uniforme des colocataires**

Pour cette dernière hypothèse de distribution, pour les mêmes raisons de disponibilité de l'information ainsi que la difficulté de segmentation. Il était difficile de bien modéliser la variable. Toutefois, à l'aide du bulletin du ministère de la famille, le comportement des québécois en matière de colocation a été étudié. L'étude fait

état d'une variation plus ou moins grande selon l'intervalle d'âge et la région. Étant donné que les intervalles d'âge sont relativement large la distribution uniforme sur la colocation semblait approprié et adéquate. Cette hypothèse a donc été retenue, ce qui permet de simplifier grandement le modèle.

## Retraite à 65 ans

Afin de faciliter le modèle, l'hypothèse de retraite à la même âge et pour tous a été retenue. En effet, il était difficile de bien modéliser la retraite progressive, la retraite anticipée, la retraite *normale* et la retraite prolongée. Le principal facteur de cette hypothèse a été le facteur temps dans la réalisation du projet. À partir des données disponibles sur la retraite de Retraite Québec, la modélisation de cette variable aurait été plus complète.

Pour conclure, les différentes hypothèses retenues ont permis de construire le modèle de façon simple et cohérente à partir des informations disponibles.

## Avantage-Inconvénient

Tout modèle n'étant pas parfait, des avantages et inconvénients sont à prendre en considération pour chacun d'eux. Le modèle présenté ici ne faisant pas exception à cette règle, celui-ci entraîne de nombreux avantages et de nombreux inconvénients.

Tout d'abord, la simplicité de l'hypothèse d'indépendance permet de facilement et capter une bonne partie de l'information et permet de modéliser les variables. Cette simplicité permet aussi d'ajouter de nombreuses variables aux modèles, il devient ainsi possible de rajouter des variables tels que le bilinguisme, le nombre d'enfants et le type de résidence. Cet avantage est la principale motivation ayant justifié l'hypothèse d'indépendance du modèle.

De plus, en supposant la distribution uniforme de certaines variables, il fut possible de modéliser des variables plus difficile à modéliser. Par exemple, la colocation aurait été beaucoup plus difficile à modéliser dans un modèle sans distribution uniforme.

La simplicité globale du modèle a aussi permis la création d'une application web, ce qui permet à toutes les agences du réseau de consulter en temps réel le modèle. Il ne devient donc plus nécessaire de faire affaire avec un intermédiaire en centre d'expertise. Situation très intéressante dans un contexte d'affaire.

Par contre, l'indépendance entre les variables entraîne une perte d'information lorsque le nombre de variable à modéliser devient important. En effet, il devient difficile de bien *supporter* l'hypothèse d'indépendance entre toutes les variables. Cette simplicité de modèle est intéressante pour les variables déjà présentes dans

les données du recensement. Par contre, lorsqu'il s'agit de représenter des variables n'ayant pas été segmenter par RTA, la qualité du modèle peut ne pas être toujours très satisfaisante.

De plus, l'utilisation de l'application étant intéressante, son interprétation peut être trompeuse et induire en erreur des décisions d'affaires des différentes agences. Pour conclure, dans son ensemble le modèle présenté représente bien la réalité de la distribution des personas. L'amélioration du modèle via les différentes pistes de solution discutées plus loin pourrait amener le modèle à mieux prédire la distribution.

## Résultats

### Piste d'amélioration

Lors de la présentation des hypothèses retenues, la problématique sur la distribution des étudiants dans les RTA a été soulevée. En effet, si on désire améliorer la fiabilité de la prédiction de cette variable, il serait intéressant de créer un modèle plus complexe pour cette variable. En utilisant les données de fréquentations des établissements, de la méthode de transport ainsi que l'établissement d'un zonage en périphérie des établissements. Ce zonage correspondant à la probabilité de retrouver un étudiant dans cette zone. Il serait alors possible d'établir l'étalement urbain des étudiants. Par la suite, en superposant les RTA sur le zonage d'étalement, il serait plus facile d'établir avec précision la distribution des étudiants par RTA.

Une seconde piste d'amélioration serait d'améliorer l'hypothèse d'âge à la retraite. Tel que discuté plus tôt, le facteur temps ayant empêché l'approfondissement de cette variable, il serait pertinent de corriger cette solution. En effet, de nombreuses données complètes existent sur la retraite. Il aurait été possible de déterminer la distribution de l'âge à la retraite et de déterminer, conjointement avec le recensement, la distribution de la retraite anticipée, de la retraite progressive et de la retraite tardive. Ainsi, la qualité de la prédiction de cette variable serait grandement améliorée.

Une dernière piste d'amélioration du modèle concerne l'ajout de variables, soit l'ajout d'une variable de canaux de communication favoris et de la variable méthode de transport. Cet ajout permettrait d'améliorer le produit offert aux agences. En effet, il serait pertinent de modéliser les canaux de communication favoris des personas, soit numérique, téléphonique et postaux. Ainsi, les agences seraient en mesure de mieux cibler la méthode pour contacter les clients potentiels. De plus, la variable de méthode de transport permet d'ajouter la distribution des personas utilisant un véhicule automobile pour se rendre quotidiennement au lieu de travail. Cette variable permettrait de mieux cibler l'offre de produit aux clients potentiels en plus de mieux cibler les offres de rabais de type assurance automobile et habitation. Pour conclure, en appliquant les pistes de solutions, la qualité de la prédiction

du modèle serait grandement améliorer ainsi que la qualité du service offert aux agences. C'est pourquoi, je recommande que l'approfondissement du projet soit considéré pour améliorer la qualité métrique d'estimation et la qualité d'affaire du projet.

## **Conclusion**

Pour conclure, blah blah