

---

# SimGen

## Équipe CREEi

avr. 30, 2021



<b>1</b>	<b>Droits d'utilisation</b>	<b>3</b>
1.1	Installation . . . . .	3
1.2	Utilisation . . . . .	5
1.3	Méthodologie . . . . .	13
1.4	Résultats . . . . .	23
1.5	Dictionnaire (classes et fonctions) . . . . .	26
1.6	Nous joindre . . . . .	30
<b>2</b>	<b>Index</b>	<b>33</b>
<b>3</b>	<b>Documentation en PDF</b>	<b>35</b>
	<b>Index des modules Python</b>	<b>37</b>
	<b>Index</b>	<b>39</b>



SimGen est un modèle de microsimulation effectuant des projections démographiques de long terme pour le Québec (2017 à 2100). Ce modèle prend en compte les transitions démographiques majeures de la fécondité (naissances), de la mortalité (décès) et des migrations (immigration et émigration). SimGen modélise également la scolarité et l'état matrimonial (formation d'unions et séparations).

SimGen peut être utilisé afin de produire des distributions démographiques très utiles dans le cadre de recherches et d'enseignements. Les résultats des simulations permettent notamment d'analyser les conséquences économiques de la taille et de la structure de la population québécoise. SimGen a été développé par l'équipe de la [Chaire de recherche sur les enjeux économiques intergénérationnels](#), une chaire conjointe [ESG UQAM](#) et [HEC Montréal](#) soutenue par le [CIRANO](#), l'[Institut sur la retraite et l'épargne \(IRE\)](#) et [Retraite Québec](#).

Pour rester informé.e des mises à jour de SimGen, inscrivez-vous à notre [liste d'envoi dédiée](#).



# CHAPITRE 1

---

## Droits d'utilisation

---

Le SimGen est fourni sous [licence MIT](#) (« MIT License ») :

A short and simple permissive license with conditions only requiring preservation of copyright and license notices. Licensed works, modifications, and larger works may be distributed under different terms and without source code.

Les conditions de la licence sont les suivantes : Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the « Software »), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions :

The copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED « AS IS », WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

## 1.1 Installation

SimGen est programmé en langage Python. Il est ainsi nécessaire de posséder la version 3.7 de Python ou une version supérieure pour faire fonctionner SimGen sur votre ordinateur. Malgré tout, si vous n'avez pas accès au logiciel Python mais que vous disposez d'un compte Google, il vous sera possible d'utiliser SimGen en accès à distance via *Google Colab*. Ainsi, il est possible d'avoir accès à SimGen selon trois méthodes présentées ci-dessous. Dans tous les cas, veuillez lire les [condition d'utilisation](#) du site internet *pypi* qui héberge le package.

---

### Important

SimGen utilise par défaut la Base de données de simulation de politiques sociales (BDSPS) comme base de données de départ. Cette base de données nécessite toutefois une licence d'utilisation gratuite octroyée sur demande par Statistique Canada.

La BDSPS est disponible par l'entremise de l'[Initiative de démocratisation des données \(IDD\)](#). Les professeurs et étudiants [des établissements postsecondaires participants](#) possèdent ainsi une licence d'utilisation de la BDSPS par l'entremise de leur établissement.

Si **vous possédez une licence**, écrivez à [yann.decarie@hec.ca](mailto:yann.decarie@hec.ca) en fournissant une preuve de licence ou d'appartenance à un établissement postsecondaire participant à l'IDD. Un fichier .csv prêt à l'emploi dans SimGen vous sera ensuite envoyé.

Si **vous ne possédez pas de licence**, vous devez faire une demande de licence pour la BDSPS en écrivant à [statcan.spsdm-bdmtps.statcan@canada.ca](mailto:statcan.spsdm-bdmtps.statcan@canada.ca) (voir également le [site internet](#) de la BDSPS). Lorsque vous aurez obtenu cette licence, il vous suffira d'écrire à [yann.decarie@hec.ca](mailto:yann.decarie@hec.ca) et un fichier .csv prêt à l'emploi dans SimGen vous sera envoyé.

---

### 1.1.1 Installation automatisée

Si Python est installé sur votre ordinateur et que vous avez accès à votre invite de commande, il est possible d'installer SimGen de manière automatisée en écrivant simplement cette commande dans l'invite de commande (terminal ou anaconda prompt) :

```
pip install simgen-creei
```

Par la suite, il est possible d'invoquer SimGen dans un notebook ou un script en tant que module de la manière suivante :

```
import simgen
```

### 1.1.2 Installation manuelle

Si Python est installé sur votre ordinateur, mais que vous ne pouvez utiliser l'invite de commande, il est possible d'installer manuellement SimGen en complétant les étapes suivantes :

1. Allez sur le site internet [Pypi](#) et faites une recherche du package « simgen-creei ».
2. Cliquez sur l'onglet « simgen-x.x.x », où « x.x.x » correspond au numéro de version.
3. Ensuite, cliquez sur « Download files » dans le menu à gauche et puis cliquez sur le nom du fichier « simgen-x.x.x.tar.gz » pour télécharger le fichier compressé.
4. Une fois le fichier téléchargé, décompressez le fichier « simgen-x.x.x.tar.gz » une première fois.
5. Ouvrez le dossier créé par l'extraction (ex. simgen-x.x.x.tar), continuez ensuite en ouvrant le dossier « dist » et décompressez le fichier « simgen-x.x.x.tar ».
6. Une fois le fichier décompressé, transférez le dossier « simgen-x.x.x » qui en résulte dans le dossier où vous entreposez vos packages (si vous n'en avez pas, créez-en un à l'endroit qui vous convient le mieux).
7. Enfin, ajoutez dans votre notebook ou votre script le chemin d'accès de votre dossier de packages et vous pourrez invoquer SimGen en tant que module de la manière suivante :

```
import sys
sys.path.append('../packages')

import simgen
```



### 1.1.3 Accès à distance

Si vous ne possédez pas ou ne pouvez pas avoir accès au logiciel Python, il est possible d'utiliser SimGen par l'entremise de Google Colab. Après avoir accédé à votre compte Google Colab ou en avoir créé un, vous n'avez qu'à utiliser la commande suivante dans un notebook ou un script pour installer SimGen :

```
pip install simgen-creei
```

Par la suite, il est possible d'invoquer SimGen en tant que module de la manière suivante :

```
import simgen
```

## 1.2 Utilisation

Cette section a pour objectif de guider les utilisateurs de SimGen dans l'utilisation de celui-ci. Dans un premier temps, les étapes d'utilisation et les différents choix possibles de paramètres sont présentés. Par la suite, un exemple de notebook/script est proposé afin de servir de point de départ aux utilisateurs pour le lancement de simulations et l'analyse des résultats.

### 1.2.1 Étapes

Lors de la rédaction d'un notebook ou d'un script Python, quatre étapes principales doivent être suivies afin d'obtenir des résultats de simulation avec SimGen :

#### 1. Choix des paramètres d'utilisation

La première étape consiste à choisir les paramètres qui guideront SimGen par rapport à la localisation de la base de données de départ et aux hypothèses de modélisation.

##### Chemin d'accès et nom du fichier de la BDSPS

Un premier paramètre d'utilisation à déterminer est le chemin d'accès et le nom du fichier .csv qui vous aura été fourni par l'équipe de la CREEi et qui correspond à une version épurée de la [Base de données de simulation de politiques sociales \(BDSPS\)](#). En début de notebook/script, il est suggéré de définir un objet qui comprend le chemin d'accès et le nom du fichier de la BDSPS selon l'endroit où vous aurez enregistré le fichier, et le nom que vous lui aurez donné :

```
donnees_brutes = '../bdsp2017_slice.csv'
```

Cet objet servira d'intrant dans la fonction de formatage des données (*bdsp\_format()*) à l'étape d'initialisation du modèle.

##### Année de fin

L'année de fin de la simulation détermine la durée des projections effectuées par SimGen. Les valeurs possibles vont de 2018 à 2100. La valeur minimale de l'année de fin correspond à 2018, puisque l'année de départ est fixée par défaut à 2017. Cette dernière valeur ne peut être choisie par l'utilisateur, puisque la base de données de départ date de 2017 et que l'année de départ doit correspondre à cette valeur.

Il est suggéré de définir, en début de notebook/script, un objet qui comprend l'année de fin choisie comme suit :

```
annee_fin = 2050
```

Cet objet sera utilisé à l'étape 2 lors du chargement des principaux intrants. Il en va de même pour tous les autres paramètres d'utilisation.

### Nombre de réplifications

Le nombre de réplifications détermine le nombre de fois où SimGen simule l'ensemble de l'horizon de temps (année du début à l'année de fin) dans une simulation. Ce paramètre est fixé par défaut à 1. Lorsque le nombre de réplifications est supérieur à 1, les résultats de la simulation correspondent à la moyenne des résultats des réplifications.

Il est suggéré d'utiliser plus d'une réplification afin d'obtenir des résultats uniformes d'une simulation à une autre. Le nombre optimal de réplifications varie selon les résultats utilisés. Les résultats précis comprenant un petit nombre d'observations (ex. : le nombre de personnes en couple de 95 ans) sont plus susceptibles de varier d'une simulation à une autre qu'un résultat global comportant un grand nombre d'observations (ex. : le nombre total de personnes en couple âgées de 15 à 65 ans). Les résultats précis nécessitent donc un plus grand nombre de réplifications pour être stables (50 réplifications) que les résultats globaux (10 réplifications).

Il est suggéré de définir, en début de notebook/script, un objet qui comprend le nombre de réplifications choisi comme suit :

```
nb_rep = 50
```

### Hypothèses

#### Fécondité

Dans SimGen, il est possible de calibrer le nombre de naissances selon trois scénarios de fécondité issus du plus récent [document de projection démographique](#) de l'Institut de la statistique du Québec (ISQ). Les scénarios de fécondité supposent la convergence de l'indice synthétique de fécondité (ISF) vers les valeurs suivantes d'ici 2026, selon le scénario :

Fécondité	ISF
Faible (weak)	1,45
Référence (reference)	1,60
Forte (strong)	1,75

Dans la version actuelle de SimGen, les termes anglais *weak*, *reference*, et *strong* doivent être utilisés comme intrant dans la fonction d'hypothèse de fécondité. Il est suggéré de définir, en début de notebook/script, un objet qui comprend le nom du scénario de fécondité choisi comme suit :

```
fecondite = 'reference'
```

#### Mortalité

Il est possible de fixer les quotients de mortalité selon trois scénarios de mortalité issus du plus récent [document de projection démographique](#) de Statistique Canada. Les scénarios de mortalité supposent l'atteinte de trois valeurs possibles de l'espérance de vie à la naissance selon le genre d'ici 2062 :

Mortalité	Hommes	Femmes
Faible (low)	89,8 ans	92,0 ans
Moyenne (medium)	87,5 ans	89,2 ans
Élevée (high)	85,9 ans	87,3 ans

En 2020, l'espérance de vie à la naissance est estimée à 80,6 ans pour les hommes et 84,0 ans pour les femmes ([ISQ, mars 2021](#)). Dans la version actuelle de SimGen, les termes anglais *weak*, *medium*, et *strong* doivent être utilisés comme intrant dans la fonction d'hypothèse de mortalité. Il est suggéré de définir, en début de notebook/script, un objet qui comprend le nom du scénario de mortalité choisi comme suit :

```
mortalite = 'low'
```

### Immigration

Dans SimGen, le nombre de nouveaux immigrants par année est déterminé par le taux prospectif d'immigration internationale (proportion de nouveaux immigrants par rapport à la population totale). Ce paramètre est fixé par défaut à une valeur de 0,0066. Cette valeur correspond à 55 000 nouveaux immigrants internationaux en 2017 (sur une population totale de 8 302 063), conformément au scénario de référence du plus récent [document de projection démographique](#) de l'ISQ.

Le paramètre de taux prospectif d'immigration internationale peut être fixé à la valeur désirée (supérieure ou égale à 0). Ce paramètre reste toutefois fixe durant toutes les années de la simulation. Cette caractéristique implique une augmentation graduelle du nombre de nouveaux immigrants suivant la croissance de la population totale.

Si l'utilisateur souhaite modifier l'hypothèse d'immigration, il est suggéré de définir, en début de notebook/script, un objet qui comprend le taux d'immigration internationale choisi comme suit :

```
taux_immigration = 0.0066
```

Il est à noter que les caractéristiques des nouveaux immigrants sont celles des immigrants récents (depuis 5 ans ou moins) dans la BDSPS de 2017. Par ailleurs, la migration interprovinciale est prise en compte dans SimGen, mais il n'existe toutefois pas de paramètre d'utilisation par rapport à cet aspect (consultez la section [Émigration](#) pour les détails méthodologiques).

## 2. Initialisation du modèle

L'initialisation du modèle vise à charger en mémoire l'ensemble des informations nécessaires au lancement de la simulation. Cette étape se divise en plusieurs sous-étapes.

### Importation des packages

SimGen utilise certains packages Python standards, qu'il est nécessaire d'importer :

```
import warnings
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
warnings.filterwarnings("ignore")
```

### Importation des fonctions et des classes de SimGen

Il est ensuite nécessaire d'importer SimGen en tant que tel :

```
import simgen
from simgen import model, formating
```

### Formatage données de départ

La fonction `bdsp_format` transforme la BDSPS de Statistique Canada afin de mettre en forme certaines variables et créer les registres des individus (dominants, conjoints et enfants). Cette fonction calibre également les poids des répondants, par âge et sexe, afin de s'arrimer à la population québécoise de 2017, selon l'ISQ. Enfin, cette fonction sauvegarde la base de données de départ en format *pkl* en lui donnant le nom de « startpop » et sauvegarde la banque de données des caractéristiques des nouveaux immigrants en format *pkl* en lui donnant le nom de « imm\_pop ». La commande à utiliser est comme suit :

```
preparation_data=formating()
preparation_data.bdsp_format(donnees_brutes)
```

où *donnees\_brutes* correspond au chemin d'accès et au nom du fichier .csv de la BDSPS dans votre système (ex. : *donnees\_brutes* = ".../bdsps2017\_slice.csv").

**Détails : bdsps\_format()**

**class** *simgen.bdsps* (*file*, *year=2017*, *iprint=False*, *file\_format='.dta'*)

Nettoyage de la BDSPS.

Fonction qui permet de mettre en forme la BDSPS.

**Paramètres**

- **year** (*int*) – année de la base de départ (défaut=2017)
- **iprint** (*boolean*) – switch pour imprimer ou non des outputs intermédiaires de cette fonction (défaut=False)

**Création de l'instance du modèle**

La commande suivante crée un gabarit permettant entre autres de stocker les résultats propres à la simulation selon les paramètres d'utilisation choisis :

```
base = model(stop_yr=annee_fin)
```

où *annee\_fin* correspond à l'année de fin de la simulation (ex. : *annee\_fin* = 2050). Si l'argument *stop\_yr* n'est pas spécifié, SimGen fixe par défaut l'année de fin à 2100.

**Détails : classe model()**

**class** *simgen.model* (*start\_yr=2017*, *stop\_yr=2100*)

Modèle de simulation SimGen.

Cette classe permet de créer une instance d'un modèle de microsimulation.

**Paramètres**

- **start\_yr** (*int*) – année de départ de la simulation (défaut=2017)
- **stop\_yr** (*int*) – dernière année de la simulation (défaut=2100)

Pour être en mesure de lancer une deuxième simulation avec des paramètres d'utilisation différents et de comparer les résultats des deux simulations, vous n'avez qu'à réutiliser cette commande en donnant un nom différent au gabarit :

```
base2 = model(stop_yr=annee_fin)
```

et de suivre les mêmes étapes de programmation que pour le premier gabarit (*base*).

**Chargement des principaux intrants**

Tout d'abord, le chargement de la base de données de départ s'effectue à l'aide de la commande suivante :

```
base.startpop('start_pop')
```

où *start\_pop* est le nom donné par défaut à la base de données de départ à la suite du formattage de la BDSPS. Ce nom exacte doit être utilisé, puisqu'un message d'erreur vous sera envoyé en cas contraire.

**Détails : fonction startpop()**

**class** `simgen.model` (*start\_yr=2017, stop\_yr=2100*)

Modèle de simulation SimGen.

Cette classe permet de créer une instance d'un modèle de microsimulation.

**Paramètres**

— **start\_yr** (*int*) – année de départ de la simulation (défaut=2017)

— **stop\_yr** (*int*) – dernière année de la simulation (défaut=2100)

**startpop** (*file*)

Charger une population de départ.

Fonction membre qui permet de charger une population de départ.

**Paramètres file** (*str*) – nom du fichier contenant la population de départ

Le chargement des hypothèses de la simulation s'effectue ensuite à l'aide des commandes suivantes et des objets définis à l'étape 1 (*taux\_immigration, fecondite, mortalite*) :

```
base.birth_assumptions(scenario=fecondite)
base.dead_assumptions(scenario=mortalite)
base.immig_assumptions(init='imm_pop', num=taux_immigration)
```

où *imm\_pop* correspond à la la banque de données des immigrants récents produite par la fonction *bdsp\_format()*. Ce nom exacte doit être utilisé pour l'argument *init*, puisqu'un message d'erreur vous sera envoyé en cas contraire. Si les arguments *scenario* des fonctions *birth\_assumptions* et *dead\_assumptions* ne sont pas spécifiés, SimGen utilise par défaut le scénario de fécondité de référence (*reference*) et le scénario de mortalité moyenne (*medium*). Pour l'immigration, SimGen fixe par défaut le taux d'immigration internationale à 0,0066, si l'argument *num* n'est pas spécifié.

**Détails : fonction \_assumptions()**

**class** `simgen.model` (*start\_yr=2017, stop\_yr=2100*)

Modèle de simulation SimGen.

Cette classe permet de créer une instance d'un modèle de microsimulation.

**Paramètres**

— **start\_yr** (*int*) – année de départ de la simulation (défaut=2017)

— **stop\_yr** (*int*) – dernière année de la simulation (défaut=2100)

**birth\_assumptions** (*scenario='reference', align=True*)

Hypothèses de fécondité.

Fonction membre qui permet de spécifier les hypothèses de fécondité.

**Paramètres**

— **scenario** (*str*) – Permet de choisir entre les différents scénarios de fécondité produits pas l'ISQ (weak, reference, strong)

— **aling** (*boolean*) – paramètre permettant d'aligner le nombre d'immigrants sur l'ISQ

**dead\_assumptions** (*scenario='medium'*)

Hypothèses de mortalité.

Fonction membre qui permet de spécifier les hypothèses de mortalité.

**Paramètres scenario** (*str*) – Permet de choisir entre les différents scénarios de mortalité produits pas l'STC (low, medium, high)

**immig\_assumptions** (*allow=True, num=0.0066, init=None*)

Hypothèses d'immigration.

Fonction membre qui permet de spécifier les hypothèses d'immigration.

**Paramètres**

- **allow** (*boolean*) – paramètre permettant d’aligner le nombre d’immigrants sur l’ISQ
- **num** (*float*) – immigration totale (nombre); par défaut, scénario de référence de l’ISQ
- **init** (*str*) – nom du fichier contenant la population d’immigrants

### 3. Lancement de la simulation

Le lancement de la simulation s’effectue à l’aide de la fonction suivante :

```
base.simulate(rep=nb_rep)
```

où *nb\_rep* correspond au nombre de réplifications de la simulation (ex. : *nb\_rep* = 50). Si l’argument *rep* n’est pas spécifié, SimGen fixe par défaut le nombre de réplifications à 1.

Il est à noter que cette commande a un temps d’exécution plus élevé que les commandes présentées précédemment. Le temps de simulation croît de manière substantielle avec l’année de fin et le nombre de réplifications.

**Détails : fonction simulate()**

**class** `simgen.model` (*start\_yr=2017, stop\_yr=2100*)

Modèle de simulation SimGen.

Cette classe permet de créer une instance d’un modèle de microsimulation.

**Paramètres**

- **start\_yr** (*int*) – année de départ de la simulation (défaut=2017)
- **stop\_yr** (*int*) – dernière année de la simulation (défaut=2100)

**simulate** (*rep=1*)

Fonction déclenchant le lancement de la simulation.

**Paramètres stratas** (*rep*) – Nombre de réplifications

### 4. Production des résultats

Tout d’abord, le tableau ci-dessous présente la liste des variables pouvant servir lors de l’affichage des résultats de SimGen :

Variable	Nom	Type	Valeurs	Étiquette
Âge	age	Entier	0 à 110	
Genre	male	Booléen	<b>True</b> <b>False</b>	Homme Femme
Statut d'études	insch	Booléen	<b>True</b> <b>False</b>	Aux études Études terminées
Scolarité complétée	educ	Caractères	<i>none</i> <i>des</i> <i>dec</i> <i>uni</i>	Sans diplôme Secondaire Collégial Universitaire (bacc. et supérieur)
Statut conjugal	married	Booléen	<b>True</b> <b>False</b>	En union Célibataire
Nombre d'enfants	nkids	Entier	0 à 3	

Il est possible de produire deux types de résultats : 1) des fréquences et 2) des proportions.

### Fréquences

La fonction `stats.freq()` calcule le nombre d'individus selon le sous-groupe spécifié. Par exemple :

```
population_hommes=base.stats.freq(sub='male==True')
```

Si l'argument `sub` n'est pas spécifié, la fonction renvoie le nombre de personnes dans l'ensemble de la population.

### Détails : fonction stats.freq()

**class** `simgen.statistics(stratas)`

Classe pour créer les statistiques provenant d'une simulation.

Cette classe permet de capturer la distribution de la population par strate durant une simulation. Elle permet ensuite de faire plusieurs tableaux dynamiques à partir de ces distributions.

**Paramètres stratas** (*list of str*) – liste des noms de variables du fichiers de dominants afin de stratifier la population et récolter les fréquences (pondérées)

**freq** (*strata=None, bins=[0], sub=None*)

Fonction de fréquences.

Fonction qui permet, à l'aide de *counts*, de calculer les fréquences pondérées pour une strate donnée. Deux options sont disponibles : l'une, *bins*, permet de modifier les catégories de la strate (par exemple le groupe d'âge), tandis que *sub* permet de définir un critère de sélection particulier pour le calcul des fréquences (en *str*).

#### Paramètres

- **strata** (*str*) – nom de la variable par laquelle on veut découper les données ; ne pas spécifier cette option revient à demander les fréquences totales
- **bins** (*list of int*) – liste de valeurs pour découper les données selon la variable *strata* ; fonctionne seulement avec des variables de types *int* (pas de *str*)
- **sub** (*str*) – condition à respecter pour un sous-échantillon, p.ex. « *age>=18* »

**Renvoie** dataframe avec les fréquences par année (ligne) et valeur de la strate (colonne)

**Type renvoyé** dataframe

### Proportion

La fonction `stats.prop()` calcule pour sa part la proportion de la population respectant les caractéristiques spécifiées. Par exemple :

```
proportion_niveau_scolaire = base.stats.prop('educ', sub="age>=25 and age<=64 and_
↳ insch==False")
```

Si l'argument `sub` n'est pas spécifié, la fonction renvoie la proportion de personnes selon les catégories de la variable spécifiée dans l'ensemble de la population.

#### Détails : fonction `stats.prop()`

**class** `simgen.statistics` (*stratas*)

Classe pour créer les statistiques provenant d'une simulation.

Cette classe permet de capturer la distribution de la population par strate durant une simulation. Elle permet ensuite de faire plusieurs tableaux dynamiques à partir de ces distributions.

**Paramètres *stratas*** (*list of str*) – liste des noms de variables du fichiers de dominants afin de stratifier la population et récolter les fréquences (pondérées)

**prop** (*strata*, *bins*=[0], *sub*=None)

Fonction de proportions.

Fonction qui permet, à l'aide de *counts*, de calculer les proportions pondérées pour une strate donnée. Deux options sont disponibles : l'une, *bins*, permet de modifier les catégories de la strate (par exemple le groupe d'âge), tandis que *sub* permet de définir un critère de sélection particulier pour le calcul des proportions (en str).

##### Paramètres

- **strata** (*str*) – nom de la variable par laquelle on veut découper les données
- **bins** (*list of int*) – liste de valeurs pour découper les données selon la variable strata; fonctionne seulement avec des variables de types int (pas de str)
- **sub** (*str*) – condition à respecter pour un sous-échantillon, p.ex. « age>=18 »

**Renvoie** dataframe avec les proportions par année (ligne) et valeur de la strate (colonne)

**Type renvoyé** dataframe

#### Sauvegarde des données des résultats

Enfin, il est possible de sauvegarder les résultats de la simulation dans un fichier .pkl à l'aide de la commande suivante :

```
base.stats.save('../resultats_simgen')
```

#### Détails : fonction `save()`

**class** `simgen.statistics` (*stratas*)

Classe pour créer les statistiques provenant d'une simulation.

Cette classe permet de capturer la distribution de la population par strate durant une simulation. Elle permet ensuite de faire plusieurs tableaux dynamiques à partir de ces distributions.

**Paramètres *stratas*** (*list of str*) – liste des noms de variables du fichiers de dominants afin de stratifier la population et récolter les fréquences (pondérées)

**save** (*file*)

Fonction pour sauvegarder les fichiers de fréquences.

**Paramètres *file*** (*str*) – nom du fichier de sauvegarde, incluant l'extension pkl (format pickle)




Pour une description complète des classes et des fonctions de SimGen, consultez la page [Dictionnaire \(classes et fonctions\)](#).

## 1.2.2 Exemple

### Simulation de base

Cet exemple de notebook permet de se familiariser avec l'utilisation de Simgen en effectuant une simulation et en présentant des résultats de base.

**Téléchargement du notebook :** Cliquez [ici](#), puis sauvegardez le fichier en format .ipynb.

**Accès au notebook via Google Colab\* :** Cliquez ici : 

— Il est à noter qu'il est nécessaire de posséder un compte Google pour utiliser Google Colab.

## 1.3 Méthodologie

À l'année de départ, la base de données de simulation de SimGen correspond à la [Base de données de simulation de politiques sociales \(BDSPS\)](#) de 2017 de Statistique Canada. Cette base de données est composée d'observations statistiquement représentatives des particuliers canadiens et québécois dans leur contexte familial.

Durant les années ultérieures, SimGen fait évoluer chaque individu selon les transitions suivantes dans l'ordre suivant :

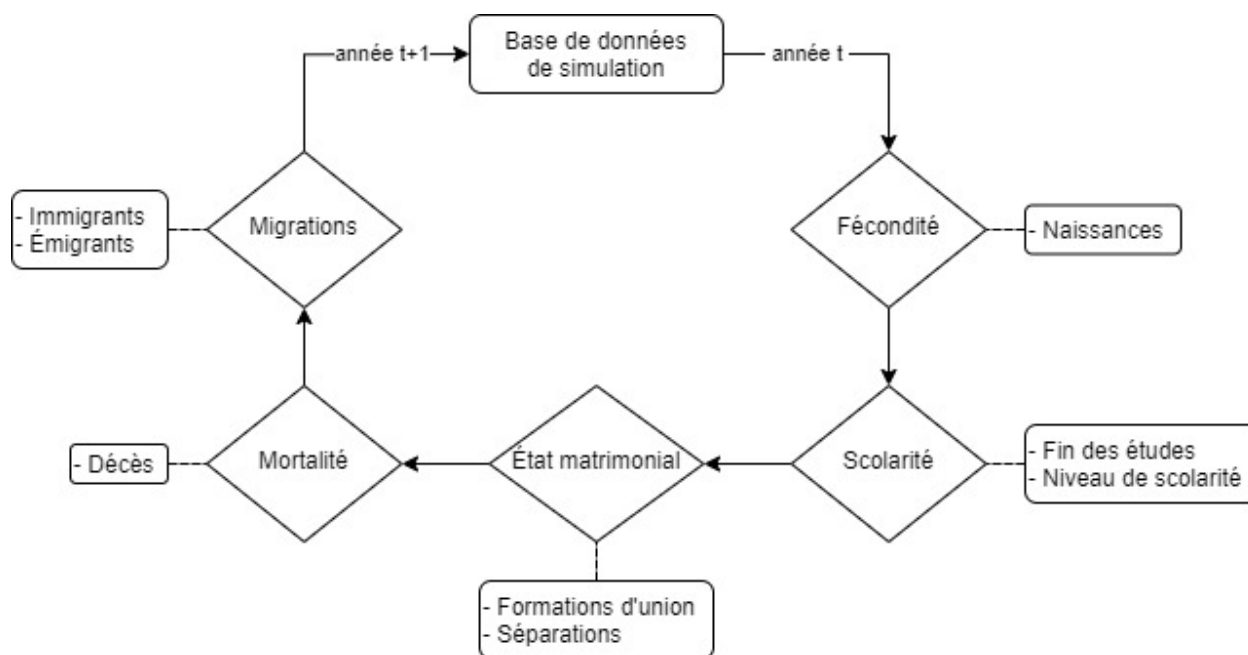
1. Fécondité
2. Scolarité
3. État matrimonial
4. Mortalité
5. Migrations

De nouvelles observations sont ajoutées dans la base de données de simulation lors des transitions de fécondité (naissances) et d'immigration. À l'opposé, certaines observations sont retirées de la base de données lors des transitions de mortalité (décès) et d'émigration. La figure suivante illustre plus explicitement la dynamique des transitions :

Lorsque SimGen arrive à l'année de fin de la simulation, celui-ci recommence le processus de simulation de l'année de départ à l'année de fin, jusqu'à ce que le nombre de réplifications sélectionné soit atteint. Une fois cette étape accomplie, la moyenne des résultats des réplifications est calculée afin de créer la base de données finale des résultats.

### Structure des données

De manière plus spécifique, la base de données de simulation de SimGen est composée de trois registres. Un premier registre contient les individus dominants, un deuxième contient les conjoints de ceux-ci et un troisième contient leurs enfants. Le registre des individus dominants représente l'échantillon principal sur lequel les transitions sont appliquées et sur lequel les sorties statistiques sont basées. Les registres des conjoints et des enfants servent essentiellement à décrire le contexte familial des individus dominants. Ils ne sont donc pas inclus dans les sorties statistiques. Cette approche a été choisie car elle permet de simplifier le processus de simulation et de rendre plus flexible la production de résultats (voir le [document suivant](#) de Statistique Canada pour des explications sur les différentes approches possibles des modèles de microsimulation).



### 1.3.1 Fécondité

#### Modèle économétrique

Pour chaque rang de naissance d'un enfant ( $k=1,2,3$ ), la probabilité d'avoir un enfant est estimée à l'aide d'un modèle logistique incluant trois groupes de variables explicatives liées à l'âge, au niveau de scolarité et à l'âge du dernier enfant, le cas échéant.

$$\mu_{i,t,k} = \mu_{0,k} + \mu_{1,k}age_{i,t} + \mu_{2,k}edu_{i,t} + \mu_{3,k}lkidage_{i,t}$$

$$\Pr(b_{i,t} = 1) = \frac{\exp(\mu_{i,t,k})}{1 + \exp(\mu_{i,t,k})}$$

#### Données et échantillon

Les effets marginaux sont calculés à partir des vagues 2006 et 2011 de l'Enquête sociale générale (ESG) menée auprès des ménages par Statistique Canada.

L'échantillon utilisé pour calculer les 3 régressions logistiques des transitions de naissance est défini en suivant plusieurs étapes :

1. Les données des vagues 2006 et 2011 de l'ESG sont regroupées dans une base unique.
2. L'échantillon est ensuite restreint aux données du Québec (variable *prv*).
3. Un fichier de pseudo panel des répondants qui recense l'historique des transitions de naissances 1, 2 et 3 est créé (calcul des naissances pour chaque année à partir des variables *agechdc1*, *agechdc2* et *agechdc3* correspondant à l'âge des enfants d'ordre 1, 2 et 3).
4. Seul l'historique des transitions du pseudo panel depuis 30 années est conservé afin d'éviter les effets des cohortes les plus anciennes (1976 à 2006 pour l'ESG de 2006 et 1981 à 2011 pour celle de 2011).
5. L'échantillon est finalement restreint aux femmes âgées de 18 à 44 ans inclusivement.

## Variables du modèle

Les variables dépendantes pour les régressions 1, 2 et 3 sont des variables indicatrices, égales à 1 lors de l'année de naissance de l'enfant d'ordre  $k=1,2,3$  et égales à 0 depuis l'année de naissance du dernier enfant (pour les naissances d'ordre 2 et 3) ou depuis 18 ans pour le premier enfant (naissance d'ordre 1).

Prenons l'exemple d'une femme qui a eu deux enfants : un enfant à 20 ans et un enfant à 30 ans. Dans ce cas, la variable dépendante pour le premier enfant sera égale à 0 de 18 ans à 19 ans, puis elle sera égale à 1 à 20 ans. La variable dépendante pour le second enfant sera égale à 0 de 21 ans à 29 ans et elle sera égale à 1 à l'âge de 30 ans.

### Variables explicatives d'âge (variables indicatrices) :

- *age1824* (référence) : la femme a entre 18 et 24 ans.
- *age2529* : la femme a entre 25 et 29 ans.
- *age3034* : la femme a entre 30 et 34 ans.
- *age3539* : la femme a entre 35 et 39 ans.
- *age40p* : la femme a entre 40 et 44 ans.

### Variables explicatives d'éducation (variables indicatrices, notées « edu » dans l'équation du modèle économétrique) :

- *insch* : la femme n'a pas terminé ses études.
- *inf* (référence) : la femme a terminé ses études, mais n'a pas complété ses études secondaires.
- *des* : la femme a terminé ses études et a un diplôme d'études secondaires ou des études partielles à l'université ou au cégep.
- *dec* : la femme a terminé ses études et a un diplôme d'études collégiales.
- *uni* : la femme a terminé ses études et a un diplôme égal ou supérieur au baccalauréat.

### Variable du dernier enfant :

- *lkidage* : âge du dernier enfant né. Cette variable est uniquement utilisée pour les naissances d'ordre 2 et 3.

## Résultats de régression

Les résultats des régressions logistiques sont présentés dans le tableau suivant :

TABLEAU 1 – Logit - Coefficients des transitions de naissances

Variables	1er enfant	2e enfant	3e enfant
<i>age2529</i>	.5111452	.20215	-.1745108
<i>age3034</i>	.0206863	.0036399	-.7896601
<i>age3539</i>	-.8569678	-.8245546	-1.621034
<i>age40p</i>	-1.939295	-2.335598	-3.388722
<i>insch</i>	-1.025785	.1545281	.4842334
<i>des</i>	-.223628	.1972002	.0498862
<i>dec</i>	-.1165148	.1987852	.2902924
<i>uni</i>	-.1596642	.4500001	.8004084
<i>lkidage</i>	0	-.0990092	-.0377307
constant	-2.532843	-1.634364	-2.288871

## Mise en œuvre

La mise en œuvre dans SimGen est réalisée par un tirage uniforme, indépendant par individu dominant. Une naissance survient lorsque le résultat de ce tirage est inférieur à la probabilité logistique prédite. Dans SimGen, les personnes à risque pour cette transition sont les femmes en couple (qu'elles soient enregistrées comme individu dominant ou conjointe) âgées de 18 à 44 ans inclusivement.

### 1.3.2 Scolarité

Tous les enfants débutent leurs études l'année de leurs 5 ans. La présente transition calcule la probabilité qu'un individu finisse ses études. S'il est déterminé que cet individu termine ses études durant l'année en cours, un niveau de scolarité lui est ensuite attribué.

#### Modèle économétrique

Deux régressions logistiques sont réalisées pour 1) calculer la probabilité de finir ses études ; 2) attribuer un niveau de scolarité aux individus qui ont complété leurs études. Une régression logistique dichotomique est appliquée pour calculer la probabilité de finir ses études et un modèle logistique multinomial est utilisé afin d'attribuer le niveau de scolarité correspondant.

- 1) probabilité d pour un individu  $i$  de finir ses études ( $f = 1$ ) à l'année  $t$  :

$$\mu_{i,t} = \mu_0 + \mu_1 age_{i,t} + \mu_2 male_{i,t} + \mu_3 father_{i,t} + \mu_4 mother_{i,t}$$

$$\Pr(f_{i,t} = 1) = \frac{\exp(\mu_{i,t})}{1 + \exp(\mu_{i,t})}$$

- 2) pour chaque niveau de scolarité  $e = 1$  (*n'a pas terminé ses études secondaires*), 2 (*diplôme d'études secondaires*) [référence], 3 (*diplôme d'études collégiales*), 4 (*diplôme égal ou supérieur au baccalauréat*) atteint par un individu  $i$  l'année de terminaison des études en  $t$  :

$$\mu_{e(i,t)} = \mu_0 + \mu_i age_{i,t} + \mu_j male_{i,t} + \mu_k father_{i,t} + \mu_l mother_{i,t}$$

$$\Pr(e_{i,t} = 1) = \frac{\exp(\mu_{1(i,t)})}{1 + \exp(\mu_{1(i,t)}) + \exp(\mu_{3(i,t)}) + \exp(\mu_{4(i,t)})}$$

$$\Pr(e_{i,t} = 2) = \frac{1}{1 + \exp(\mu_{1(i,t)}) + \exp(\mu_{3(i,t)}) + \exp(\mu_{4(i,t)})}$$

$$\Pr(e_{i,t} = 3) = \frac{\exp(\mu_{3(i,t)})}{1 + \exp(\mu_{1(i,t)}) + \exp(\mu_{3(i,t)}) + \exp(\mu_{4(i,t)})}$$

$$\Pr(e_{i,t} = 4) = \frac{\exp(\mu_{4(i,t)})}{1 + \exp(\mu_{1(i,t)}) + \exp(\mu_{3(i,t)}) + \exp(\mu_{4(i,t)})}$$

#### Données et échantillon

Les régressions logistiques sont réalisées à l'aide des vagues 2006 et 2011 de l'Enquête sociale générale (ESG) menée auprès des ménages par Statistique Canada.

L'échantillon utilisé pour calculer les transitions de scolarité est défini en suivant plusieurs étapes :

- 1) Les données des vagues 2006 et 2011 de l'ESG sont regroupées (*pooled*) dans une base unique.
- 2) L'échantillon est restreint aux données du Québec (variable *prv*).
- 3) Un fichier de pseudo panel des répondants qui recense l'historique des transitions de fin d'études et le niveau de scolarité associé est créé.

- 4) Seul l'historique des transitions du pseudo panel depuis 30 années est conservé afin d'éviter les effets des cohortes les plus anciennes (1976 à 2006 pour l'ESG de 2006 et 1981 à 2011 pour celle de 2011).
- 5) L'échantillon est restreint aux individus âgés de 17 à 35 ans inclusivement.
- 6) Les années qui suivent l'année de terminaison des études sont supprimées.
- 7) Pour la régression logistique multinomiale du niveau de scolarité, l'échantillon est restreint à l'année de terminaison des études.

### Variables du modèle

La variable dépendante *schldone* définissant la probabilité de finir ses études est égale à 1 lorsque l'individu a terminé ses études, et elle est égale à 0 lorsque l'individu n'a pas encore terminé ses études. Cette variable indicatrice est calculée à partir de la variable *agecmplt* (âge du répondant à la fin des études) de l'ESG.

La variable dépendante et indicatrice du niveau de scolarité, « educ », est utilisée dans une régression logistique multinomiale. Elle inclut 4 niveaux de scolarité :

- *inf* : n'a pas terminé ses études secondaires.
- *des* (référence) : a obtenu un diplôme d'études secondaires ou des études partielles à l'université ou au cégep.
- *dec* : a obtenu un diplôme d'études collégiales.
- *uni* : a obtenu un diplôme égal ou supérieur au baccalauréat.

Les variables explicatives et indicatrices de la fin des études, « schldone », et du niveau de scolarité atteint sont les suivantes :

- *male* : égal à 1 si le répondant est un homme et égal à 0 si le répondant est une femme.
- *father* : égal à 1 si le répondant est un homme avec des enfants, 0 sinon.
- *mother* : égal à 1 si le répondant est une femme avec des enfants, 0 sinon.
- *agex* : égal à 1 si l'individu a  $x$  ans, 0 sinon, avec  $x = 17$  à 35 ans (la catégorie de référence est constituée des individus âgés de 17 ans).

### Résultats de régression

Les résultats des régressions logistiques sont présentés dans le tableau suivant :

TABLEAU 2 – Logit - Coefficients de la transition de fin d'études (colonne 2) et d'attribution du niveau de scolarité (colonne 3 à 5)

Variables	Fin etudes	Inf. secondaire	Collegial	Universitaire
male	.18436	.410925	-.190319	-.598743
mother	-.184402	.317556	.170107	-1.0691
father	-.182383	-.609737	-.423941	-.688694
age18	-.108408	2.5816	2.69061	1.86966
age19	.376351	2.18569	3.05998	2.89381
age20	.832675	-35.5138	3.13082	4.12123
age21	.849631	-34.2392	3.51179	5.73524
age22	1.13566	-33.5076	3.37148	6.74679
age23	1.33625	-34.0375	2.75714	6.94923
age24	1.23258	-17.3182	2.7163	7.07475
age25	1.08433	-17.6177	2.45016	6.60651
age26	.960907	-17.5423	2.60372	6.72941
age27	1.06736	-17.2421	3.01988	6.7807
age28	1.02315	-17.1569	3.16667	6.96461
age29	1.00341	-17.3664	2.61141	7.11651
age30	.875271	-17.8062	2.1641	6.38981
age31	.994532	-17.6105	2.34907	6.83184
age32	1.36367	-17.3727	2.69095	6.95154
age33	1.52	-16.9129	3.37866	7.38962
age34	1.95557	-1.56612	2.70337	6.27352
age35	2.4045	-2.35543	2.26993	6.41581
constant	-3.1736	-1.133584	-1.66927	-5.01949

## Mise en œuvre

La mise en œuvre dans SimGen est réalisée à l'aide d'un tirage uniforme, indépendant par individu dominant, et la fin des études et le niveau de scolarité associé sont déterminés lorsque le résultat de ce tirage est inférieur à la probabilité logistique prédite.

Dans SimGen, les personnes à risque pour cette transition sont les individus dominants âgés de 17 à 35 ans qui sont encore aux études. Les individus âgés de 35 ans ont une probabilité de terminer leurs études fixée à 100%. Avant l'année de fin des études, les individus sont considérés sans scolarité (aucun niveau ne leur est attribué). Le niveau de scolarité obtenu l'année de fin des études est attribué aux individus jusqu'à la fin de leur vie. Aucun retour aux études n'est possible après la fin des études.

## 1.3.3 État matrimonial

### Modèle économétrique

Deux régressions logistiques sont réalisées pour 1) calculer la probabilité d'entrer dans une union conjugale (union libre ou mariage, indistinctement); 2) calculer la probabilité de se séparer. La probabilité d'entrer en union et de se séparer dépend de variables similaires liées à l'âge du répondant, à son genre et à son niveau de scolarité. De plus, la probabilité de se séparer dépend également de la présence d'au moins un enfant âgé de moins de 18 ans.

1) probabilité  $c$  pour un individu  $i$  de se mettre en couple l'année  $t$  :

$$\mu_{i,t} = \mu_0 + \mu_1 age_{i,t} + \mu_2 male_{i,t} + \mu_3 educ_{i,t}$$

$$\Pr(c_{i,t} = 1) = \frac{\exp(\mu_{i,t})}{1 + \exp(\mu_{i,t})}$$

2) probabilité  $s$  pour un individu  $i$  de se séparer l'année  $t$  :

$$\mu_{i,t} = \mu_0 + \mu_1 age_{i,t} + \mu_2 male_{i,t} + \mu_3 educ_{i,t} + \mu_4 kid_{i,t}$$

$$\Pr(s_{i,t} = 1) = \frac{\exp(\mu_{i,t})}{1 + \exp(\mu_{i,t})}$$

## Données et échantillon

Les modèles logistiques sont estimés à partir des vagues 2006 et 2011 de l'Enquête sociale générale (ESG) réalisée auprès des ménages par Statistique Canada.

L'échantillon utilisé pour calculer les transitions matrimoniales est défini en suivant plusieurs étapes :

- 1) Les données des vagues 2006 et 2011 de l'ESG sont regroupées dans une base unique.
- 2) L'échantillon est ensuite restreint aux données de la province du Québec (variable *prv*).
- 3) Un fichier de pseudo panel des répondants qui recense l'historique des transitions d'unions et de séparations d'ordre 1 à 4 est créé (jusqu'à 4 unions et séparations sont donc permises tout au long de la vie).
- 4) Seul l'historique des transitions du pseudo panel depuis 30 années est conservé afin d'éviter les effets des cohortes les plus anciennes (1976 à 2006 pour l'ESG de 2006 et 1981 à 2011 pour celle de 2011).

## Variables du modèle

Pour calculer la transition de mise en union, la variable dépendante est égale à 0 lorsque l'individu est célibataire et la variable dépendante est égale à 1 à partir de l'année de la mise en couple. Symétriquement, pour le calcul de la transition de séparation, la variable dépendante est égale à 0 lorsque l'individu est en couple et la variable dépendante est égale à 1 à partir de l'année de la séparation. Il faut préciser que le fait de devenir veuf n'est pas considéré comme une transition de séparation dans le modèle logistique.

### 1) Variables explicatives des transitions de mise en couple

**Genre (variable indicatrice) :**

- *male* : égal à 1 si le répondant est un homme et égal à 0 si le répondant est une femme.

**Âge (variables indicatrices) :**

- *age1619* : le répondant a entre 16 et 19 ans.
- *age2024* : le répondant a entre 20 et 24 ans.
- *age2529* : le répondant a entre 25 et 29 ans.
- *age3034* (référence) : le répondant a entre 30 et 34 ans.
- *age3539* : le répondant a entre 35 et 39 ans.
- *age4044* : le répondant a entre 40 et 44 ans.
- *age4549* : le répondant a entre 45 et 49 ans.
- *age5054* : le répondant a entre 50 et 54 ans.
- *age5559* : le répondant a entre 55 et 59 ans.
- *age6065* : le répondant a entre 60 et 65 ans.

**Scolarité (variables indicatrices) :**

- *insch* : le répondant n'a pas encore terminé ses études.
- *inf* (référence) : le répondant a terminé ses études mais n'a pas complété ses études secondaires.
- *des* : le répondant a terminé ses études et a un diplôme d'études secondaires ou des études partielles à l'université ou au cégep.
- *dec* : le répondant a terminé ses études et a un diplôme d'études collégiales.
- *uni* : le répondant a terminé ses études et a un diplôme égal ou supérieur au baccalauréat.

### 2) Variables explicatives des transitions de séparation

**Genre (variable indicatrice) :**

- *male* : égal à 1 si le répondant est un homme et égal à 0 si le répondant est une femme.

**Âge :**

- *mage* : âge du répondant si c'est un homme, 0 sinon.
- *mage2* : âge au carré du répondant si c'est un homme, 0 sinon.
- *mage3* : âge au cube du répondant si c'est un homme, 0 sinon.
- *wage* : âge du répondant si c'est une femme, 0 sinon.
- *wage2* : âge au carré du répondant si c'est une femme, 0 sinon.
- *wage3* : âge au cube du répondant si c'est une femme, 0 sinon.

**Scolarité (variables indicatrices) :**

- *insch* : le répondant n'a pas encore terminé ses études.
- *inf* (référence) : le répondant a terminé ses études mais n'a pas complété ses études secondaires.
- *des* : le répondant a terminé ses études et a un diplôme d'études secondaires ou des études partielles à l'université ou au cégep.
- *dec* : le répondant a terminé ses études et a un diplôme d'études collégiales.
- *uni* : le répondant a terminé ses études et a un diplôme égal ou supérieur au baccalauréat.

**Enfants (variable indicatrice) :**

- *kid* : égal à 1 si le répondant a au moins un enfant de moins de 18 ans, 0 sinon.

Cette variable contrôle pour la présence d'enfants mineurs, potentiellement résidants au domicile parental ou bien à la charge de leurs parents. La présence d'enfants majeurs n'est pas prise en compte car ceux-ci ne sont potentiellement plus à la charge de leurs parents.

**Résultats de régression**

Les résultats du modèle logistique de mise en couple sont présentés dans le tableau suivant :

TABLEAU 3 – Logit - Coefficients des transitions de mise en couple

Variables	Coefficients
male	-.1837456
age1619	-.6663195
age2024	.3110995
age2529	.4030818
age3539	-.3277921
age4044	-.399524
age4549	-.5470281
age5054	-.4505239
age5559	-.8944283
age6065	-.8323357
insch	-.732394
des	.1500707
dec	.3124948
uni	.3111567
constant	-2.408129

Les résultats du modèle logistique de séparation sont présentés dans le tableau suivant :



TABLEAU 4 – Logit - Coefficients des transitions de séparations

Variables	Coefficients
male	-.7359777
mage	-.277098
mage2	.0087923
mage3	-.0000819
wage	-.327683
wage2	.0098783
wage3	-.0000913
insch	.6755069
des	-.1025277
dec	-.1902154
uni	-.4476943
kid	-.5016676
constant	.2193172

### Mise en œuvre

La mise en œuvre des transitions de formation d’union et de séparation dans SimGen est réalisée par un tirage uniforme, indépendant par individu dominant, et ces événements surviennent lorsque le résultat de ce tirage est inférieur à la probabilité prédite.

Lorsqu’un individu dominant *D1* est sélectionné pour former une union, une banque d’individus dominants est créée à partir des individus en couple, ayant les mêmes caractéristiques que ce dernier quant à l’âge, le genre et le niveau de scolarité et ayant une différence d’âge avec leur conjoint de moins de 5 ans. Si aucun individu dominant en couple avec ces caractéristiques n’est trouvé, une nouvelle banque d’individus dominants est créée à partir des individus dominants en couple, ayant les mêmes caractéristiques que l’individu dominant *D1* quant au genre et au niveau de scolarité, et ayant une différence d’âge avec leur conjoint de moins de 20 ans. Par la suite, les caractéristiques du conjoint *C1* de l’individu dominant *D1* sont obtenues en attribuant à ce conjoint les mêmes caractéristiques que le conjoint *C2* d’un individu dominant *D2* sélectionné aléatoirement à partir de la banque d’individus dominants créée à cet effet.

### 1.3.4 Mortalité

Chaque année  $t$ , un individu d’âge  $a$  et de genre  $g$  a une probabilité  $P(t,a,g)$  de décéder. Cette probabilité, définie comme un taux de mortalité, est calculée à partir des quotients prospectifs de mortalité selon l’âge et le sexe estimés par Statistique Canada entre 2013-2014 et 2062-2063 (juillet-juin) pour les provinces et territoires. Le [rapport technique](#) de Statistique Canada présente la méthodologie et les hypothèses utilisés pour construire ces quotients prospectifs.

L’âge, le genre et la cohorte de naissance sont donc les seuls déterminants de l’espérance de vie des individus. Notons également que les immigrants et les natifs ont des probabilités équivalentes de décès en fonction de leur âge, de leur genre et de leur cohorte.

### 1.3.5 Migrations

#### Immigration

Le taux prospectif d'immigration internationale est égal à 6,6‰. Ce taux est calculé en divisant le nombre d'immigrants projeté dans le scénario de référence de l'ISQ (55 000) par la population québécoise enregistrée par Statistique Canada en 2017 (8 302 063). Les caractéristiques socio-économiques et démographiques des nouveaux immigrants internationaux sont attribuées en fonction des immigrants internationaux récents issus de la BDSPS de Statistique Canada pour l'année 2017. Chaque année  $t$ , on tire aléatoirement dans la base de départ 6,6 pour mille de la sous-population des immigrants récents (depuis 5 ans ou moins). Les caractéristiques socio-économiques et démographiques des nouveaux immigrants sont alors celles des immigrants tirés de la BDSPS de 2017 : l'âge, le genre, le niveau de scolarité, la présence de conjoint et le nombre d'enfants.

#### Émigration

Les caractéristiques des émigrants dépendent uniquement de l'âge. L'émigration intègre les émigrants internationaux ainsi que le solde migratoire interprovincial. À chaque âge donné, la probabilité d'émigrer est égale pour toutes les personnes dominantes. Les émigrants d'un âge donné sont tirés de manière aléatoire. De plus, on considère que le(la) conjoint(e) du dominant ainsi que tous ses enfants âgés de moins de 18 ans émigrent avec la personne dominante. Le taux d'émigration par âge est calculé à partir du nombre d'émigrants interprovinciaux par classe d'âge en 2018-2019 du tableau 17-10-0015-01 « Estimations des composantes de la migration interprovinciale, par âge et sexe, annuelles », du nombre d'émigrants internationaux par classe d'âge en 2018-2019 du tableau 17-10-0014-01 « Estimations des composantes de la migration internationale, par âge et sexe, annuelles » et de la population québécoise par classe d'âge au 1er juillet 2018 du tableau 17-10-0005-01 « Estimations de la population au 1er juillet, par âge et sexe ». À noter que le nombre d'émigrants interprovinciaux à chaque âge a été normalisé sur les hypothèses du solde interprovincial annuel de l'ISQ (9 000 personnes). Les taux d'émigration par classe d'âge sont les suivants :

Classe d'âge	Taux d'émigration (‰)
15 à 19 ans	1,37
20 à 24 ans	3,08
25 à 29 ans	5,03
30 à 34 ans	4,90
35 à 39 ans	3,74
40 à 44 ans	2,72
45 à 49 ans	2,00
50 à 54 ans	1,38
55 à 59 ans	0,99
60 à 64 ans	0,84
65 à 69 ans	0,82
70 à 74 ans	0,64
75 à 79 ans	0,65
80 à 84 ans	0,62
85 à 89 ans	0,55
90 ans et plus	0,41

## 1.4 Résultats

Cette section présente brièvement les résultats du modèle SimGen et les compare aux données officielles du Québec.

### 1.4.1 Données de comparaison

Ces données proviennent de différentes sources officielles :

- Données pour la population totale de 1998 à 2018 (Figure 1) : les données proviennent de séries de l'ISQ.
- Données de projections de population (Figures 1 & 2) : les projections de population sont basées sur le scénario moyen de l'ISQ à partir des données corrigées du recensement de 2016. Pour plus d'information concernant la méthodologie utilisée pour le calcul des projections de population, se référer au rapport « [Perspectives démographiques du Québec et des régions, 2016-2066, édition 2019](#) » produit par l'ISQ.
- Données par niveau de scolarité (Figure 3) : les données concernant le plus haut niveau de scolarité atteint proviennent des fichiers de microdonnées à grande diffusion des recensements de 2006, 2011 et 2016. Ces données sont disponibles par l'entremise de l'Initiative de démocratisation des données (IDD) de Statistique Canada.
- Données pour personnes en couple (Figure 4) : ces données proviennent des estimations de la population au 1er juillet, selon l'état matrimonial ou l'état matrimonial légal, l'âge et le sexe (Tableau : 17-10-0060-01), qui sont produites par [Statistique Canada](#).

### 1.4.2 Données de simulation

Comme mentionné précédemment, la base de données de départ de SimGen est tirée de la [Base de données de simulation de politiques sociales \(BDSPS\)](#).

Pour ce qui est des résultats analysés, ceux-ci proviennent d'une simulation de 2017 à 2040 utilisant les hypothèses suivantes :

```
reference = model(start_yr=2017, stop_yr=2040)
reference.startpop('startpop')
reference.immig_assumptions(init='newimmpop', num=0.0066)
reference.birth_assumptions(scenario='reference')
reference.dead_assumptions(scenario='low')
```

### 1.4.3 Résultats des comparaisons

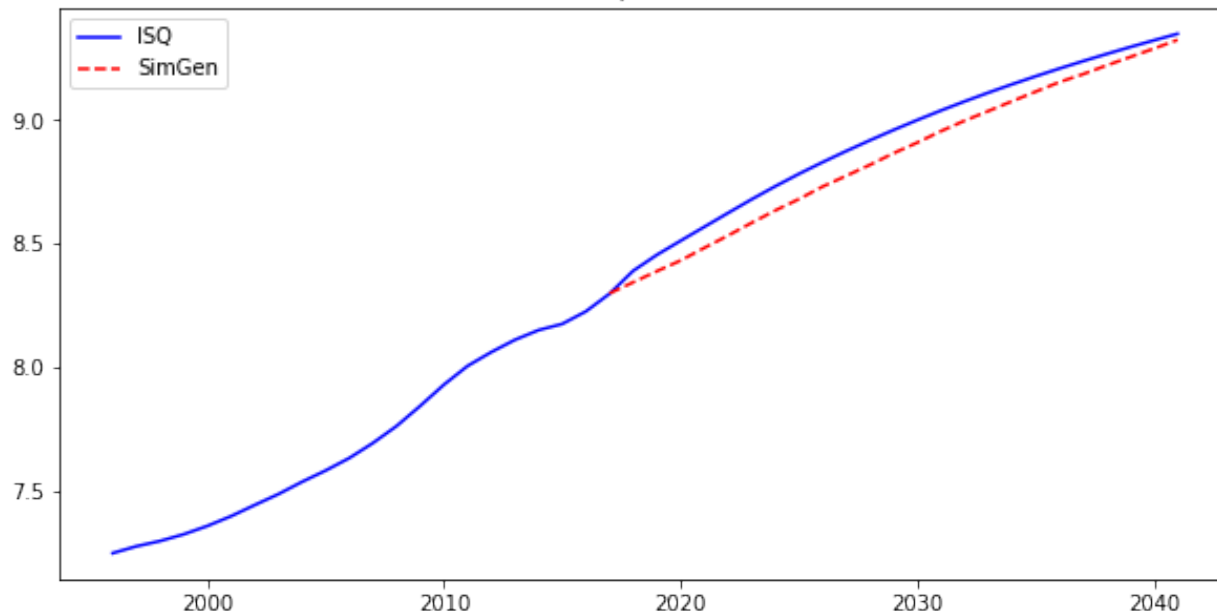
Il est important de noter que l'objectif de cet exercice n'est pas de reproduire exactement les projections des différentes agences statistiques, mais d'illustrer les différences afin de mieux comprendre les éventuels impacts sur les différents modules et les modèles utilisant les résultats de SimGen comme intrant.

#### Population totale

Les données de l'année d'initialisation du modèle SimGen en 2017 sont calibrées sur les données de population par âge et par genre de l'ISQ pour cette même année. La Figure 1 compare les projections de population totale du modèle SimGen (2017-2040) avec les projections réalisées par l'Institut de la statistique du Québec (ISQ) à partir de l'année 2017.

Le modèle SimGen reproduit avec fidélité les projections réalisées par l'ISQ. En 2040, la population totale obtenue par SimGen (9,29 millions d'habitants) est similaire à la population totale obtenue par l'ISQ (9,32 millions d'habitants).

Figure 1. Comparaison des projections de population totale  
(en millions), Québec, 2017-2040



### Population par groupes d'âge

La Figure 2 compare les projections de population par groupes d'âge réalisées avec SimGen avec les projections de l'ISQ pour les années 2017 à 2040. On remarque que les deux séries de projections sont relativement similaires.

Les projections de la population âgée de 65 ans et plus sont quasiment identiques entre l'ISQ et SimGen. Cette population devrait évaluer 2,45 millions en 2040 selon l'ISQ et elle devrait évaluer 2,43 millions la même année selon SimGen. En revanche, les projections de population pour les 0-24 ans et pour les 25-64 ans présentent de légères différences entre l'ISQ et SimGen. En 2040, la population âgées de 0 à 24 ans devrait évaluer 2,37 millions selon l'ISQ et elle devrait évaluer 2,08 millions selon SimGen. La même année, la population âgées de 25 à 64 ans devrait évaluer 4,51 millions selon l'ISQ et elle devrait évaluer 4,78 millions selon SimGen.

### Niveau de scolarité

Premièrement, on remarque à la Figure 3 un saut entre les données du recensement de 2016 et celles projetées par SimGen pour 2017 en ce qui concerne les proportions de population selon le plus haut niveau de scolarité atteint. Cet écart s'explique par le fait que la variable de scolarité n'est pas catégorisée de la même manière dans la base de données initiale et dans les données publiques des recensements. Il faudra donc porter une attention particulière à cette variable pour tous projets ayant pour objectif d'étudier le système québécois d'éducation.

Pour ce qui est des tendances générales, on remarque une augmentation de la proportion de personnes ayant obtenu un diplôme de niveau universitaire et une diminution de la proportion des trois autres niveaux de scolarité.

Figure 2. Comparaison des projections de population par groupes d'âge  
(en millions), Québec, 2017-2040

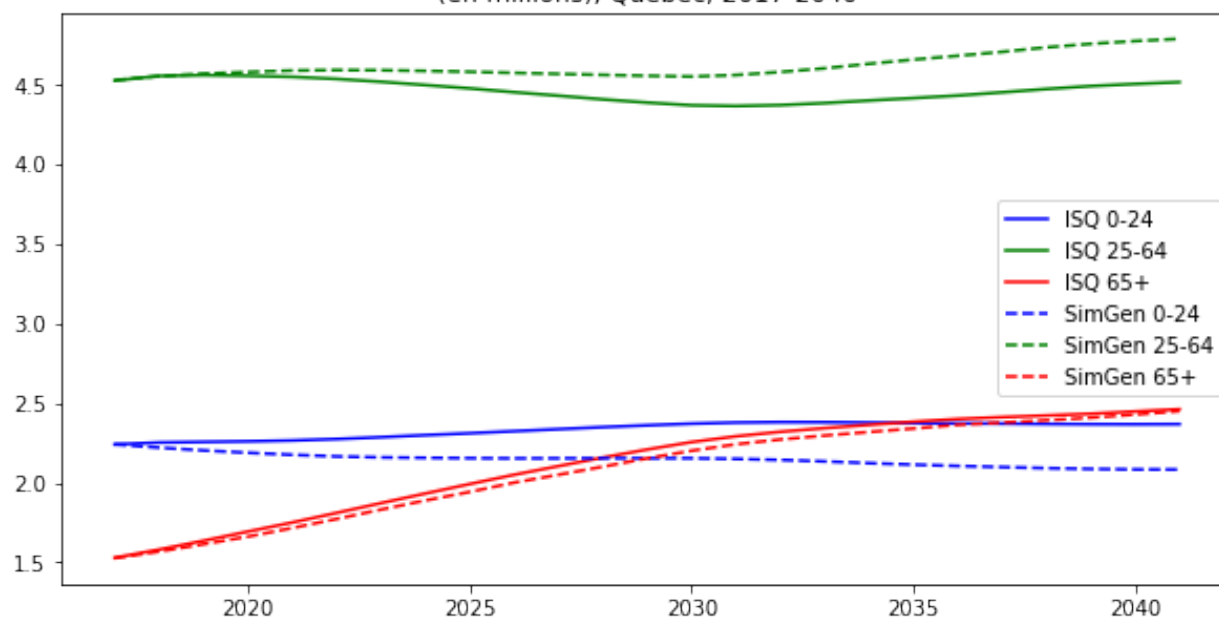
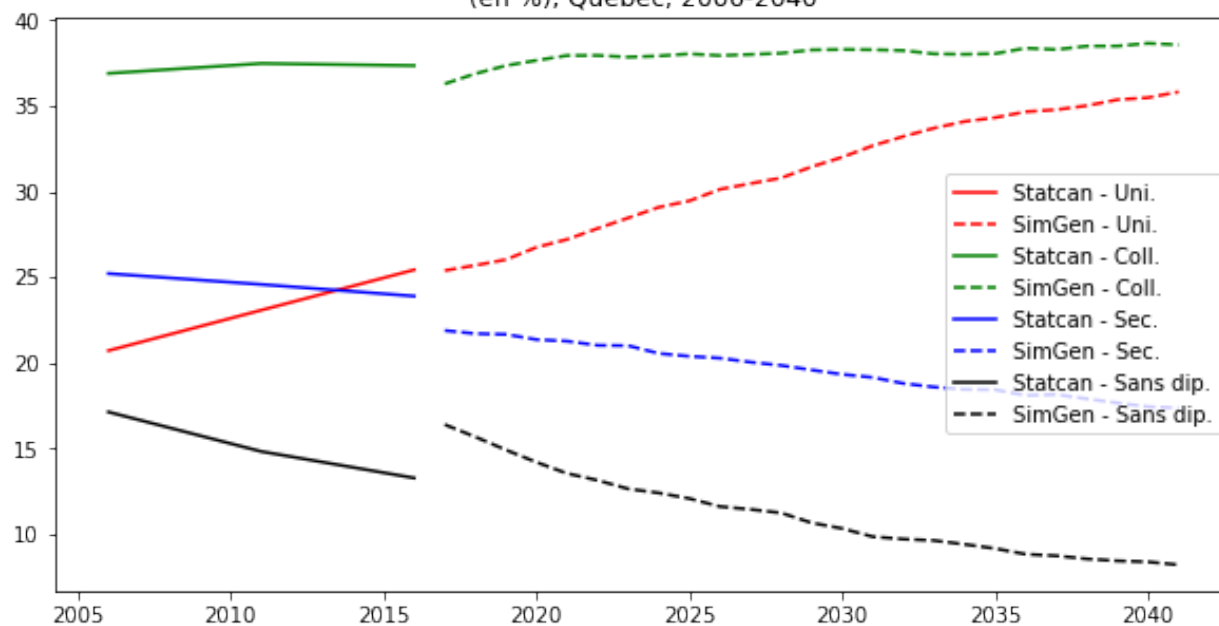
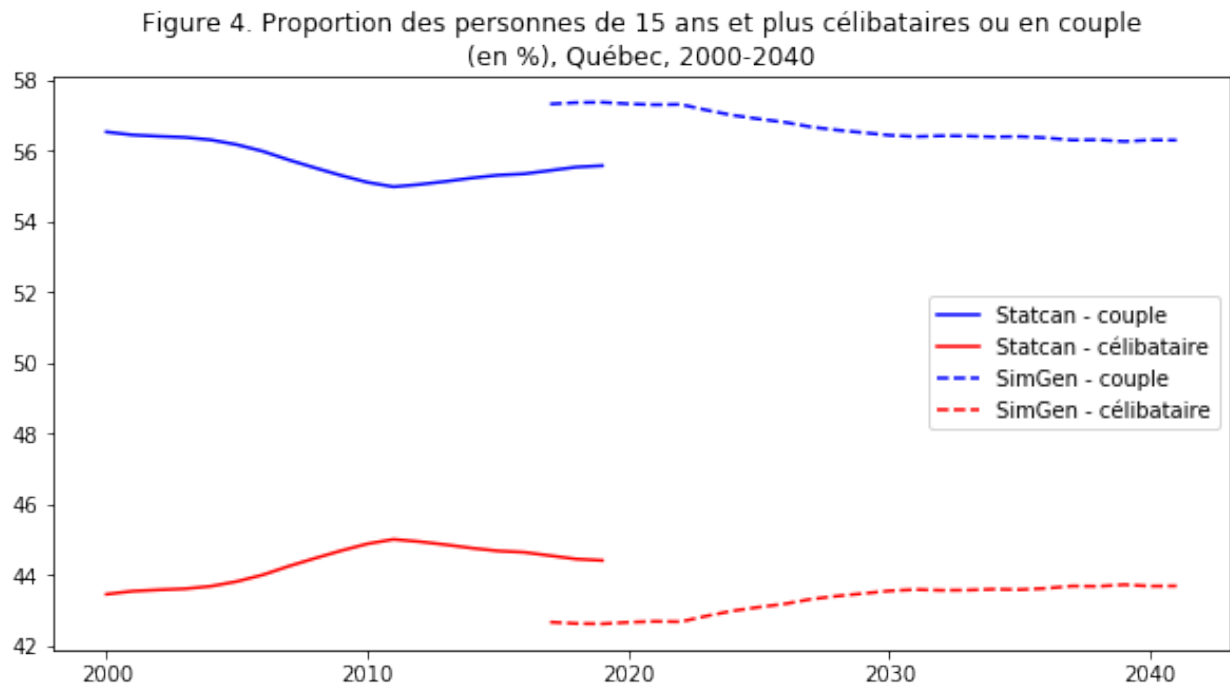


Figure 3. Proportion de la population par niveaux de scolarité  
(en %), Québec, 2006-2040



## Personnes en couple



Concernant la part de personnes en couple, la Figure 4 met en évidence un léger décalage entre les données historiques de Statistique Canada et les projections de SimGen à la fin des années 2010. Cet écart est comparable à ceux observés entre les estimations de population annuelles et le recensement.

Les résultats de SimGen mettent en avant une légère diminution de la proportion de personnes de 15 ans et plus en couple au Québec pour l'ensemble de la période de projection.

## 1.5 Dictionnaire (classes et fonctions)

### 1.5.1 Données

Les fonctions de données permettent de préparer les données pour la simulation.

`simgen.bdsps (file, year=2017, iprint=False, file_format='.dta')`

Nettoyage de la BDSPPS.

Fonction qui permet de mettre en forme la BDSPPS.

#### Paramètres

- **year** (*int*) – année de la base de départ (défaut=2017)
- **iprint** (*boolean*) – switch pour imprimer ou non des outputs intermédiaires de cette fonction (défaut=False)

`simgen.isq (year)`

Population par âge de l'ISQ.

Fonction qui permet d'obtenir la population par âge de l'ISQ.

**Paramètres** **year** (*int*) – année pour la population

**Renvoie** dataframe *pandas* contenant la population par âge (hommes et femmes)

**Type renvoyé** dataframe

**class** `simgen.parse`

Mise en forme des variables pour référence de SimGen.

Classe qui permet de prendre un dataframe provenant d'une base de données particulière et retourner un dataframe propre interprétable par SimGen. On peut faire correspondre les noms de variables avec l'initialisation de la classe en utilisant les dictionnaires *map\_hh*, *map\_sp* et *map\_kd* pour les trois registres.

**dominants** (*data*)

Mise en forme des dominants.

Fonction membre qui permet de prendre un dataframe dominant et d'appliquer les dictionnaires *map\_hh* pour les noms de variables qui concordent avec SimGen.

**Paramètres** **data** (*dataframe*) – dataframe de dominants

**Renvoie** dataframe avec les noms de variables de SimGen

**Type renvoyé** dataframe

**kids** (*data*)

Mise en forme des enfants.

Fonction membre qui permet de prendre un dataframe enfants et d'appliquer les dictionnaires *map\_kd* pour les noms de variables qui concordent avec SimGen.

**Paramètres** **data** (*dataframe*) – dataframe d'enfants

**Renvoie** dataframe avec les noms de variables de SimGen

**Type renvoyé** dataframe

**spouses** (*data*)

Mise en forme des conjoints.

Fonction membre qui permet de prendre un dataframe conjoint et d'appliquer les dictionnaires *map\_sp* pour les noms de variables qui concordent avec SimGen.

**Paramètres** **data** (*dataframe*) – dataframe de conjoints

**Renvoie** dataframe avec les noms de variables de SimGen

**Type renvoyé** dataframe

**class** `simgen.population`

Structure de population.

Cette classe permet d'abriter sous un seul toit les dominants, conjoints et enfants et permet certaines opérations.

**input** (*hh*, *sp*, *kd*)

Fonction pour entrer les registres.

Fonction qui permet d'entrer les registres dominants, conjoints et enfants qui ont été préalablement passés dans *parse()*.

**Paramètres**

— **hh** (*dataframe*) – dataframe des dominants

— **sp** (*dataframe*) – dataframe des conjoints

— **kd** (*dataframe*) – dataframe des enfants

## 1.5.2 Transitions

**class** `simgen.update`

Classe pour les transitions.

Classe permettant d'effectuer différentes transitions d'une année à l'autre.

**birth** (*pop*, *year*, *ntarget*)

Fonction de transitions pour les naissances.

**Paramètres**

— **pop** (*population*) – population (instance de la classe *population*)

- **year** (*int*) – année de la transition
- **ntarget** (*int*) – nombre de naissances visé (si alignement)

**Renvoie** instance de la classe population

**Type renvoyé** *population*

**dead** (*pop*, *year*)

Fonction de transitions pour les décès.

**Paramètres**

- **pop** (*population*) – population (instance de la classe population)
- **year** (*int*) – année de la transition

**Renvoie** instance de la classe population

**Type renvoyé** *population*

**divorce** (*pop*, *year*)

Fonction de transitions pour les dissolutions d'unions.

**Paramètres**

- **pop** (*population*) – population (instance de la classe population)
- **year** (*int*) – année de la transition

**Renvoie** instance de la classe population

**Type renvoyé** *population*

**educ** (*pop*, *year*)

Fonction de transitions pour changements de niveau d'éducation.

**Paramètres**

- **pop** (*population*) – population (instance de la classe population)
- **year** (*int*) – année de la transition

**Renvoie** instance de la classe population

**Type renvoyé** *population*

**emig** (*pop*, *year*)

Fonction de transitions pour gérer l'émigration.

**Paramètres**

- **pop** (*population*) – population (instance de la classe population)
- **year** (*int*) – année de la transition

**Renvoie** instance de la classe population

**Type renvoyé** *population*

**marriage** (*pop*, *year*)

Fonction de transitions pour les formations d'unions.

**Paramètres**

- **pop** (*population*) – population (instance de la classe population)
- **year** (*int*) – année de la transition

**Renvoie** instance de la classe population

**Type renvoyé** *population*

### 1.5.3 Simulation

La classe permettant de réaliser les simulations est *model*. Voici sa description.

**class** `simgen.model` (*start\_yr=2017*, *stop\_yr=2100*)

Modèle de simulation SimGen.

Cette classe permet de créer une instance d'un modèle de microsimulation.



**Paramètres**

- **start\_yr** (*int*) – année de départ de la simulation (défaut=2017)
- **stop\_yr** (*int*) – dernière année de la simulation (défaut=2100)

**birth\_assumptions** (*scenario='reference', align=True*)

Hypothèses de fécondité.

Fonction membre qui permet de spécifier les hypothèses de fécondité.

**Paramètres**

- **scenario** (*str*) – Permet de choisir entre les différents scénarios de fécondité produits pas l'ISQ (weak, reference, strong)
- **aling** (*boolean*) – paramètre permettant d'aligner le nombre d'immigrants sur l'ISQ

**dead\_assumptions** (*scenario='medium'*)

Hypothèses de mortalité.

Fonction membre qui permet de spécifier les hypothèses de mortalité.

**Paramètres scenario** (*str*) – Permet de choisir entre les différents scénarios de mortalité produits pas l'STC (low, medium, high)

**immig\_assumptions** (*allow=True, num=0.0066, init=None*)

Hypothèses d'immigration.

Fonction membre qui permet de spécifier les hypothèses d'immigration.

**Paramètres**

- **allow** (*boolean*) – paramètre permettant d'aligner le nombre d'immigrants sur l'ISQ
- **num** (*float*) – immigration totale (nombre); par défaut, scénario de référence de l'ISQ
- **init** (*str*) – nom du fichier contenant la population d'immigrants

**set\_statistics** (*stratas=['age', 'male', 'insch', 'educ', 'married', 'nkids', 'risk\_iso']*)

Fonction déterminant les variables de sortie.

**Paramètres stratas** (*list*) – Liste des variables de sortie

**simulate** (*rep=1*)

Fonction déclenchant le lancement de la simulation.

**Paramètres stratas** (*rep*) – Nombre de réplifications

**startpop** (*file*)

Charger une population de départ.

Fonction membre qui permet de charger une population de départ.

**Paramètres file** (*str*) – nom du fichier contenant la population de départ

## 1.5.4 Statistiques

Cette classe permet de produire des statistiques dans le cadre d'une simulation.

**class** `simgen.statistics` (*stratas*)

Classe pour créer les statistiques provenant d'une simulation.

Cette classe permet de capturer la distribution de la population par strate durant une simulation. Elle permet ensuite de faire plusieurs tableaux dynamiques à partir de ces distributions.

**Paramètres stratas** (*list of str*) – liste des noms de variables du fichiers de dominants afin de stratifier la population et récolter les fréquences (pondérées)

**add** (*pop, year*)

Fonction pour ajouter une année à la distribution.

À chaque année d'une simulation, cette fonction est invoquée afin de récolter la distribution par strate dans l'année en cours. Cette population est ajoutée à *counts*.

**Paramètres**

- **pop** (*population*) – population de départ (instance de la classe population)

— **year** (*int*) – année de départ de la simulation

**freq** (*strata=None, bins=[0], sub=None*)

Fonction de fréquences.

Fonction qui permet, à l’aide de *counts*, de calculer les fréquences pondérées pour une strate donnée. Deux options sont disponibles : l’une, *bins*, permet de modifier les catégories de la strate (par exemple le groupe d’âge), tandis que *sub* permet de définir un critère de sélection particulier pour le calcul des fréquences (en *str*).

**Paramètres**

- **strata** (*str*) – nom de la variable par laquelle on veut découper les données ; ne pas spécifier cette option revient à demander les fréquences totales
- **bins** (*list of int*) – liste de valeurs pour découper les données selon la variable *strata* ; fonctionne seulement avec des variables de types *int* (pas de *str*)
- **sub** (*str*) – condition à respecter pour un sous-échantillon, p.ex. « *age* >= 18 »

**Renvoie** dataframe avec les fréquences par année (ligne) et valeur de la strate (colonne)

**Type renvoyé** dataframe

**prop** (*strata, bins=[0], sub=None*)

Fonction de proportions.

Fonction qui permet, à l’aide de *counts*, de calculer les proportions pondérées pour une strate donnée. Deux options sont disponibles : l’une, *bins*, permet de modifier les catégories de la strate (par exemple le groupe d’âge), tandis que *sub* permet de définir un critère de sélection particulier pour le calcul des proportions (en *str*).

**Paramètres**

- **strata** (*str*) – nom de la variable par laquelle on veut découper les données
- **bins** (*list of int*) – liste de valeurs pour découper les données selon la variable *strata* ; fonctionne seulement avec des variables de types *int* (pas de *str*)
- **sub** (*str*) – condition à respecter pour un sous-échantillon, p.ex. « *age* >= 18 »

**Renvoie** dataframe avec les proportions par année (ligne) et valeur de la strate (colonne)

**Type renvoyé** dataframe

**save** (*file*)

Fonction pour sauvegarder les fichiers de fréquences.

**Paramètres file** (*str*) – nom du fichier de sauvegarde, incluant l’extension *pkl* (format *pickle*)

**start** (*pop, year*)

Initialisation de la distribution sur l’année de départ.

Le membre de la classe qui contient les fréquences (*counts*) est populé pour l’année de départ.

**Paramètres**

- **pop** (*population*) – population de départ (instance de la classe *population*)
- **year** (*int*) – année de départ de la simulation

## 1.6 Nous joindre

### 1.6.1 Principaux contributeurs

Nicholas-James Clavet, Yann Décarie, Pierre-Carl Michaud, Julien Navaux

## 1.6.2 Personne-contact

Yann Décarie

## 1.6.3 Problèmes et améliorations

SimGen est un modèle dit « open source ». Les utilisateurs sont donc invités à signaler les problèmes liés à SimGen et à soumettre des propositions d'ajout au code en cliquant sur le [lien](#) suivant, et en cliquant sur le bouton vert « New issue ».



## CHAPITRE 2

---

### Index

---

— genindex



## CHAPITRE 3

---

### Documentation en PDF

---

Ficher pdf





### S

`simgen`, [26](#)



## A

`add()` (méthode *simgen.statistics*), 29

## B

`bdsps` (classe dans *simgen*), 8

`bdsps()` (dans le module *simgen*), 26

`birth()` (méthode *simgen.update*), 27

`birth_assumptions()` (méthode *simgen.model*), 9, 29

## D

`dead()` (méthode *simgen.update*), 28

`dead_assumptions()` (méthode *simgen.model*), 9, 29

`divorce()` (méthode *simgen.update*), 28

`dominants()` (méthode *simgen.parse*), 27

## E

`educ()` (méthode *simgen.update*), 28

`emig()` (méthode *simgen.update*), 28

## F

`freq()` (méthode *simgen.statistics*), 11, 30

## I

`immig_assumptions()` (méthode *simgen.model*), 9, 29

`input()` (méthode *simgen.population*), 27

`isq()` (dans le module *simgen*), 26

## K

`kids()` (méthode *simgen.parse*), 27

## M

`marriage()` (méthode *simgen.update*), 28

`model` (classe dans *simgen*), 8–10, 28

module  
    *simgen*, 26

## P

`parse` (classe dans *simgen*), 27

`population` (classe dans *simgen*), 27

`prop()` (méthode *simgen.statistics*), 12, 30

## S

`save()` (méthode *simgen.statistics*), 12, 30

`set_statistics()` (méthode *simgen.model*), 29

`simgen`  
    module, 26

`simulate()` (méthode *simgen.model*), 10, 29

`spouses()` (méthode *simgen.parse*), 27

`start()` (méthode *simgen.statistics*), 30

`startpop()` (méthode *simgen.model*), 9, 29

`statistics` (classe dans *simgen*), 11, 12, 29

## U

`update` (classe dans *simgen*), 27