
SimGen

Équipe CREEi

nov. 10, 2020

1	Premiers pas avec SimGen	3
1.1	Installation de SimGen	3
1.2	Importer SimGen dans un notebook ou un script	3
1.3	Rouler une première simulation	4
2	Base de départ	5
2.1	Exemple de mise en forme des données	5
2.2	Nettoyage de la BDSPS	6
2.3	Adapter les noms de variables pour SimGen	6
2.4	Création des structures de population et sauvegarde des données	7
3	Documentation des fonctions de SimGen	9
3.1	Données	9
3.2	Transitions	10
3.3	Simulation	11
3.4	Statistiques	11
4	Modèles de transition	13
4.1	Naissances	13
4.2	Fin des études et niveau de scolarité associé	15
4.3	Mises en couple et séparations	17
4.4	Décès	20
4.5	Migrations	20
5	Résultats	21
5.1	Données de comparaison	21
5.2	Données de simulation	22
5.3	Comparaison	22
6	Index et Tables	25
	Index des modules Python	27
	Index	29

Le modèle de microsimulation SimGen permet de faire des simulations démographiques pour le Québec en prenant en compte les naissances, les unions et les divorces, la scolarité, l'immigration et finalement la mortalité. Il peut être utilisé afin de produire des distributions démographiques qui peuvent servir à faire des analyses économiques. Il 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 et Retraite Québec.

CHAPITRE 1

Premiers pas avec SimGen

1.1 Installation de SimGen

On peut installer facilement SimGen en suivant deux étapes. La première est de télécharger le simulateur depuis Github

```
git clone https://github.com/creei-models/simgen simgen
```

Ensuite on doit l'installer au terminal en allant a la racine du répertoire *simgen*

```
python setup.py install
```

Par la suite, on ne devrait pas travailler dans ce répertoire d'installation. Pour obtenir une nouvelle version, il suffit de répéter les étapes qui précèdent.

1.2 Importer SimGen dans un notebook ou un script

Pour importer SimGen dans un notebook ou un script Python, on ajoute :

```
import simgen
```

On peut aussi importer des sous-modules spécifiques en utilisant :

```
from simgen import model, update, parse
```

1.3 Rouler une première simulation

On importe le modèle :

```
from simgen import model
```

On déclare une instance avec année de départ et année d'arrêt.

```
base = model(start_yr=2017, stop_yr=2040)
```

On donne le nom du fichier *pickle* qui contient la base de départ (un exemple se trouve dans `simgen/params`). Il peut être copié dans le répertoire de travail.

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

On donne les hypothèses d'immigration ainsi que le nom du fichier pour la population de nouveaux immigrants. Un exemple peut être copié dans le répertoire de travail et se trouve sous `simgen/params`.

```
base.immig_assumptions(init='newimmpop')
```

On prend les hypothèses par défaut pour les naissances :

```
base.birth_assumptions()
```

On prend aussi les hypothèses par défaut pour la mortalité :

```
base.dead_assumptions()
```

On peut toujours réinitialiser à l'aide de *reset*.

```
base.reset()
```

```
base.pop.size()
```

```
8298827.000000236
```

On peut faire une seule année en utilisant *next()*.

```
base.next()
```

Pour faire rouler la simulation de l'année de départ à l'année de fin, on lance *simulate* :

```
time base.simulate()
```

```
2040
```


2.1 Exemple de mise en forme des données

Il est important de noter que les données populationnelles de base pour SimGen proviennent de la [Base de données de simulation de politiques sociales \(BDSPS\)](#) qui est produite par Statistique Canada. La BDSPS est une base de données non confidentielle et statistiquement représentative de particuliers canadiens dans leur contexte familial, contenant suffisamment de renseignements sur chaque particulier pour calculer les impôts payés au gouvernement et les transferts de fonds versés par ce dernier.

La BDSPS est disponible par l'entremise de l'Initiative de démocratisation des données (IDD). Les professeurs et étudiants des établissements postsecondaires participants peuvent accéder au modèle par l'entremise de leur contact auprès de l'IDD. Pour plus de renseignements sur le programme de l'IDD, consulter [son site Web](#).

Dans cette section, on illustre comment effectuer la mise en forme des données provenant de la BDSPS. Toutefois, le modèle pourrait fonctionner avec une base de données différentes permettant une mise en forme identique.

Une version de base des fichiers de départ est fournie dans le répertoire params, soit *startpop.pkl* et *newimmpop.pkl*, tous deux en format « pickle ».

On utilise trois fonctions de SimGen.

- *bdsps* : manipule la BDSPS de Statistique Canada pour mettre en forme certaines variables et créer les registres avec NAS d'individus dominants.
- *parse* : reformule les noms de variables à l'aide de dictionnaires.
- *population* : crée la structure de population.

```
from simgen import bdsps, population, parse
```

2.2 Nettoyage de la BDSPS

SimGen vient avec une fonction, *bdsps*, qui nettoie les données de la BDSPS, crée les NAS et les trois registres. Cette fonction peut être adaptée pour d'autres bases de données. Fait important à noter, la fonction *bdsps* calibre aussi les poids des répondants, par âge et sexe, pour s'arrimer sur la population québécoise de 2017, selon l'ISQ.

```
hh, sp, kd = bdsps('../raw/bdsps2017.dta')
```

Voici à quoi ressemblent les registres des individus dominants, conjoints et enfants avant la mise en forme finale.

```
hh.head()
```

```
sp.head()
```

```
kd.head()
```

On partitionne pour obtenir seulement les nouveaux immigrants dans des registres différents.

```
imm = hh[hh.newimm]
imm_nas = imm.index
sp_imm = sp.loc[sp.index.isin(imm_nas),:]
kd_imm = kd.loc[kd.index.isin(imm_nas),:]
```

2.3 Adapter les noms de variables pour SimGen

Une étape importante consiste à faire la correspondance entre les noms de variables des registres de la BDSPS et ceux dans SimGen. Pour ce faire, nous avons la classe *parse()*. Cette classe permet, à l'aide d'un dictionnaire, de faire cette correspondance pour chaque registre. Par défaut, *parse()* suppose les mêmes noms de variables que dans SimGen.

```
parsing = parse()
parsing.map_hh
```

```
{'wgt': 'wgt',
 'byr': 'byr',
 'male': 'male',
 'educ': 'educ',
 'insch': 'insch',
 'nkids': 'nkids',
 'married': 'married'}
```

On fait ensuite notre correspondance propre à la BDSPS.

```
parsing.map_hh['educ'] = 'educ4'
parsing.map_hh['insch'] = 'inschool'
parsing.map_sp['educ'] = 'educ4'
parsing.map_sp['insch'] = 'inschool'
parsing.map_kd['insch'] = 'inschool'
```

On ajuste ensuite les trois registres à l'aide de ces dictionnaires.

```
hh = parsing.dominants(hh)
sp = parsing.spouses(sp)
kd = parsing.kids(kd)
```

```
hh.head()
```

```
imm = parsing.dominants(imm)
sp_imm = parsing.spouses(sp_imm)
kd_imm = parsing.kids(kd_imm)
```

2.4 Création des structures de population et sauvegarde des données

Après avoir ajusté les registres, on peut les introduire dans des structures de population. La classe *population* permet de le faire. On peut ensuite sauvegarder les objets en format « pickle » en donnant le nom du fichier.

```
pop = population()
pop.input(hh, sp, kd)
pop.save('startpop')
```

```
newimm = population()
newimm.input(imm, sp_imm, kd_imm)
newimm.save('newimmpop')
```

Documentation des fonctions de SimGen

3.1 Données

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

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

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)

`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

3.2 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`

params_birth ()

Chargement des paramètres pour transitions de naissances.

Le chargement est fait automatiquement avec la création d'une instance de la classe.

3.3 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)

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*) – switch pour 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

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

3.4 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

4.1 Naissances

4.1.1 Modèle économétrique

Pour chaque rang de naissance $k=1,2,3$, 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.

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

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

4.1.2 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 (*pooled*) dans une base unique.
2. On restreint l'échantillon aux données du Québec (variable *prv*).
3. On crée un fichier de pseudo panel des répondants qui recense l'historique des transitions de naissances 1, 2 et 3 (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. On conserve seulement l'historique des transitions du pseudo panel depuis 30 années afin d'éviter les effets des cohortes les plus anciennes (2006-30 pour la première vague ESG et 2011-30 pour la seconde vague).
5. On restreint l'échantillon aux femmes âgées de 18 à 44 ans inclusivement.

4.1.3 Variables dépendantes et variables explicatives

Les variables dépendantes pour les régressions 1, 2 et 3 sont des variables indicatrices (*dummies*), égales à 1 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) et depuis 18 ans pour l'aîné des enfants (naissance d'ordre 1).

Variables explicatives d'âge (variables indicatrices) :

- *dage1824* (référence) : la femme a entre 18 et 24 ans.
- *dage2529* : la femme a entre 25 et 29 ans.
- *dage3034* : la femme a entre 30 et 34 ans.
- *dage3539* : la femme a entre 35 et 39 ans.
- *dage40p* : la femme a entre 40 et 44 ans.

Variables explicatives d'éducation (variables indicatrices) :

- *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.

4.1.4 Résultats de la régression logistique et mise en œuvre dans le modèle démographique

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

Tableau 1 – Logit - Coefficients des transitions de naissances

var	kid1	kid2	kid3
dage2529	.5111452	.20215	-.1745108
dage3034	.0206863	.0036399	-.7896601
dage3539	-.8569678	-.8245546	-1.621034
dage40p	-1.939295	-2.335598	-3.388722
lkidage	0	-.0990092	-.0377307
insch	-1.025785	.1545281	.4842334
des	-.223628	.1972002	.0498862
dec	-.1165148	.1987852	.2902924
uni	-.1596642	.4500001	.8004084
constant	-2.532843	-1.634364	-2.288871

L'implémentation dans le modèle démographique est réalisée par un tirage uniforme, indépendant par individu dominant, et une naissance survient quand le résultat de ce tirage est inférieur à la probabilité logistique prédite.

Dans le modèle de simulation démographique, les personnes à risque pour cette transition sont les femmes en couple (qu'elles soient enregistrées comme l'individu dominant ou la conjointe dans la BDSPS) âgées de 18 ans à 44 ans inclusivement.

Il faut préciser que les effets marginaux obtenus pour le logit appliqué au 3e enfant (*kid3*) est utilisé dans la simulation démographique pour calculer l'occurrence de la naissance du 3e enfant, mais également des enfants suivants.

4.2 Fin des études et niveau de scolarité associé

Tous les enfants débutent leurs études l'année de leurs 5 ans. La présente transition calcule la probabilité de finir ses études ainsi que le niveau de scolarité correspondant.

4.2.1 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 ordinaire est appliquée pour calculer la probabilité de finir ses études et un modèle logistique multinomial est utilisé pour définir le niveau de scolarité correspondant.

- 1) probabilité d'un individu i de finir ses études f 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 d'éducation $e = 1$ (*inf*), 2 (*des*) [référence], 3 (*dec*), 4 (*uni*) atteint par un individu i l'année de terminaison des études en t :

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

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

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

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

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

4.2.2 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 d'éducation 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) On restreint l'échantillon aux données du Québec (variable *prv*).
- 3) On crée un fichier de pseudo panel des individus répondants qui recense l'historique des transitions de fin d'études et le niveau de scolarité associé.
- 4) On conserve seulement l'historique des transitions du pseudo panel depuis 30 années afin d'éviter les effets des cohortes les plus anciennes (2006-30 pour la première vague ESG et 2011-30 pour la seconde vague).
- 5) On restreint l'échantillon aux individus âgés de 17 à 35 ans inclusivement.
- 6) On supprime les années qui suivent l'année de terminaison des études.
- 7) Pour la régression logistique multinomiale du niveau de scolarité, l'échantillon est restreint à l'année de terminaison des études.

4.2.3 Variables dépendantes et variables explicatives

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 (*dummy*) 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.
- *age_x* : 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).

4.2.4 Résultats des régressions logistiques et mise en œuvre dans le modèle démographique

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 des niveaux d'études (colonne 3 à 5)

var	schldone	inf	dec	uni
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	-.133584	-1.66927	-5.01949

La mise en œuvre dans le modèle démographique 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 le modèle de simulation démographique, les personnes à risque pour cette transition sont les individus dominants âgés de 18 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 éducation (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.

4.3 Mises en couple et séparations

4.3.1 Modèle économétrique

Deux régressions logistiques sont réalisées pour 1) calculer la probabilité d'entrer dans une union (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 d'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 d'un individu i de se séparer l'année t :

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

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

4.3.2 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 maritales 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) On restreint l'échantillon aux données de la province du Québec (variable *prv*).
- 3) On crée un fichier de pseudo panel des individus répondants qui recense l'historique des transitions d'unions et de séparations d'ordre 1 à 4 (jusqu'à 4 unions et séparations sont possibles tout au long de la vie).
- 4) On conserve seulement l'historique des transitions du pseudo panel depuis 30 années afin d'éviter les effets des cohortes les plus anciennes (2006-30 pour la première vague ESG et 2011-30 pour la seconde vague).

4.3.3 Variables dépendantes et variables explicatives

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.

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

Éducation (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.

4.3.4 Résultats des régressions logistiques et mise en œuvre dans le modèle démographique

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

var	union
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

var	separation
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

La mise en œuvre des transitions de mise en couple et de séparation dans le modèle démographique est réalisée par un tirage uniforme, indépendant par individu dominant, et une mise en couple ou une séparation survient lorsque le résultat de ce tirage est inférieur à la probabilité logistique prédite.

Pour l'individu dominant *D1* nouvellement en couple, les caractéristiques du nouveau conjoint *C1* sont attribuées en tirant d'abord aléatoirement un autre individu dominant *D2* ayant le même genre et le même niveau de scolarité que *D1*. Les caractéristiques du conjoint *C2* (âge, genre et scolarité) sont alors attribuées au nouveau conjoint *C1* du dominant *D1* nouvellement en couple. Dans un premier temps, on restreint le bassin de tirage aux dominants *D2* qui ont un écart d'âge avec leur conjoint *C2* qui est inférieur à 5 ans (en valeur absolue). Si, pour un dominant *D1*, aucun conjoint n'est identifié en appliquant cette restriction d'écart d'âge, alors on réalise un second tirage dans le bassin des dominants *D2* qui ont le même genre et le même niveau de scolarité que *D1*, et qui ont un écart d'âge avec leur conjoint *C2* inférieur à 20 ans (en valeur absolue).

Il faut également préciser que dans la version actuelle de SimGen, le nouveau conjoint *C1* qui est attribué au dominant *D1* est systématiquement du sexe opposé. De futures versions du modèle pourront intégrer une représentation plus diversifiée des unions matrimoniales.

4.4 Décès

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 de 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.

4.5 Migrations

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 $0,0066 * \text{Population}(t)$ nouveaux immigrants parmi ceux de l'année 2017. 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.

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

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

5.1 Données de comparaison

Ces données proviennent de différentes sources officielles.

5.1.1 Données historiques

Pour la population de 2018-2018, il s'agit d'estimations de population constituant une série historique de populations comparables ayant servi à la construction des projections de population basées sur le recensement de 2016.

5.1.2 Données de projections de population

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.

5.1.3 Données par niveau de scolarité

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.

5.1.4 Données pour personnes en couple

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](#).

5.1.5 Base de données de départ

Pour cet exemple les données populationnelles de base pour ce modèle proviennent de la Base de données de simulation de politiques sociales (BDSPS). Pour plus de détails, consulter la section [Base de départ](#).

5.2 Données de simulation

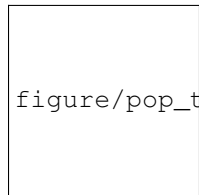
Pour ce qui est des résultats obtenus à l'aide de SimGen, ils proviennent d'une simulation de 2017 à 2040 utilisant le scénario de référence.

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

5.3 Comparaison

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 modèles utilisant SimGen.

5.3.1 Population totale

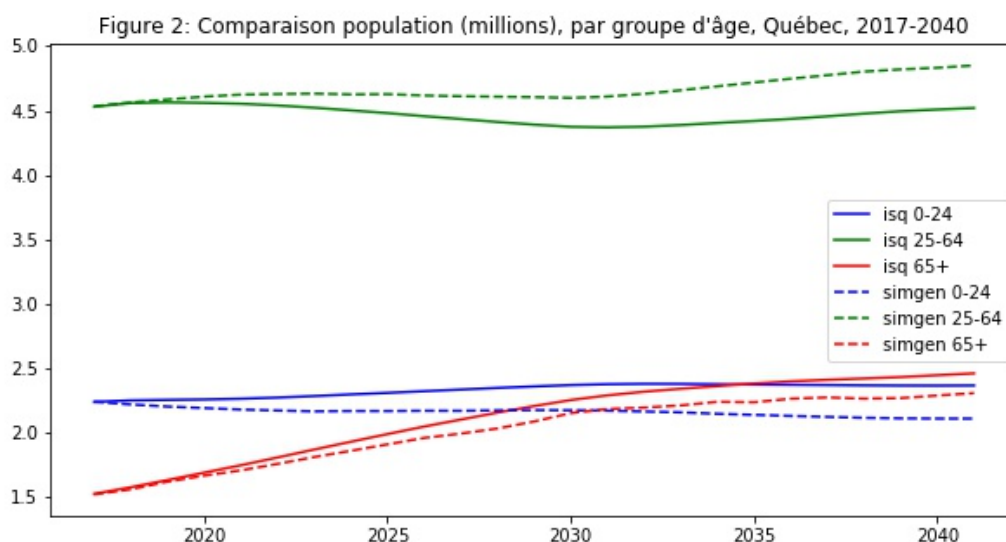


figure/pop_total.jpg

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 (9,27 millions d'habitants) est similaire à la population totale obtenue par l'ISQ (9,32 millions d'habitants).

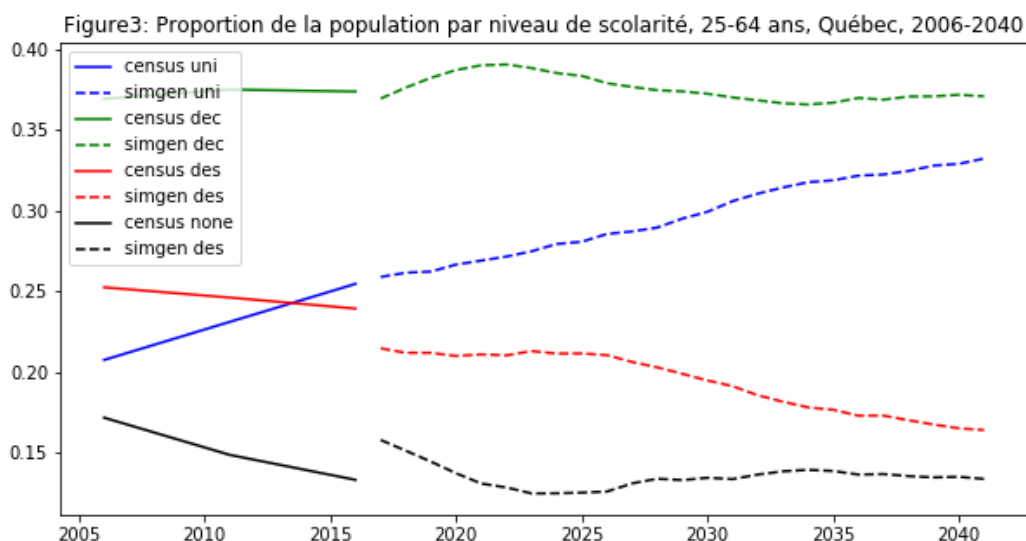
5.3.2 Population par groupe d'âge



La Figure 2 compare les projections de population par groupe d'âge réalisées avec SimGen avec les projections de l'ISQ pour les années 2017 à 2040.

On remarque que les deux séries de projection sont similaires. Par contre, les projections de population de SimGen pour les 0-24 ans et pour les 65 ans et plus sont relativement inférieures aux projections de l'ISQ. En 2040, la population âgée de 0 à 24 ans (respectivement 65 ans et plus) serait de 2,1 millions (respectivement 2,34 millions) selon SimGen, alors qu'elle serait égale à 2,37 millions (respectivement 2,45 millions) selon l'ISQ.

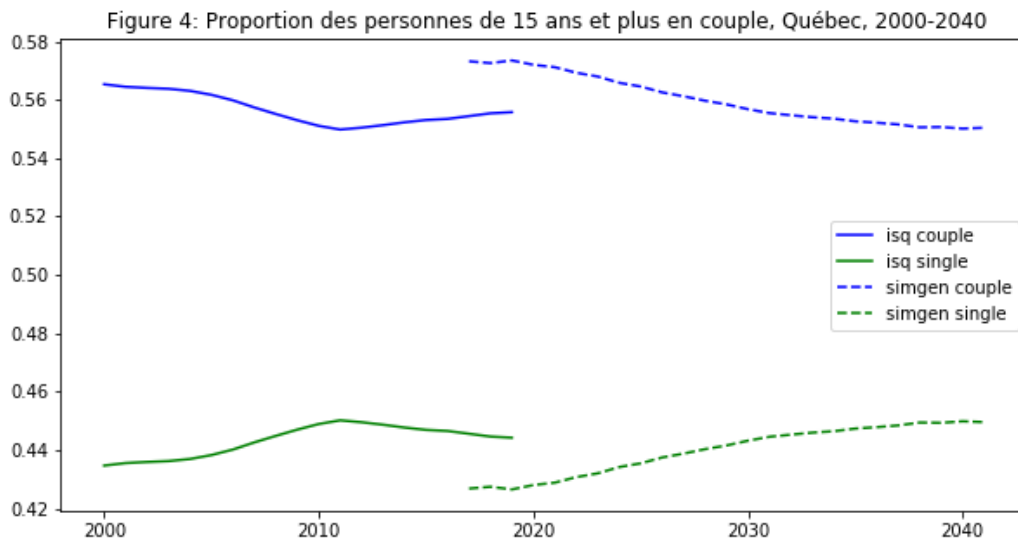
5.3.3 Niveau de scolarité



Premièrement, on remarque à la Figure 3 un saut entre les données du recensement de 2016 et celles projetées par SimGen pour 2017 pour 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 façon dans la base de données initiale et dans les données de publiques des recensements. Il faudra donc porter une attention particulière à cette variable pour tous projets voulant étudier le système québécois d'éducation.

Pour ce qui est des tendances général, on remarque une augmentation de la proportion de personnes obtenant un diplôme de niveau universitaire et une baisse pour les trois autres niveaux.

5.3.4 Personnes en couple



Pour ce qui est de la part de personnes en couple, on remarque aussi à la Figure 4 que les niveaux ont un petit décalage par rapport aux estimations de Statistique Canada. L'écart observé ici est comparable à ce qui est observé dans certaines analyses plus poussées des familles au [Canada](#).

Pour ce qui concerne la tendance générale, on remarque que la proportion de personnes en couple reste stable chez les 15 ans et plus au Québec pour l'ensemble de la période de projection.

CHAPITRE 6

Index et Tables

- `genindex`
- `modindex`
- `search`

S

`simgen`, 9

A

`add()` (méthode *simgen.statistics*), 11

B

`bdsps()` (dans le module *simgen*), 9

`birth()` (méthode *simgen.update*), 10

D

`dominants()` (méthode *simgen.parse*), 9

F

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

I

`immig_assumptions()` (méthode *simgen.model*), 11

`input()` (méthode *simgen.population*), 10

`isq()` (dans le module *simgen*), 9

K

`kids()` (méthode *simgen.parse*), 10

M

`model` (classe dans *simgen*), 11

P

`params_birth()` (méthode *simgen.update*), 10

`parse` (classe dans *simgen*), 9

`population` (classe dans *simgen*), 10

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

S

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

`simgen` (module), 9

`spouses()` (méthode *simgen.parse*), 10

`start()` (méthode *simgen.statistics*), 12

`startpop()` (méthode *simgen.model*), 11

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

U

`update` (classe dans *simgen*), 10