Simulation of Belgium with Advanced Stride

Cedric De Schepper[1], Thanh Danh Le[2], Wannes Marynen[3], and Stijn Vissers[4]

[1] University of Antwerp, Antwerp, Belgium
Cedric.DeSchepper@student.uantwerpen.be
[2] University of Antwerp, Antwerp, Belgium
ThanhDanh.Le@student.uantwerpen.be
[3] University of Antwerp, Antwerp, Belgium
Wannes.Marynen@student.uantwerpen.be
[4] University of Antwerp, Antwerp, Belgium
Stijn.Vissers@student.uantwerpen.be

Abstract. This paper presents an overview of the new features of Stride as well as a case study of disease spread in Flanders and Belgium with the extended Stride simulator.

Keywords: simulation · infectious diseases · Belgium · Flanders · comparison

1 Introduction

This paper is split into two main sections: Extended features & Simulation of Belgium. The features section details the changes made to the stride simulator:

- What have we added and why is it useful?
- What is the reasoning behind the implementation choices?
- What is the impact of the feature on the simulations?

While the first two questions are important to provide background information, the last one is critical. Adding/Using a feature without knowing what impact it will have on our results is detrimental to the simulator.

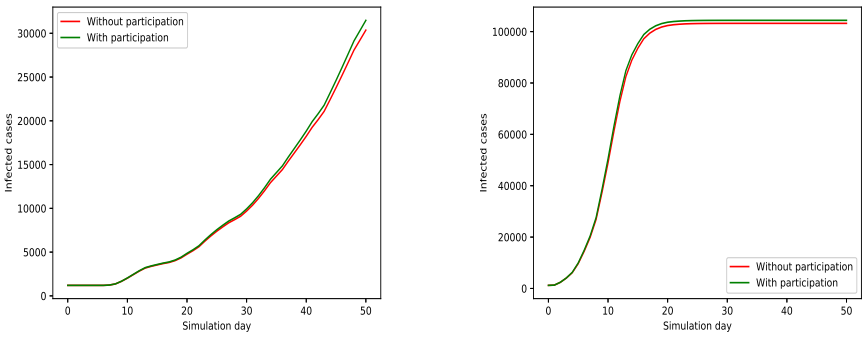
2 Extended Features

2.1 Daycare & Preschool

Introduction The functionality of the geopop component has been expanded by adding new contact types: Daycare and Preschool. This allows us to create additional contactpools of type Daycare and PreSchool for persons from 0 to 3 and 3 to 6, respectively. These contact types also hold a participation ratio that indicates how many persons of either age segment are part of one of the contactpools.

Implementation The implementation of both Daycare and Preschool are identical to K12School (which again is based on the household algorithm) since the only difference between these contact types is the amount and size of the different pools in a single Daycare/Preschool.

Impact on simulations As can be suspected, the amount of cumulative cases ends up to be higher since persons can come into contact with more individuals. This can be seen at 2. However the impact is less visible as one might assume. This is because these daycare/preschools are fairly small which diminishes the impact it has.



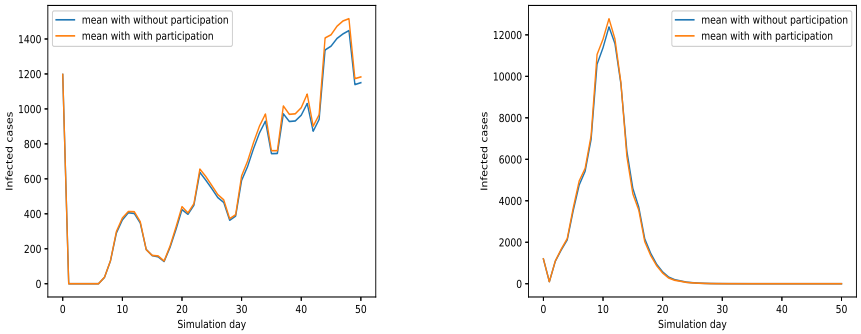
(a) measles (b) influenza

Fig. 1: Cumulative cases of 500 simulations with daycare and preschool participation at 0.45 and 0.99, respectively

2.2 Data Formats

Introduction The support of multiple data formats for our input/output data files allows us to choose the optimal format for each file needed by the simulator. To do this, we updated the JSON readers/writers and also added support for HDF5 to work alongside the existing Protobuf implementation.

Implementation HDF5 does not support automatic conversion to a stream. Therefore, the structure of GeoGridWriter and GeoGridReader had to be changed. We redefined the GeoGridReader/Writer classes into two classes, the GeoGridStreamReader/Writer and GeoGridFileReader/Writer. The former is the interface used by the JSON and Protobuf formats, and the latter is the interface for the HDF5 format.



(a) measles (b) influenza

Fig. 2: New cases of 500 simulations with daycare and preschool participation at 0.45 and 0.99, respectively

Data Format Comparison In this section, we will discuss the performance of the different file formats available for the GeoGrid in Stride.
To make this comparison, we ran the default generate and import scenario 100 times, then averaged the runtimes. We also ran these scenario's with double the population to check the impact of the population size to the file size and I/O times.

	Default scenario			**Double population**		
	Gen time	**Import time**	**File size**	**Gen time**	**Import time**	**File size**
Proto	1.27s	1.00s	13.8 MB	2.81s	1.91s	27.1 MB
JSON	6.55s	6.97s	117.2 MB	14.23s	12.99s	235.6 MB
HDF5	18.40s	10.48s	147.8 MB	37.31s	22.99s	294.5 MB

Table 1: Data formats performance

Table 1 indicates that for each data format both file size and I/O times scale linearly with the population size. The protobuf format seems to excel in I/O times, probably due to it's minimal size in comparison to the JSON or HDF5 formats. The JSON and HDF5 formats are comparable in file size, but the JSON format seems faster in I/O times, this can be explained by the greater complexity of HDF5.
The choice of data format should be based on your intended use. When planning to use the files in stride again, it's best to use protobuf because of the performance reasons mentioned previously. If readability is important, JSON is the most fitting while HDF5 excels at re-usability since other tools can easily process it.

2.3 Data Visualisation

Introduction Stride generates a large amount of data. In its raw data form, it's becomes very difficult to keep an overview. This Feature makes this large amount of data comprehensible for everyone. By showing the data on a map it gives a clear representation what a disease/virus does when introduced in an environment.

Implementation This feature is implemented using a Model-View-Controller pattern. The Model is the data that has been generated by a stride simulation. This data can be in JSON, HDF5 or protobuf form. The Viewer is a simple Qt interface with a map as the main aspect. This map is the visual representation of the provided data. The controller is implemented using c++.

To make the implementation of this possible the old location was split up in to 2 classes: Location and EnhancedCoordinate. The Location class contains all data that is related to the simulation and the EnhancedCoordinate consists of the coordinates and a pointer to the data. This allows us to use the EnhancedCoordinate in the visualizer. The Geogrid class has also been split up into 2 classes, 1 reusable for any location based application(LocationGrid) and the specific for this application class Geogrid. The LocationGrid contains a tree of EnhancedCoordinates which makes it possible to quickly find neighbouring locations while still keeping a separation for re-usability.

2.4 Demographic Profiling

Introduction Stride has the possibility to generate a population for the simulated area. Due to the possible large size of the simulated area, there can be mayor differences between regions of the simulated area, and thus this generated population is not always an accurate representation.
With this feature, one can specify certain parameters per region in the configuration file, most important of which are the population and the households profile. Furthermore, the user can provide different households profiles for central cities. The ID's of cities which are defined as major can be found in an external CSV file. All of these alterations will help the simulator to generate a more correct population.

Implementation The implementation of this feature expands on the implementation of GeoGridConfig. The old way of defining parameters is still used, these now serve as default values. Apart from the default values, one can specify the following parameters per region.

- population_size
- fraction_workplace_commuters
- fraction_college_commuters
- participation_preschool
- participation_daycare
- participation_college
- participation_workplace
- household_file
- major_household_file

If any of these parameters are not present, the default value will be used.
The generating of the contact pools will now be done per region. This way, regions with a younger population will have more daycares/preschools/k12schools and less workplaces and vice versa. The exception to this is the colleges, since there are a limited number of colleges for the entire area, and the number of regions is virtually unlimited.
To calculate the number of pools to be generated, the target fraction of the reference population is multiplied by the number of people in the region. The generators will differentiate between major cities and the rest, so that a possible younger/older population in the reference households for major cities will influence the number of schools/workplaces in major cities.
It goes without saying that the households of major cities will now be chosen from the major households profile, and that the other cities will pick households from the general households profile.

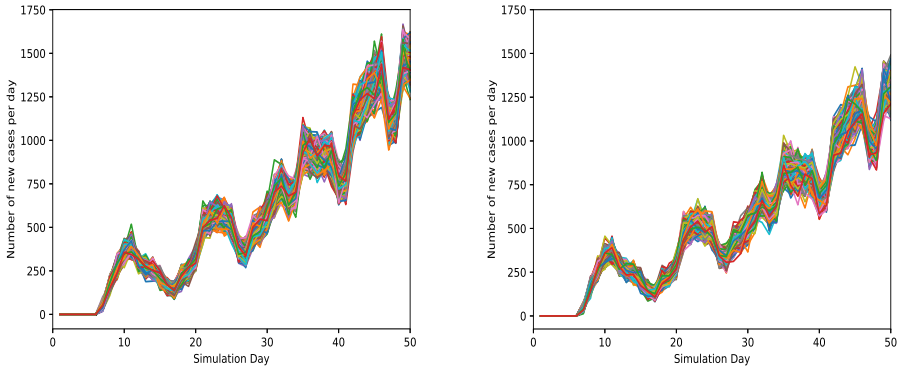
Impact on simulations Globally speaking, these alterations do not drastically change the outcome of our simulations, which is to be expected in this scenario. When computing the average number of cumulative cases after 50 days of 1024 simulation runs, we observe that the total number of cases is slightly higher when simulating without different regions, but this difference is minimal (26 996 with 4 regions, 29 899 without regions). The same result can be observed when viewing the number of new cases per day, as depicted in figure 3.

This small difference in outcomes could possibly be explained by the difference in ages of the households profiles. According to our data, the average age was higher in every region, thus relatively more people will be falling in the age bracket of 65+, and will not be working, thus having less possibility to contract the disease.

The configurations used to obtain these results are demographic_profile_default.xml and demographic_profile_regions.xml.

2.5 Workplace Size Distribution

Introduction Previously all generated workplaces had a default size of 20 and were populated using a uniform distribution. This default implementation is still available but has been extended to allow the user to specify different workplace size distributions using a CSV file. Additionally we also added two options: fast or accurate. When choosing the accurate version, the actual ratios are very close to the config ratios but at the cost of a higher running time. The fast version speeds up the process significantly but at the cost of accuracy. The accurate



(a) without regions (1 actual region) (b) 4 regions

Fig. 3: Cases per day over 50 days of 1024 simulations

version is the default when dealing with size distributions. The fast version has to be configured explicitly.

Implementation Input CSV files have to be of the form: size_min, size_max, ratio. Size_min and size_max respectively being the minimum and maximum target of persons in a workplace of a certain type. $Ratio*100$ shows the distribution in terms of percentage of the different workplace types. Currently, only CSV files are supported but the general ReaderFactory can easily be adapted to support other file readers as well.

After reading the config file, we must generate an appropriate amount of Workplace contactpools. This is done by calculating the average workplace size using:

$$\sum_{n=1}^{k} ratio_n * (max_n + min_n)/2$$

This guarantees we have enough workplaces to approximate the ratios accurately.

After generating the workplaces, they have to be populated. We first decide whether the person is active (correct age, fraction of the age bracket that is a student, fraction of the age bracket that is active). We assign a workplace to an active person that works close to home in an analogous way as the assignment of K-12 schools to students. For the commuting workers we use an algorithm analogous to that of the commuting college students. When taking workplace size distribution into account, the assignment of workplaces does not happen uniformly.

Firstly, every workplace gets assigned a workplace type, using the different ratios

as weights, which adds a min and max value. For the accurate implementation, the workplace receives a weight w based on the current size compared to the min/max for every active person:

- pool size < minimum size: w = 1 - ratio
- minimum < pool size < maximum: w = (1 - ratio)/10
- pool size ≥ max (with pool not largest workplace type): w = 0.00000000001
- pool size ≥ max (with pool largest workplace type): w = 0.00000000001 / (amount over max size)

These weights make sure the workplaces are populated as accurately as possible. It attempts to fill a workplace so the size is between its min and max value while only upgrading to a larger workplace type when absolutely necessary. These weights are then used to discretely assign a workplace to a person. Since the weights have to be recalculated for every person, it increases the running time significantly.

The fast implementation is less complex. As long as it finds workplaces that aren't full, it'll add persons to it. Only when all pools are full, will it calculate weights to find the best fit.

Accuracy vs Speed By running run_generate_default.xml, we can compare the ratios from the config file with the actual ratios, calculated during the simulation. By calculating the mean ratio and its standard deviations after 50 runs, we can clearly show that the accuracy is very high.

[min, max]	**config ratio**	**actual ratio (mean)**	**standard deviation**	**accuracy (mean)**
[1, 9]	0.77853284	0.77602516	0.00284574	99.68 %
[10, 49]	0.17190112	0.17426484	0.00275862	98.64 %
[50, 199]	0.04100390	0.04234333	0.00166064	96.84 %
[200, 400]	0.00856214	0.00736667	0.00059997	86.04 %

Table 2: Accurate implementation after 50 runs.

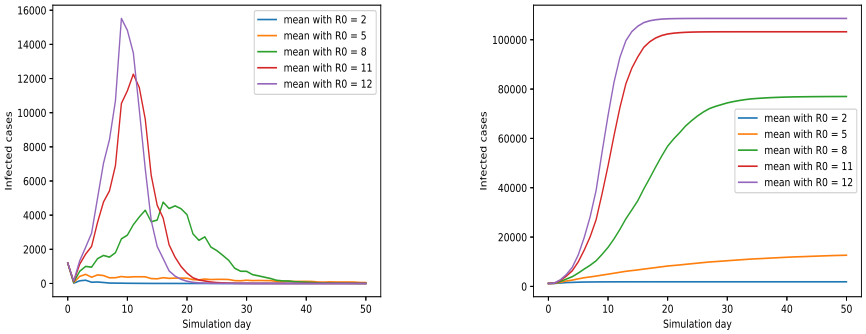
[min, max]	**config ratio**	**actual ratio (mean)**	**standard deviation**	**accuracy (mean)**
[1, 9]	0.77853284	0.86146304	0.00231454	90.37 %
[10, 49]	0.17190112	0.09855179	0.00282680	57.33 %
[50, 199]	0.04100390	0.03018688	0.00105180	73.62 %
[200, 400]	0.00856214	0.00979836	0.00049904	87.38 %

Table 3: Fast implementation after 50 runs.

If we define the weighted average as:

$$\sum_{n=1}^{k} ratio_n * avg_n$$

This gives us 99.27 % for the accurate implementation and 83.98 % for the fast implementation. While the fast implementation loses some accuracy, it's 9x as fast as the accurate implementation.



(a) new cases

(b) cumulative cases

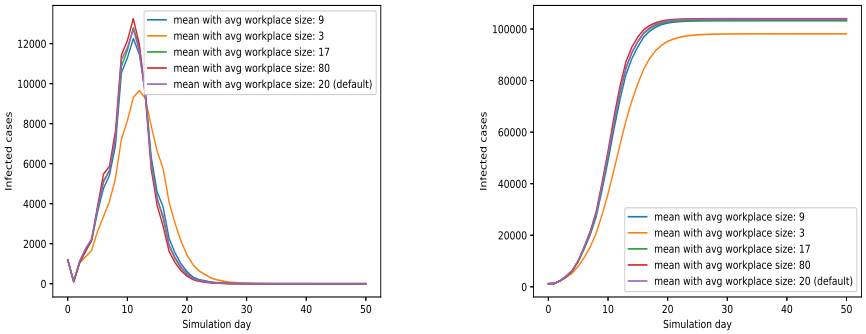
Fig. 4: Results after 500 simulations with different R0 values but constant workplace size

Impact on simulations As can be seen in Fig.4, both the amount of new and cumulative cases increases when using an increasing R0. This is very logical since R0 indicates number of individuals one infected person would infect in a completely susceptible population. This means that the higher the R0 value is, the faster a workplace gets infected. However, Fig.5 shows that an increasing average size of workplace doesn't necessarily result in more new/cumulative cases. The amount of cases eventually converges to a fairly constant number even when increasing the size significantly (20 vs 80).

3 Simulation of Belgium

3.1 Configuration of Belgium

Correspondingly to section 2.4, a simulation for Belgium can now be configured per province. For every province in Flanders and Wallonia, the parameters as seen in Listing 1.1 can be configured to match its demographic profile. The



(a) new cases (b) cumulative cases

Fig. 5: Results after 500 simulations with different average workplace sizes calculated using distribution files

Brussels-Capital Region will be considered as part of Flanders – more specifically, as part of the province Flemish Brabant. However, due to lack of recent data, some parameters aren't available. Therefore, an estimation has to be derived from data that is available and related. In this section, these parameters are discussed.

```
<fraction_college_commuters></fraction_college_commuters>
<fraction_workplace_commuters></fraction_workplace_commuters>
<household_file></household_file>
<participation_preschool></participation_preschool>
<participation_daycare></participation_daycare>
<participation_college></participation_college>
<participation_workplace></participation_workplace>
<population_size></population_size>
```

Listing 1.1: config parameters

The Flemish government publishes yearly a statistical report detailing higher education in Flanders. From this report [5], the percentage of 18-25 year olds who attend higher education (participation_college) can be calculated by taking the amount of registrations divided by the total of 18-25 year olds in Flanders [4]. However, this approach does not take older or younger students into account but compared to Wallonia [6], this result can be considered a good approximation.

Commuters, individuals who travel to another city/town, are also taken into consideration. The distance between workplace and residence can indicate if someone is a commuter. Consider the fraction of workplace commuters as the group of individuals who live more than 5-7km away from their workplace [3]. Unfortunately, the same data is not available for students. Here, the fraction of college commuters can be seen as those who travel by car or train [3]. These

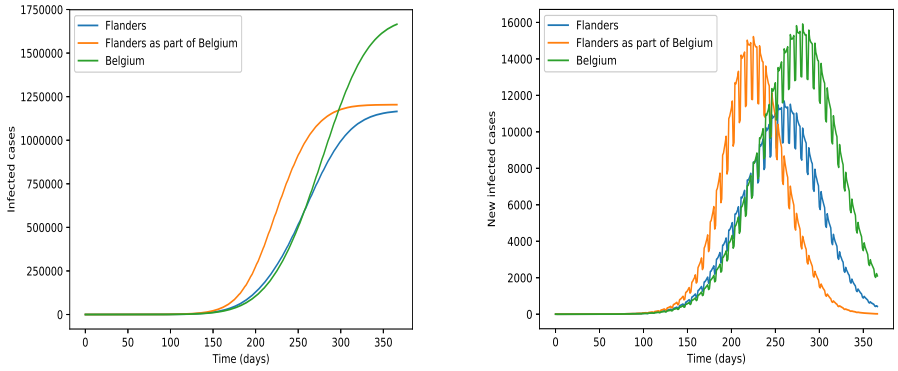
numbers are for Flanders but will also be expand to Wallonia because of similar college and workplace participation within a margin of error.

3.2 Configuration of Flanders

For Flanders we have two separate cases that we want to simulate. Firstly Flanders as an isolated region, secondly Flanders as part of Belgium. For the latter configuration, the configuration of Belgium is adjusted so that it only contains the provinces of Flanders. The configuration of the former is an specification on the latter; the cities, commuting and workplace distribution files are changed to match Flanders.

3.3 Results

In each simulation, 1 infected was introduced into the population. We performed 100 simulations for each case for the measles transmission disease.



(a) Cumulative Cases (b) New Cases per Day

Fig. 6: Average of 100 simulation without the outbreaks

Figure 6(a) shows the average of the development of infected individuals in function of the time. This average does only include the simulations where an outbreak occurred of the 100 simulation. Same for Figure 6(b), that represents the average of new infected cases in function of the time. We observe that total amount of infected of "Flanders" are similar in both cases. Also, we notice that the peak of the average new infected cases for "Flanders as part of Belgium", reaches similar heights as the peak for "Belgium". Although this peak happens earlier, even earlier than that of "Flanders".

The percentage of infected cases in the population for "Flanders"(0,150), "Flanders as part of Belgium"(0,155) and "Belgium"(0,146), indicates that there is little difference between simulations results. However, the context in which

the configuration was set up, has an impact on the simulation course as seen in Figure 6. "Flanders" and "Belgium" show similar curves but "Flanders" reaches a plateau sooner due to a smaller population. "Flanders of Belgium" is a fuse of the two previous cases. It reaches the same infected count as "Flanders" but behaves more like "Belgium".

4 Conclusion

During this semester we spent numerous hours implementing extra features and extensively running simulations to confirm their correct behaviour.

Although some of our features did not inherently change the ins and outs of the stride simulator, they did make it easier to use it, e.g. the data formats and data visualisation feature.On the other hand there were some features that changed the way stride worked on the inside. These features altered the way stride generates the population that is going to be used in the simulation, and the way they interact with each other.

Even though these alterations did not affect the results of our simulations significantly, stride will now be able to simulate the spreading of diseases more accurately, thanks to our improvements. This ultimately was the goal of our bachelor thesis, and we find that this goal is hereby reached.

References

1. BEVOLKINGSCIJFERS PER PROVINCIE EN PER GEMEENTE OP 1 JANUARI 2019, https://www.ibz.rrn.fgov.be/nl/bevolking/statistieken-van-bevolking/. Last Accessed June 19, 2019.
2. Leeftijdspiramide van België, de gewesten en provincies, https://statbel.fgov.be/nl/nieuws/belgie-telde-op-1-januari-2018-11376070-inwoners. Last Accessed June 19, 2019.
3. Janssens, D., Declercq, K., Wets, G. (2018). ONDERZOEK VERPLAATSINGSGEDRAG VLAANDEREN 5.3 (2017-2018) (p. 30, 75). Hasselt. Retrieved from https://www.mobielvlaanderen.be/pdf/ovg53/analyserapport.pdf.
4. Bevolking naar geslacht en leeftijdsgroep - Belgi, laatste jaar, https://statbel.fgov.be/nl/themas/bevolking/structuur-van-de-bevolking#panel-13. Last Accessed June 19, 2019.
5. Agentschap Hoger Onderwijs, Volwassenenonderwijs, Kwalificaties en Studietoelagen. (2018). Hoger onderwijs in cijfers. Academiejaar 2018-2019 (p. 8). Brussel. Retrieved from https://www.vlaanderen.be/publicaties/hoger-onderwijs-in-cijfers-academiejaar-2018-2019.
6. Fdration Wallonie-Bruxelles / Ministre, Administration gnrale de l'Enseignement. (2018). Les indicateurs de l'enseignement 2018 (p. 13). Retrieved from http://www.enseignement.be/index.php?page=28126&navi=4551.
7. LTB01-Werkzaamheidsgraad, http://www.werk.belgie.be/moduleDefault.aspx?id=21166. Last Accessed June 19, 2019.
8. Cijfers op maat, https://www.kindengezin.be/cijfers-en-rapporten/cijfers/kinderopvang-baby-peuter/cijfers-op-maat/#1-Meest-recente-cijfers-i. Last Accessed June 19, 2019.
9. Werkgelegenheid en werkloosheid, https://statbel.fgov.be/nl/themas/werk-opleiding/arbeidsmarkt/werkgelegenheid-en-werkloosheid#figures. Last Accessed June 19, 2019.
10. Federale Overheidsdienst Mobiliteit en Vervoer. (2017). Federale diagnostiek woon-werkverkeer 2017-2018 (p. 10-19). Brussel. Retrieved from https://mobilit.belgium.be/nl/resource/rapport_2017.
11. IWEPS. (2018). Chiffres-cls de la Wallonie Edition 2018 (p. 150). Bruxelles. Retrieved from https://www.iweps.be/publication/cc2018/
12. Jaarevolutie van de btw-plichtige ondernemingen, https://statbel.fgov.be/nl/themas/ondernemingen/btw-plichtige-ondernemingen/jaarevolutie-van-de-btw-plichtige-ondernemingen#panel-12. Last accessed 19 June 2019
13. Part des lves du maternel frquentant une cole de leur commune, https://walstat.iweps.be/walstat-catalogue.php?niveau_agre=C&indicateur_id=243500&ordre=0&periode=Ann%C3%A9e%20scolaire%202015-2016&niveau_agre=P&sel_niveau_catalogue=T. Last accessed 19 June 2019.
14. Knack.be. (2018, 22 januari). Vlaanderen internationaal koploper voor kleuterparticipatie op school. Knack. Retrieved from https://www.knack.be/nieuws/belgie/vlaanderen-internationaal-koploper-voor-kleuterparticipatie-op-school/article-normal-954397.html?cookie_check=1559821157.
15. Actieve (werkende en werkloze) en inactieve bevolking sinds 2017 op basis van de hervormde Enquête naar de ArbeidsKrachten, per kwartaal, gewest, leeftijdsklasse en onderwijsniveau,

https://bestat.statbel.fgov.be/bestat/crosstable.xhtml?view=577e7eeb-dc77-4461-b190-9164ad1da877. Last Accesed June 19, 2019.
16. TF_SOC_POP_STRUCT_2018, https://statbel.fgov.be/nl/open-data/bevolking-naar-woonplaats-nationaliteit-burgerlijke-staat-leeftijd-en-geslacht-8 . Last Accessed June 19, 2019.