CIM Virtual Population

an introduction

What is the CIM Virtual Population (VP) ?

1. The best possible sociodemographic description of Belgian population
2. Extended with (media) device possession and useage categories
3. Based on the best and most solid available sources

What is it used for ?

1. To define universes for different CIM studies in a consistent manner
2. To give research institutes the best possible Golden Standard 12+
3. To validate other VP’s already used by research partners, e.g. in Out-of-Home
4. To offer research partners a ready made & transparant ‘receiver’ database, e.g. to combine Classic TV and Online Video data

What is its structure ?

The VP is a database with >11 Mio lines that all contain the following variables

Individual Sociodemos	**Household Sociodemos**	**Location**	**Household possession (HHp)**	**Personal Use (PU)**	**Personal Frequency (PFr)**
Gender	Household size	Municipality	HHp # TV		PFr TV Viewing
Age	# Children	Arrondissement	*HHp # Smart TV*		*PFr Smart TV Viewing*
Education*	Age Category Children	Province	HHp PC	PU PC	PFr Radio Listening
Professional active (or not)	Social Group	Nielsen	HHp Laptop	PU Laptop	PFr Internet Use
Profession (current or last)		CIM Habitat	HHp Tablet	PU Tablet	
MS (main shopper)		CIM Inhabitant	HHp Smartphone	PU Smartphone	
MRR (main resp. for revenues)			HHp Other mobile	PU Other Mobile	
Belgian (or not)			HHp Landline	PU Landline	
CIM language (Nl, Fr)			HHp Portable Game Console	PU Portable Game Console	
			HHp Other Game Console	PU Other Game Console	
				PU PC/Laptop last 30d Home	
				PU PC/Laptop last 30d Work/School	
				PU Tablet last 30d	
				PU Smartphone last 30d	
				Visit Belgian Sites last 30D	

What are the sources ?

The source information differs between variables

Individual Sociodemos	Household Sociodemos	Location	Household possession (HHp)	Personal Use (PU)	Personal Frequency (PFr)
Gender	Household size	Municipality	HHp # TV	PU PC	PFr TV Viewing
Age	# Children	Arrondissement	HHp PC	PU Laptop	PFr Radio Listening
Education	Age Category Children	Province	HHp Laptop	PU Tablet	PFr Internet Use
Professional active (or not)	Social Group	Nielsen	HHp Tablet	PU Smartphone	
Profession (current or last)		CIM Habitat	HHp Smartphone	PU Other Mobile	
MS (main shopper)		CIM Inhabitant	HHp Other mobile	PU Landline	
MRR (main resp. for revenues)			HHp Landline	PU Portable Game Console	
Belgian (or not)			HHp Portable Game Console	PU Other Game Console	
CIM language (Nl, Fr)			HHp Other Game Console		
				PU PC/Laptop last 30d Home	
				PU PC/Laptop last 30d Work/School	
				PU Tablet last 30d	
				PU Smartphone last 30d	
				Visit Belgian Sites last 30D	

Sources

Fgov - Structure of Population	CIM 40k (Radio + Press + Establishment Survey)
Fgov - Workforce Survey	CIM 6K (Establishment Survey)

* Master – Bachelor – Other: Fgov; Distribution of Other educational levels = CIM 40K

What are the periods of the sources? (VP 2021)

Source	Period
Fgov	Population on 01 January 2020
CIM Radio	From May 2019 – April 2020
CIM Press	From June 2019 – May 2020
CIM Establishment Survey	From October 2018 – September 2019

Where can I find it ?

The sociodemographic description (the Golden Standard) is available on CIM.be

https://www.cim.be/nl/golden-standard

https://www.cim.be/fr/golden-standard

Currently, media device and useage data are for internal CIM purposes only

(but subscribers will find the source information in the Establishment Survey)

Is this database correct ?

- If you mean: is this an exact photo of the Belgian population, NO

  No public authority or private enterprise has all of these data

- Variables are modelled and combined using the best available data, like the full fgov data as publicly released (= aggregated reports) or the best possible combination of CIM data

  Individual lines may therefore not completely exist as such, that’s why it is called a VIRTUAL population

  But YES, aggregated results give a fair insight in the Belgian population

How is the Virtual Population made ?

If you really want to know, you will enjoy the next 15 slides!

They explain the 7 steps in the creation of the Virtual Population.

1. Hard socio-demographical variables
2. Creation of households
3. Soft socio-demographical variables
4. Household possession and personal use of devices
5. Personal use of media
6. Optimisations
7. Validation

Sources

- Hard variables : fgov
  - Fgov Structure of population : Age x Gender x Municipality (581) x HH size (5)
  - Fgov Workforce survey : Active or not, Bachelor/Master/Others

- Soft variables : CIM
  - Aggregate all CIM field studies
    (6K ES + 10K Press + 24K Radio = 40k, reweighted)
  - Reproduce in the VP the distribution of the variables (40k), in the best possible way

1. Creation of the VP with hard variables (fgov)

- Establish the number of persons on 1st January for
  - Age (0-110) x Gender (2) x Municipality (581) x HH size (5)

- Creation of 11.431.406 persons to represent each of these crosses

- In the dbase, 1 line = 1 person with variables :
  - Age, Gender, Municipality and HH size
  - Arrondissement, Province, Nielsen, CIM Habitat, CIM Inhabitant

2. Household creation

- Need to add the family level :
  - Technical help to create some variables such as Social Groups
  - Important for the description of the TV universe
  - Logical constraint for the devices (see later)

- Reproduction, in the VP, of the household structures observed in 40k
  - By municipality
  - By household size
  - Taking into account of the profile combinations Age x Gender observed in the 40k households.

3. Creation of soft variables by prediction

<u>Step 1</u> Selection of the best predictors already existing in the VP

Perform regression analysis of these variables in 40k (= determine their impact)

<u>Step 2</u> Calculation of quotas for the 3-4 best predictors

Reproduce exactly in the VP the cross-distributions observed in 40k

<u>Step 3</u> Calculation of probabilities

Determine cross probabilities of the new variable with the 3-4 predictors that follow
= take also into account of some additional good predictors, as much as possible

<u>Step 4</u> Creation of missing individual variables

Respect of remaining quotas

Random selection of a modality of the new variable on the basis of probabilities obtained in step 3

<u>Step 5</u> Creation of missing household variables (same value for each member of a family)

Constitution of household quotas and random selection of a modality on the basis of probabilities, the difference being that the probability is the average of the probabilities of the individual members

Fictive example (education)

Education	Step 2	Quotas (best predictors)			
	M 12-17	M 18-24	**M 25-34**	…	W 75+
Primary	10	5	**5**		14
Secondary	10	15	**18**		10
University	10	10	**7**		6

Imagine that age and gender are the best predictors.

Then, we will reproduce exactly the distribution of the education observed in 40k for each profile age * gender.

In the example: Man 12-17 = 1/3 Primary, 1/3 Secondary, 1/3 University observed in 40k
=> 10 – 10 – 10 in the VP (if 30 Men 12-17)

Fictive example (education)

Education	Step 2	Quotas (best predictors)				Step 3	Probabilities (good predictors)			
	M 12-17	M 18-24	**M 25-34**	…	W 75+	HH1	HH2	**HH3**	HH4	HH5+
Primary	10	5	**5**		14	10%	10%	**20%**	30%	30%
Secondary	10	15	**18**		10	50%	40%	**40%**	20%	50%
University	10	10	**7**		6	40%	50%	**40%**	50%	20%

Imagine that household size is a good predictor (but not one of the best).

Then, we will reproduce, as good as possible, the distribution of the education observed in 40k per HH size

In the example: HH3 = 1/5 Primary, 2/5 Secondary, 2/5 University observed in 40k
=> we will come the best possible closer to 6 - 12 - 12

Fictive example (education)

	Step 2		Quotas (best predictors)			Step 3	Probabilities (good predictors)			
Education	M 12-17	M 18-24	**M 25-34**	…	W 75+	HH1	HH2	**HH3**	HH4	HH5+
Primary	10	5	**5**		14	10%	10%	**20%**	30%	30%
Secondary	10	15	**18**		10	50%	40%	**40%**	20%	50%
University	10	10	**7**		6	40%	50%	**40%**	50%	20%

Step 4

P1 : Man 25-34 HH3	Primary	5	Primary	0,2	
	Secondary	18	Secondary	0,4	
	University	7	University	0,4	v

For the first man 25-34 in HH3 (selected randomly),
the education level is randomly selected according the probabilities 20% - 40% - 40% obtained in step 3

Hence, we decrease by 1 the number of men with university education, aged 25-34 category (7 → 6)

cim

3. Soft variables – Socio-Demographic

- Creation of the variables Education, Profession, Professional Status, MS, MRI, CIM Language, Children (<15 years) and Social Groups

- Use of reweighted 40k as reference

- Extra-primary constraints :
  - At least one MS and one MRI by household
  - If only one 12+ in the household → systematically MS and MRI
  - Maximum 2 MRI by household
  - If a household contains 2 MRI, they are both either active or inactive
  - Respect of quotas established in step 2, as much as possible

- Social Groups are calculated according to education and profession of the MRI

4. Devices

- From here, use of 6k from ES re-weighted as reference

- Creation of the variables TV, Smartphone, PC, Laptop, Tablet, Console, Portable Console, Other Mobile, Landline in terms of :
  - Household possession first
  - Personal use second
  - Personal use on internet the last 30 days third (except other mobile and landline)

    \+ distinction between PC/Laptop home and work
  - Constraint : avoid inconsistencies

    (e.g. possession smartphone = 0 and Use of smartphone = 1)

5. Use of TV/Internet/Radio

- Creation of the variables :
  - ➢ Frequency of TV viewing (# days/week)
  - ➢ Frequency of internet use (# days/week)
  - ➢ Visit of Belgian sites the last 30 days
  - ➢ Frequency of radio listening (# days/week)

- Use of reweighted 6k from Establishment Survey as reference

6. Optimizations

- Control of the distributions of the PREDICTORS variables in VP and 40k

  Check if the distributions are very well reproduced.

- Control of the distributions of the OTHERS variables in VP and 40k

  Check if the divergences are not too important.

Solution : if absolute and relative deviations are too large,
the corresponding variable is included either in quotas or in probabilities
(if possible)

- Example : Nielsen and Landline possession (before optimization)

		VP			6k reweighted			Difference		
		HH Landline Yes	HH Landline No	Total	HH Landline Yes	HH Landline No	Total	HH Landline Yes	HH Landline No	Total
Nielsen	Nielsen 1	15,3%	8,6%	24,0%	15,7%	8,3%	24,0%	-0,4%	0,4%	0,0%
	Nielsen 2	20,4%	11,6%	32,0%	21,0%	11,0%	32,0%	-0,6%	0,6%	0,0%
	Nielsen 3	7,2%	4,9%	12,1%	5,9%	6,2%	12,1%	1,3%	-1,3%	0,0%
	Nielsen 4	9,5%	5,8%	15,3%	9,8%	5,5%	15,3%	-0,3%	0,3%	0,0%
	Nielsen 5	10,2%	6,3%	16,5%	10,3%	6,2%	16,5%	-0,1%	0,1%	0,0%

- After optimization

		VP			6k reweighted			Difference		
		HH Landline Yes	HH Landline No	Total	HH Landline Yes	HH Landline No	Total	HH Landline Yes	HH Landline No	Total
Nielsen	Nielsen 1	15,7%	8,3%	24,0%	15,7%	8,3%	24,0%	0,0%	0,1%	0,0%
	Nielsen 2	21,0%	11,0%	32,0%	21,0%	11,0%	32,0%	0,0%	0,0%	0,0%
	Nielsen 3	6,1%	6,1%	12,1%	5,9%	6,2%	12,1%	0,1%	-0,1%	0,0%
	Nielsen 4	9,6%	5,7%	15,3%	9,8%	5,5%	15,3%	-0,2%	0,2%	0,0%
	Nielsen 5	10,3%	6,3%	16,5%	10,3%	6,2%	16,5%	-0,1%	0,1%	0,0%

Limitations

It’s important to keep in mind that :

- We cannot take into account of all predictors in quotas
- Consequently, some important predictors are used in probabilities
- The cross-distribution of a good predictor with the new variable may be different in the VP than the one observed in 40k