Final Project

Step 1: Import and pre-process HBS data from Eurostat

Eurostat provides annual data on Household Final Consumption Expenditure broken down by COICOP categories and by certain cross-sectional variables. The data is available for most EU countries and is collected ~every 5 years.

For more information on COICOP categories (Classification of Individual Consumption According to Purpose) please see here:
https://en.wikipedia.org/wiki/Classification_of_Individual_Consumption_According_to_Purpose

For your project you will use HBS data for the year 2015. Each student has to select:

1. **One** cross-sectional variable out of
    a. income quintile (“quintile”)
    b. type of household (“hhtyp”)
    c. age of the reference person (“age”)
    d. degree of urbanization (“deg_urb”)
2. **One** country

Each student needs to have a unique combination of both.

On Ilias we uploaded **two** datasets for each variable/country combination:

1. The mean consumption expenditure of private households (datasets starting with hbs_exp_). Unit: Euro
2. The structure of mean consumption expenditure (datasets starting with hbs_struc_). Unit: Per mille

Step 1: Pre-process data

1. Merge both datasets (mean consumption and structure of consumption) to calculate the **absolute** expenditures by cross-sectional variable and COICOP category.
2. Convert the expenditure data into a pandas DataFrame of shape (n x m) where n is the number of COICOP categories and m the number of different classes from your cross-sectional variable (e.g. income quintiles, age groups, etc.). For an example of what the data should look like, see below.
3. **Question 1: What COICOP category shows the highest expenditure for each different group from your cross-sectional variable? If the category differs between groups, why do you think this is the case?**
4. **Question 2: What is the total expenditure by group? How do you explain the differences between groups?**
5. Save the DataFrame in a format convenient for you (e.g. .xlsx)

Expected result: The resulting DataFrame should look like this (example for degree of urbanization, not all rows are shown):

deg_urb	DEG1	DEG2	DEG3
coicop			
CP011	2168.088	2237.025	2176.860
CP012	229.317	255.660	248.784
CP021	208.470	191.745	207.320
CP022	145.929	127.830	124.392
CP023	0.000	0.000	0.000
CP031	771.339	766.980	704.888
CP032	208.470	213.050	186.588
CP041	2251.476	1214.385	787.816
CP042	2272.323	3472.715	3773.224

Step 2: Map COICOP categories to EXIOBASE products

To be able to attach the final consumption matrix to EXIOBASE you need to map the COICOP categories to EXIOBASE products.

We have already created a correspondence table (CT) that shows which COICOP categories correspond to which EXIOBASE products in which region (file: `CT_coicop_exiobase.xlsx`). The CT is formatted as a matrix with the COICOP categories on the rows and the EXIOBASE products and regions on the columns. The number 1 means that there is a correspondence, a zero indicates no correspondence.

However, instead of just binary information on if there is a correspondence or not, what we want is quantitative information on what **share** of the expenditure of one COICOP category corresponds to which EXIOBASE product in which region. For example, we want to know what share of the total expenditure for category c01.1 (Food) shall be allocated to which of the EXIOBASE food sectors (Paddy rice, Wheat, …) in which region.

The information on which share of total expenditure by COICOP category shall be allocated to which EXIOBASE sector/region, we take from EXIOBASE’s Y-matrix.

Please follow these steps and answer the questions:

1. Fill CT with quantitative information on shares (see above)
    a. Read the CT into python
    b. Question: What is the shape of the CT? Are there more COICOP categories than EXIOBASE products or vice versa? What does this mean for the mapping?
    c. Parse EXIOBASE3 for the year 2015 in the product-by-product variant (pxp). If you haven’t downloaded it yet, you first need to download it.
    d. Extract the Final Demand (Y) for the “Final consumption expenditure by households” for your country of choice (the one you also have downloaded the HBS data for).
    e. Now, you have to create a new CT of the same shape as the binary CT, but filled with the respective shares instead. Note, that each row should sum to 1. See example below.
    f. **Question: Why is it important that each row of the filled CT sums to one?**

Example for sub-step 1e: CT mapping [A,B,C] to [a,b,c,d].

binary CT:

	a	b	c	d
A	1	0	0	1
B	0	0	1	0
C	0	1	0	0

Final demand vector:

a	1
b	2
c	3
d	4

filled CT:

	a	b	c	d
A	0.2	0.0	0.0	0.8
B	0.0	0.0	1.0	0.0
C	0.0	1.0	0.0	0.0

Expected result of sub-step 1:

region										AT	...
sector	Paddy rice	Wheat	Cereal grains nec	Vegetables, fruit, nuts	Oil seeds	Sugar cane, sugar beet	Plant-based fibers	Crops nec	Cattle	Pigs	...
coicop											
c01.1	0	0.000093	0.000057	0.000664	0.000161	0	0	0.000102	0	0	...
c01.2	0	0.000146	0.000089	0.001042	0.000252	0	0	0.000160	0	0	...
c02.1	0	0.000000	0.000000	0.000000	0.000000	0	0	0.000000	0	0	...
c02.2	0	0.000000	0.000000	0.000000	0.000000	0	0	0.000000	0	0	...
c02.3	0	0.000179	0.000110	0.000000	0.000310	0	0	0.000196	0	0	...
c03.1	0	0.000000	0.000000	0.000000	0.000000	0	0	0.000000	0	0	...

2. Then, we combine both, the HBS data from Step 1 and the filled CT:
    a. Load the HBS from step 1 into python.
    b. Use the power of linear algebra to map your HBS from the COICOP classification to EXIOBASE products/regions (NB: any other method is totally ok. Just the results need to be correct). The result is a final demand matrix with the EXIOBASE regions/sectors on the rows and the socio-economic groups on the columns (see below).
    c. Calculate the total expenditure by group (of your cross-sectional variable) for both (1) the original HBS data in COICOP classification, and (2) the expenditure data mapped to EXIOBASE sectors. Both should deliver the same results!
    d. **Question: What EXIOBASE sector-region pair shows the highest expenditure for each different group from your cross-sectional variable? If the category differs between groups, why do you think this is the case?**
    e. Save your resulting Final demand matrix in a format convenient for you.

Expected results of sub-step 2 (example):

region	sector	DEG1	DEG2	DEG3
AT	Paddy rice	0.000000	0.000000	0.000000
	Wheat	0.407748	0.408888	0.384162
	Cereal grains nec	0.248907	0.249603	0.234509
	Vegetables, fruit, nuts	2.100742	2.130526	2.051978
	Oil seeds	0.705050	0.707021	0.664266
...	...	...	...	...
WM	Membership organisation services n.e.c. (91)	0.000000	0.000000	0.000000
	Recreational, cultural and sporting services (92)	0.000000	0.000000	0.000000
	Other services (93)	0.000000	0.000000	0.000000
	Private households with employed persons (95)	0.000000	0.000000	0.000000
	Extra-territorial organizations and bodies	0.000000	0.000000	0.000000

9800 rows × 3 columns

Step 3: Calculate footprints

Now, you are ready to calculate footprints!

1. Parse EXIOBASE and calculate all missing elements.
2. Load the final demand matrix created in step 2.
3. Select 4 different **impact** categories (such as “GHG emissions (GWP100) | Problem oriented approach: baseline (CML, 2001) | GWP100 (IPCC, 2007)” etc.). Filter the S-matrix for those rows.
4. Calculate footprints:
    a. Total footprints by group
        i. Plot your results using an adequate plot type.
        ii. **Question: Does the total footprint substantially differ between the different groups? If yes, how do you explain the differences between the groups?**
        iii. *Bonus: Calculate the total national footprints by summing over the groups and compare the results to the EXIOBASE national footprint. Which of the two is higher? Why?*
    b. Footprints by source sector (&source region) and group
        i. Rank the source sectors by their contribution (for each group)
        ii. Rank the source regions by their contribution (for each group)
        iii. **Question: Which are the 3 <u>sectors</u> with the highest contribution per group? If it differs between groups, how do you explain the differences?**
        iv. **Question: Which are the 3 <u>regions</u> with the highest contribution per group? If it differs between groups, how do you explain the differences?**
        v. Plot your results using an adequate level of aggregation (or level of filtering) and an adequate plot type.
        vi. *Bonus: Show a piece of results from your breakdown by source sectors that you find interesting and interpret it.*
    c. Footprints by final product (&country of origin) and group
        i. Rank the final products by their contribution (for each group)
        ii. Rank the country of origin by their contributions (for each group)
        iii. **Question: Which are the 3 <u>final products</u> with the highest contribution per group? If it differs between groups, how do you explain the differences?**
        iv. **Question: Which are the 3 <u>countries of origin</u> with the highest contribution per group? If it differs between groups, how do you explain the differences?**
        v. Plot your results using an adequate level of aggregation (or level of filtering) and an adequate plot type.
        vi. *Bonus: Show a piece of results from your breakdown by final product that you find interesting and interpret it.*