Documentation Programme Indépendant

0. Principe et utilité

Le fichier de code indépendant permet de produire une base de données agrégées complète qui contient les valeurs à représenter dans l'application Shiny, pour toutes les dimensions :

- Données pour tous les niveaux NUTS (0, 1, 2)
- Données pour toutes les années (si celles-ci sont disponibles)
- Données pour toutes les périodes considérées (si celles-ci sont disponibles). Les périodes sont construites de façon à couvrir la période entre deux changements de la norme NUTS[1]. On a actuellement :
    - 2004 à 2005
    - 2006 à 2009
    - 2010 à 2012
    - 2013 à 2015
- Données pour tous les pays de l'Union Européenne (si celles-ci sont disponibles).

La liste des variables retenues sera détaillée plus tard.
Le pourcentage de valeurs manquantes est également disponible dans cette base de données.
Ceci permet de répertorier toutes les données nécessaires pour le fonctionnement de l'application sans avoir à reprendre les données individuelles et exécuter des calculs d'agrégation.

Ce programme n'a vocation d'être exécuté qu'à chaque nouvelle enquête récoltée.

1. Paramètres initiaux globaux

Il est demandé au détenteur des données individuelles de rentrer 4 options au préambule du programme :

- *importer_carte* : 0 afin de ne pas importer les données géographiques, 1 sinon. Cet option sert uniquement à économiser du temps si ces données sont déjà chargées dans la session R.
- *precision* (nécessite *importer_carte* = 1) : s'il est demandé d'importer les données géographiques, la précision des polygones tracés peut être choisi :
    - tracé précis mais moins rapide : *precision* = 1;
    - tracé moins précis mais plus rapide : *precision* = 60.
- *date_premiere_enquete* : année en format numérique YYYY de la première enquête à considérer.
- *date_derniere_enquete* : année en format numérique YYYY de la dernière enquête à considérer.

Ces deux dernières valeurs sont importantes pour le bon déroulement de la création du fichier de données.

[1] Historique des changements de la norme : https://ec.europa.eu/eurostat/fr/web/nuts/history

2. Importations

2.1. Dictionnaires correctifs

Plusieurs “dictionnaires” (en réalité des fichiers .csv avec deux colonnes) sont importés. Ceux-ci sont utiles pour relever les exceptions (dans le sens péjoratif) dans les données individuelles afin de pouvoir les traiter. Par conséquent, si de nouvelles exceptions apparaissent dans de futures enquêtes, il faudra suivre le même procédé pour corriger les incohérences.

2.2. Listes de variables sélectionnées

Plusieurs listes sont importées. Elles répertorient les variables retenues (fichier liste_variable.csv dans le jeu dico_variable_import) pour l’application et distinguent les variables binaires (fichier liste_variable_binaire.csv dans le jeu liste_variable_binaire) et les variables qualitatives (à plus de 2 modalités) (fichier liste_variable_quali.csv dans le jeu liste_variable_quali).

2.3. Données géographiques

Les données géographiques sont importées si demandé. Les polygones représentant les régions NUTS0, NUTS1 et NUTS2 et les segments représentant les frontières des pays (donc NUTS0) pour les différentes normes précédentes et celle en vigueur sont importés.

2.4. Données des enquêtes

Les enquêtes étudient d’un côté les ménages, et de l’autre les individus. Il n’y a pas de lien entre ces deux entités (on ne peut pas connecter un individu à un ménage qui serait le sien). Par conséquent, les données à importer prennent deux directions différentes.

Avant d’étudier le processus d’importation, il est important de connaître la structure des dossiers contenant les données :

a. L’ensemble des fichiers (des archives .zip) sont dans un même répertoire ./data/enquete
b. Chaque fichier zip correspond aux données d’un pays pour une année en particulier. Les noms de ces archives zip sont harmonisés et contiennent le code iso2 du pays et l’année en question. Par exemple, on a : “ES_2011_EUSILC.zip” pour les données de l’Espagne en 2011.
c. À l’intérieur de chaque fichier zip, on retrouve 4 fichiers .csv. Ce sont les 4 bases qui structurent l’enquête (registre des ménages (h), données des ménages (d), registre des individus (r), données des individus(p)). Ces 4 fichiers sont également nommés de façon harmonisée. Par exemple, on a, au sein de l’archive “ES_2011_EUSILC.zip”, 4 fichiers du type “ES_2011x_EUSILC.csv”, avec x qui peut être d, h, p ou r.

La procédure d’importation est la même pour les deux cas et est la suivante :

1. Initialiser les 2 tables qui vont contenir les données et qui vont être complétées au fur et à mesure des importations.
2. Procéder à une année (par exemple 2011) à la fois :
    a. Filtrer et Identifier les archives qui correspondent à l'année en cours
    b. Procéder à un archive valide (par exemple ES_2011_EUSILC.zip) à la fois :
        i. Importer les 4 fichiers .csv
    c. Joindre la table (r) avec (p) et la table (h) avec (d)
    d. Faire des traitements (section suivante)
    e. Concaténer avec la table initiale (remplissage - mise à jour)

L'harmonisation des noms de fichiers permet une importation automatique de l'ensemble des fichiers de données. De plus, R permet l'importation des fichiers .csv détenus dans les archives .zip sans avoir à dézipper ces derniers, ce qui permet un gain de temps et d'espace de stockage non négligeable.

3. Traitements

Après jointure des 4 fichiers pour n'en avoir que 2, on traite directement les données. Les jointures se font grâce à l'identifiant (du ménage ou de l'individu), de la région et de l'année. L'identifiant seul ne suffit pas parce qu'un même ménage garde le même identifiant entre deux enquêtes.

3.1. Recodage des modalités

Les valeurs prises par les variables peuvent être incohérentes. Il est nécessaire de nettoyer ces valeurs. Ces exceptions ne sont pas les mêmes selon qu'il s'agisse d'une variable qualitative ou d'une variable binaire.
À partir des listes de variables (qui distinguent les variables binaires aux variables qualitatives), on peut traiter les deux cas distinctement :

- Les variables binaires sont codées en
    - 1 : Yes 2 : No.

  Il est nécessaire de recoder en
    - 1 : Yes 0 : No.

  Il n'y a pas de valeurs parasites (valeurs autre que 1 ou 2 avant recodage) dans les données actuelles.
- Les variables qualitatives présentent des valeurs parasites :
    - "" (chaine vide),
    - "0" (le 0 n'est jamais une modalité pour les variables qualitatives, à l'exception d'une variable qui est traitée à part),
    - "0-1", "0 - 1" et "NAs".

  Ces valeurs sont inexploitables, et sont donc considérées comme valeur manquante (NA).

3.2. Tableau disjonctif complet pour les variables qualitatives

Maintenant que les variables qualitatives prennent uniquement des valeurs qui ont du sens, il est possible de construire le tableau disjonctif complet pour chacune de ces variables afin d'étudier chaque modalité séparément.

3.3. Gestion des erreurs/incohérences au niveau des codes régionales NUTS

Des exceptions (erreurs, mauvaises normes) sont retrouvées dans la variable indiquant la région NUTS2. Celles-ci empêchent d'associer l'individu statistique (ménage ou individu) à la bonne région.

Il est important de noter que la norme NUTS change régulièrement (2003, 2006, 2010, 2013, 2016) et donc une région correctement codée dans une année T peut être incorrecte dans une autre année. De plus, les codes NUTS1 et NUTS0 sont déduites des codes NUTS2 (FR11 devient FR1 puis FR par exemple). Donc une erreur au niveau du code NUTS2 induit souvent une erreur au niveau NUTS1 puis NUTS0. Ainsi, pour procéder à la correction de ces exceptions, il faut gérer les erreurs en partant du NUTS2 pour arriver à NUTS0 tout en considérant la norme NUTS en vigueur.

L'association des "fausses" et "bonnes" normes a été faite à partir de superposition de cartes (d'un côté nous avons une carte régionale avec les normes de l'enquête et de l'autre, nous avons une carte régionale avec les normes cartographiques).

3.3.1. NORME NUTS2

Problème	Solution
Croatie : La norme NUTS 2 de la carte en 2010 change et ne correspond pas à celle utilisée dans l'enquête. C'est à dire qu'il y a un regroupement de deux régions en une.	Nous avons utilisé un dictionnaire pour ajuster la norme utilisée dans l'enquête à celle utilisée pour afficher la carte. Nous remplaçons les valeurs : HR01 -> HR04 HR02 -> HR04 Fichier : `dico_enquete_croatie_apres_2010` Modification effectuée avant tout agrégat.
Royaume-Uni (Londres) : La norme NUTS 2 de la carte en 2013 change et ne correspond pas à celle utilisée dans l'enquête. Londres est divisée en plusieurs région. UKI1, UKI2	Nous avons regardé la correspondance dans la norme NUTS2 en 2013. Nous avons regroupé les deux régions en une UKI1 -> UKIX UKI2 -> UKIX Fichier : `dico_enquete_uk_apres_2013` Modification effectuée avant tout agrégat. Après calcul dans la région UKIX, nous

Grèce : Avant 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. La carte contient des GR. L'enquête contient des EL. Ceci n'est nécessaire qu'avant 2010.	Utilisation d'un dictionnaire avant tout agrégat. EL51 -> GR11 EL52 -> GR12 EL53 -> GR13 EL61 -> GR14 EL54 -> GR21 EL62 -> GR22 EL63 -> GR23 EL64 -> GR24 EL65 -> GR25 EL30 -> GR30 EL41 -> GR41 EL42 -> GR42 EL43 -> GR43 Fichier : `dico_codage_avant_2010`
Grèce : Après 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. Les chiffres derrière l'indice du pays ne sont pas bons. Ceci n'est nécessaire qu'après 2010.	Utilisation d'un dictionnaire avant tout agrégat. EL51 -> EL11 EL52 -> EL12 EL53 -> EL13 EL61 -> EL14 EL54 -> EL21 EL62 -> EL22 EL63 -> EL23 EL64 -> EL24 EL65 -> EL25 Fichier : `dico_codage_apres_2010`
Italie : Avant 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. La carte contient des ITD et ITE. L'enquête contient des ITH et ITI.	Utilisation d'un dictionnaire avant tout agrégat. ITH1 -> ITD1 ITH2 -> ITD2 ITH3 -> ITD3 ITH4 -> ITD4 ITH5 -> ITD5 ITI1 -> ITE1 ITI2 -> ITE2 ITI3 -> ITE3 ITI4 -> ITE4 Fichier : `dico_codage_avant_2010`
Bulgarie : Avant 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. Les chiffres derrière l'indice du pays ne sont pas bons.	Utilisation d'un dictionnaire avant tout agrégat. BG01 -> BG31 BG02 -> BG32 BG03 -> BG33 BG04 -> BG41 BG05 -> BG42 BG06 -> BG34

	Fichier : `dico_codage_avant_2010`
Roumanie : Avant 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. Les chiffres derrière l'indice du pays ne sont pas bons.	Utilisation d'un dictionnaire avant tout agrégat. RO01 -> RO21 RO02 -> RO22 RO03 -> RO31 RO04 -> RO41 RO05 -> RO42 RO06 -> RO11 RO07 -> RO12 RO08 -> RO32 Fichier : `dico_codage_avant_2010`
Finlande : Avant 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. Les chiffres derrière l'indice du pays ne sont pas bons.	Utilisation d'un dictionnaire avant tout agrégat. FI1D -> FI13 FI1B -> FI18 FI1C -> FI1A Fichier : `dico_codage_avant_2010`

3.3.2. NORME NUTS1

Problème	Solution
Italie, Bulgarie, Roumanie, Finlande	Les modifications faites en NUTS2 permettent d'avoir les bonnes normes.
Grèce : Avant 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. La carte contient des GR. L'enquête contient des EL. Ceci n'est nécessaire qu'avant 2010.	Utilisation d'un dictionnaire avant tout agrégat. EL51 -> GR11 EL52 -> GR12 EL53 -> GR13 EL61 -> GR14 EL54 -> GR21 EL62 -> GR22 EL63 -> GR23 EL64 -> GR24 EL65 -> GR25 EL30 -> GR30 EL41 -> GR41 EL42 -> GR42 EL43 -> GR43 Fichier : `dico_codage_avant_2010` Ensuite dans le code, pour la norme NUTS1, nous couperons manuellement l'indice région pour ne garder que les 3 premiers termes qui correspondront à

	l'indice de la région NUTS1. GR11 -> GR1 GR12 -> GR1 etc… pour toutes les régions Ainsi, nous pourrons effectuer un agrégat, prenant en compte les valeurs de toute la région NUTS1.
Grèce : Après 2010, la norme de l'enquête ne correspond pas à celle de l'enquête. Les chiffres derrière l'indice du pays ne sont pas bons. Ceci n'est nécessaire qu'après 2010.	Utilisation d'un dictionnaire avant tout agrégat. EL51 -> EL11 EL52 -> EL12 EL53 -> EL13 EL61 -> EL14 EL54 -> EL21 EL62 -> EL22 EL63 -> EL23 EL64 -> EL24 EL65 -> EL25 Fichier : `dico_codage_apres_2010` Ensuite dans le code, pour la norme NUTS1, nous couperons manuellement l'indice région pour ne garder que les 3 premiers termes qui correspondront à l'indice de la région NUTS1. EL11 -> EL1 EL12 -> EL1 etc… pour toutes les régions Ainsi, nous pourrons effectuer un agrégat, prenant en compte les valeurs de toute la région NUTS1.

3.3.3. NORME NUTS0

Problème	Solution
Allemagne : L'Allemagne n'a pas de région enregistré.	Nous affichons l'Allemagne qu'en NUTS 0. Les données sont sur l'ensemble du pays. Nous ne pouvons pas agréger les données par région.
Grèce : Avant 2010, il s'agit d'un cas particulier. En effet, la norme des cartes avant 2010 est correcte et correspond à celle des enquêtes. Mais à cause des modifications précédentes (NUTS2), nous avons changé :	[Voir modification à l'étape NUTS2] -> Résultat : EL51 -> GR11 EL52 -> GR12 EL53 -> GR13

EL en GR. Il est donc nécessaire d'ajuster à nouveau la norme : GR -> EL. Nous avons fait ce choix de manière à optimiser le code.	… Suite aux modifications précédentes, nous devons, avant 2010, ajuster la norme de la Grèce : GR11 -> EL

4. Variables sélectionnées

La sélection des variables a été faite en amont du code. On ne garde que les variables qui nous intéressent, c'est-à-dire la REGION, celles retenues en amont et celles construites par la disjonction des variables qualitatives. On aura alors que des variables d'identification, quantitatives ou binaires 0-1. Les variables qui indiquent l'année et le niveau NUTS sont construites en fin d'itération de boucle. Ces deux variables sont simplement déduites de l'itération en cours (pour rappel, le traitement des données se fait avec double boucle, l'une sur l'année en cours et l'autre sur le niveau NUTS étudié).

5. Calcul des agrégats

5.1. Calcul des moyennes/effectifs par région et par année

Maintenant qu'on a deux tables correctement structurées et sans valeurs indésirables, on peut calculer les moyennes pour chaque variable (proportions pour les variables binaires).
En parallèle, la proportion de NA est également calculée pour chaque variable.
Ces calculs sont faits pour chaque variable dans chaque année et dans chaque région (toutes échelles confondues).

5.2. Calcul des moyennes/effectifs par région et par période

Dans la même logique, on calcule les moyennes et proportions de valeurs manquantes pour toutes les régions (à différentes échelles) pour les périodes possibles. En marge, on construit également une variable ANNEES_PRESENTES pour indiquer quelles années sont considérées dans la moyenne de la période. Pour illustration fictive, l'Espagne peut n'avoir qu'une seule enquête faite (2007) dans la période 2006-2009, mais une moyenne sera quand même calculée. L'information détenue dans ANNEES_PRESENTES permet d'informer l'utilisateur de la signification de la valeur qu'il observe.

Ces deux tables calculées (l'une pour les années, l'autre pour les périodes) sont rassemblées dans une même table (concaténation).

6. Fichier final

Au bout de l'étape 5, on obtient deux tables, une pour les ménages et l'autre pour les individus. Puisqu'en finalité, l'individu statistique est une région européenne, il est

raisonnable de joindre ces deux tables pour avoir une table centrale structurée en 4 morceaux.

Ce fichier final contient toutes les valeurs nécessaires pour alimenter l’application Shiny sans avoir à recalculer les agrégations.