

TP2 Analyse Bivariée en Python

2024-2025

0.1 Les données

0.1.1 Revenus, pauvreté et niveau de vie en 2019 - Données carroyées Dispositif Fichier localisé social et fiscal (Filosofi)

Les données proviennent de l'INSEE ([url](#)).

Ce sont des données qui décrivent de nombreuses informations socio-économiques, et qui sont quadrillés sur des carreaux de 200m par 200m (ou 1km par 1km).

Néanmoins, comme ces données peuvent être à caractère personnel, lorsqu'un carreau comprend moins de 11 ménages, les données sont « **imputées** » : plusieurs carreaux de moins de 11 ménages sont regroupés jusqu'à obtenir au moins 11 ménages, puis la moyenne pondérée des variables est affectée à chaque carreau.

Elles sont disponibles en plusieurs formats (SHP, GPKG, CSV)

Les variables du jeu de données sont :

- *idcar_1km* : Identifiant Inspire du carreau de 1 km
- *i_est_1km* : Vaut 1 si le carreau est imputé par une valeur approchée, 0 ou 2 sinon.
- *lcog_geo* : Code officiel géographique au 1er janvier 2020 de la ou des commune(s) dans laquelle (lesquelles) se trouve le carreau.
- *ind* : Nombre d'individus
- *men* : Nombre de ménages
- *men_pauv* : Nombre de ménages pauvres
- *men_1ind* : Nombre de ménages d'un seul individu
- *men_5ind* : Nombre de ménages de 5 individus ou plus
- *men_prop* : Nombre de ménages propriétaires
- *men_fmp* : Nombre de ménages monoparentaux
- *ind_snv* : Somme des niveaux de vie winsorisés des individus
- *men_surf* : Somme de la surface des logements du carreau
- *men_coll* : Nombre de ménages en logements collectifs
- *men_mais* : Nombre de ménages en maison
- *log_av45* : Nombre de logements construits avant 1945
- *log_45_70* : Nombre de logements construits entre 1945 et 1969
- *log_70_90* : Nombre de logements construits entre 1970 et 1989
- *log_ap90* : Nombre de logements construits depuis 1990
- *log_inc* : Nombre de logements dont la date de construction est inconnue
- *log_soc* : Nombre de logements sociaux
- *ind_0_3* : Nombre d'individus de 0 à 3 ans
- *ind_4_5* : Nombre d'individus de 4 à 5 ans
- *ind_6_10* : Nombre d'individus de 6 à 10 ans
- *ind_11_17* : Nombre d'individus de 11 à 17 ans
- *ind_18_24* : Nombre d'individus de 18 à 24 ans
- *ind_25_39* : Nombre d'individus de 25 à 39 ans
- *ind_40_54* : Nombre d'individus de 40 à 54 ans
- *ind_55_64* : Nombre d'individus de 55 à 64 ans
- *ind_65_79* : Nombre d'individus de 65 à 79 ans

- *ind_80p* : Nombre d'individus de 80 ans ou plus
 - *ind_inc* : Nombre d'individus dont l'âge est inconnu
- Recherchez dans la documentation et expliquez le mot « winsorisé »

0.1.2 Données d'occupation du sol (OCS GE)

L'OCS GE est une base de données de référence pour la description de l'occupation du sol de l'ensemble du territoire métropolitain et des départements et régions d'outre-mer (DROM). Elle est produite à partir des prises de vues aériennes, des données existantes extraites des bases de l'IGN, et de toutes autres données mobilisables issues de référentiels nationaux ou locaux.

Elle couvre deux aspects :

- La couverture du sol : une description physique des matériaux du territoire
- L'usage du sol : une description des activités socio-économiques

Elle est disponible à l'adresse <https://geoservices.ign.fr/ocsgc>.

Choisissez le département et l'année de votre choix parmi ceux disponibles de la nouvelle génération.

Les variables disponibles sont :

- *ID* : Identifiant de l'objet surfacique
- *CODE_CS* : Code de couverture du sol
- *CODE_US* : Code d'usage du sol
- *MILLESIME* : Année de référence de l'objet surfacique
- *SOURCE* : Type de données sources
- *OSSATURE* : Appartenance de l'objet surfacique à « l'Ossature ».
- *ID_ORIGINE* : Identifiant « Ossature »
- *CODE_OR* : Code « Ossature »

Les classes de couverture et d'usage sont les suivantes :

CS 1 Sans végétation	CS 1.1 Surfaces anthropisées	CS 1.1.1 Zones imperméables	CS 1.1.1.1 Zones bâties	US 1 Production primaire		US 1.1 Agriculture
			CS 1.1.1.2 Zones non bâties			US 1.2 Sylviculture
		CS 1.1.2 Zones perméables	CS 1.1.2.1 Matériaux minéraux			US 1.3 Activité d'extraction
			CS 1.1.2.2 Matériaux composites			US 1.4 Pêche et aquaculture
	CS 1.2 Surfaces naturelles		CS 1.2.1 Sols nus			US 1.5 Autres prod. primaires
			CS 1.2.2 Surfaces d'eau	US 235 Production secondaire, tertiaire et usage résidentiel		US 235 Production secondaire, tertiaire et usage résidentiel
			CS 1.2.3 Nèves et glaciers			US 4.1.1 Routier
		CS 2.1.1 Formations arborées	CS 2.1.1.1 Feuillus	US 4 Réseaux de transport et d'utilité publique, logistique	US 4.1 Réseaux de transport	US 4.1.2 Ferré
			CS 2.1.1.2 Conifères			US 4.1.3 Aérien
			CS 2.1.1.3 Mixte			US 4.1.4 Navigable
CS 2 Avec végétation	CS 2.1 Végétation ligneuse		CS 2.1.2 Formations arbustives, sous-arbrisseaux			US 4.1.5 Autres
			CS 2.1.3 Autres formations ligneuses			US 4.2 Services logistiques et de stockage
	CS 2.2 Végétation non ligneuse		CS 2.2.1 Formations herbacées	US 6 Autre usage		US 4.3 Réseau d'utilité publique
			CS 2.2.2 Autres formations non ligneuses			US 6.1 Zones en transition
						US 6.2 Zones abandonnées
						US 6.3 Sans usage
						US 6.6 Inconnu

1 Analyse bivariée entre deux variables quantitatives

1.1 Choisissez deux variables quantitatives du jeu de données Filosofi

Étapes :

- Choisir deux variables du jeu de données Filosofi, pour lesquelles le lien entre ces variables vous semble intéressant. Vous pouvez éventuellement créer de nouvelles variables en combinant d'autres (e.g. proportion d'individus de moins de 25 ans). Si vous ne savez pas choisir, choisissez deux variables au hasard.
- Quelle est la première chose à faire avec ces données ?
- Y a-t-il des données aberrantes ? Faut-il filtrer ces données ?
- Semble-t-il exister une relation entre les données ?

1.2 Calculer la corrélation entre les deux variables

Étapes :

- Calculer les corrélations de Pearson et de Spearman entre les deux variables.
- Existe-t-il une corrélation ? Dans quelle sens ? Est-elle significative ?
- Commenter

1.3 Régression entre les variables

Étapes :

- Choisissez une variable à inférer en fonction de l'autre
- Y a-t-il besoin de transformer une des variables ?
- Calculer et afficher la droite de régression.
- Commenter la validité de la régression.

2 Analyse bivariée entre deux variables qualitatives

Les distributions de *CODE_CS* et *CODE_US* sont-elles indépendantes ?

Faites un test du Chi 2 pour le vérifier.

3 Analyse bivariée entre une variable quantitative et une variable qualitative

Explorer le lien entre la population et l'usage dans un département.

Étapes :

- Faites une jointure spatiale pour calculer la population par polygone de l'OCS GE.
- Affichez un boxplot de la variable *ind* en fonction des classes d'usage.
- Commentez.