UNIVERSITÉ DE NANTES

IAE NANTES
ÉCONOMIE & MANAGEMENT

EKAP
Master Économétrie Appliquée

Master Économétrie et Statistiques, parcours Économétrie Appliquée

Mémoire de Master 2

Analyse de la situation open data en France :

Quels facteurs influencent l'ouverture de données des territoires ?

THIERRY Diane

Sous la direction de Muriel TRAVERS et Mathieu MOREY

Août 2021

1 Remerciements

Avant de commencer ce mémoire, je tiens à remercier sincèrement tous ceux qui m'ont aidé, de manière directe ou indirecte à la rédaction de cette étude que j'ai eu la chance de réaliser en partie pendant mon stage, comme un sujet de recherche en interne pour l'entreprise.

Je tiens tout d'abord à remercier Mr. Darné pour ces 2 années de master, les enseignements et son aide dans la recherche de stage, qui m'a permis de postuler dans la coopérative Datactivist, sans laquelle je n'aurai pas été amenée à analyser ce thème fort intéressant.

Je voudrais ensuite remercier Mme Travers pour la formation économétrique solide, enseignée tout au long du master, ainsi que sa disponibilité pour toute aide dans les projets. Merci pour le temps dédié à l'évaluation de ce travail.

Un grand merci à Mathieu, mon maître de stage, grâce à qui les 5 mois se sont si bien déroulés, merci pour la disponibilité, la patience, la rigueur et l'accompagnement notamment dans cette analyse qui représentait un long travail de composition.

J'aimerais aussi exprimer ma gratitude à Joël et Samuel, cofondateurs de Datactivist, pour leur confiance accordée et leur soutien tout au long du projet.

Merci à Anne-Laure du temps consacré pour comprendre l'outil wikidata ; pour le partage et la patience. Merci également à tous les Datactivistes, pour leur réactivité face à mes questions et leur partage d'expertise sur le sujet, qui m'ont été très utiles pour mener à bien ce travail.

Enfin, je tiens à remercier Teodoro pour son éternel soutien dans ce que j'entreprends, son aide précieuse et ses conseils affûtés.

2 Résumé

Cette étude a pour objectif d'analyser la situation open data en France en 2021. Plus précisément, nous cherchons les déterminants de l'ouverture de données des territoires, en axant l'analyse économétrique sur les départements. Nous disposons pour cela de 17 variables explicatives du nombre de publications open data recensé, pouvant être groupées autour des thématiques suivantes : économie, politique, démographie et géographie. L'analyse exploratoire révèle l'importance de variables socio-économiques telles que la part des diplômés d'un BAC+5 ou plus dans la population, le niveau de densité, la couleur politique, ou de manière plus générale le dynamisme et le niveau d'activité. Nous avons ensuite cherché à modéliser le nombre de publications en incluant les variables non corrélées entre elles et retenues par les méthodes de sélection de variables. Dans un premier temps nous avons estimé le modèle linéaire généralisé qu'est la régression de Poisson, étant dans le cas d'une variable de comptage nécessairement positive et entière. Cependant, nous avons vu qu'il existe une sur-dispersion importante dans la série, rendant invalide ce type de modèle. Sachant que 36% des départements n'ouvrent pas de données, nous estimons dans un second temps plusieurs modèles qui prennent en compte l'inflation en zéro, par la loi de Poisson ou par la Binomiale Négative. Finalement, le modèle que nous gardons est un *double hurdle* en régression de Poisson, qui a l'avantage de décomposer Y en une partie binaire (ouvre des données ou non) et en une partie de comptage (nombre de publications). Les résultats montrent qu'un département qui a au moins 5.65% de diplômés dans sa population a une probabilité d'ouvrir ses données 5.09 fois supérieure à un département ayant moins de 5.65% de diplômés. De même, un département politiquement à gauche a 1.56 fois plus de publications qu'un département de droite. Enfin, un département ayant 5.65% de diplômés ou plus, a 2.33 fois de publications qu'un département ayant moins de 5.65% de diplômés.

Mots clefs : open data, publications, OODT, départements, R, GLM, comptage.

3 Abstract

The aim of this paper is to analyze the open data situation in France in 2021. More specifically, we are looking for the determinants of the opening up of data from the territories, by focusing the econometric analysis on the departments. To do so, we have 17 explanatory variables for the number of open data publications, which can be grouped into 4 themes : economics, politics, demographics and geographics. The exploratory analysis reveals the importance of socio-economic variables such as the proportion of graduates with a BAC + 5 or more in the population, the level of density, the political persuasion, or more generally the dynamism and the activity level. We then sought to model the number of publications by including the uncorrelated variables and those retained by the selection methods. First, we estimated a Poisson regression from the generalized linear model, given that we work with a positive and integer counting variable. However, we have seen that there is a significant over-dispersion in the series, making this type of model invalid. Knowing that 36% of the departments do not open data, we then estimate several models that take into account zero inflation, by Poisson's or by Negative Binomial's law. Finally, the model we keep is a *double hurdle* with Poisson regression, which has the advantage of decomposing Y into a binary part (open data or not) and a counting part (number of publications). The results show that a department which has at least 5.65% of graduates in its population has a probability of opening its data 5.09 times higher than a department with less than 5.65% of graduates. Likewise, a politically left-wing department has 1.56 more publications than a right-wing department. Finally, a department with 5.65% of graduates or more has 2.33 more publications than a department with less than 5.65% of graduates.

Key words : open data, publications, OODT, departments, R, GLM, counting.

Sommaire

4 Liste de sigles

OD : open data

CADA : commission d'accès aux documents administratifs

OCDE : organisation de coopération et de développement économiques

OODT : observatoire open data des territoires

EPCI : établissements publics de coopération intercommunale

CA : communauté d'agglomération

CU : communauté urbaine

CC : communauté de communes

INSEE : institut national de la statistique et des études économiques

OFGL : observatoire des finances et de la gestion publique locales

RNE : répertoire national des élus

CSP : catégorie socio-professionnelle

COG : code officiel géographique

CART : classification and regression tree

ACP : analyse en composantes principales

GLM : modèles linéaires généralisés

ZIP : zero-inflated Poisson

ZINB : zero-inflated Negative Binomial

HP : hurdle Poisson

HNB : hurdle Negative Binomial

5 Introduction

En 2021 selon l'observatoire open data des territoires (OODT), seuls 11% des collectivités locales concernées par la loi **République Numérique** ouvrent réellement leurs données.

Cette loi, aussi appelée "loi Lemaire" et promulguée le 7 octobre 2016, a changé le rapport qu'ont les territoires au numérique. Effectivement, cette dernière spécifie l'obligation d'ouvrir les données pour les administrations et collectivités locales de plus de 3.500 habitants ou de plus de 50 agents.

Pourquoi cette loi, et qu'implique-t-elle réellement ? Pour répondre à ces questions, revenons sur l'origine et les fondements de l'**open data** (OD) en France. Depuis 1789, dans l'article 15 de la Déclaration des droits de l'Homme et du citoyen, on peut lire :« La Société a le droit de demander compte à tout Agent public de son administration. », ce qui établit déjà le **droit d'accès à l'information**. Par la suite, c'est en 1978 que la CADA, commission d'accès aux documents administratifs, pose véritablement les bases de la **transparence administrative** en France, avec la "loi CADA". Celle-ci institue le droit à toute personne qui le souhaite, d'**accéder** aux documents administratifs et même de les **reproduire**, grâce aux photocopies en plein essor en ce début des années 80. [1]

Le développement d'Internet dans les années 90 facilite la création et la diffusion des données publiques, questionnant alors sur la **réutilisation** de celles-ci. [2] En effet, de plus en plus, les utilisateurs attendent que les données et les informations soient publiées en ligne par défaut. Ce n'est pourtant pas la logique des organisations susceptibles de publier, qui s'interrogent sur les usages qui pourraient être faits des données ouvertes. Cette **crise de confiance** prend fin en 2009 pendant une conférence TED, lorsque l'inventeur du web Sir Tim Berners-Lee exprime son désir et celui des usagers :« We want raw data now ! » [3]. Finalement, en 2013 a été publiée une **charte sur l'open data**, suite à la réunion du G8, qui a été généralisée dans un second temps au contexte international, fixant 5 grands principes pour l'ouverture des données. Cette dernière doit être la pratique par défaut des

1. Benyayer et Chignard, *La CADA fête ses 40 ans : retour sur les premiers pas de la transparence administrative*.
2. Goëta, "Rapport sur l'Open Data".
3. Chignard, "Une brève histoire de l'Open Data".

administrations en publiant des données de qualité, accessibles, centralisées, comparables et interopérables, qui permettent d'améliorer la gouvernance et d'encourager la participation citoyenne en donnant le pouvoir d'agir à tous. [4]

On comprend ainsi l'enjeu et l'importance de la loi de 2016 qui **oblige** désormais les organisations concernées à publier leurs données, garantissant alors une nouvelle dimension à l'open data. L'***open data*** désigne l'ensemble des données accessibles et ouvertes à tous, qui peuvent être utilisées et partagées librement. Ces dernières peuvent porter sur n'importe quel sujet allant des comptes administratifs aux actes d'état civil, en passant par les horaires des administrations d'une commune. Tout sujet ouvert peut permettre une réutilisation et ainsi un potentiel bénéfice, à la fois pour les acteurs qui ouvrent les données et pour ceux qui les réutilisent par la suite. La Commission européenne estime ainsi les bénéfices économiques issus des données ouvertes à 40 milliards d'euros par an. [5]

Les principaux avantages de l'open data sont les suivants :

- **pour les pouvoirs publics** : augmentation de la qualité des données, renforcement de la légitimité et plus grande ouverture sur l'extérieur ;
- **pour l'économie** : gains d'efficacité dans les services publics et réutilisation des données mises à disposition ;
- **pour les citoyens** : plus grande transparence et accès facilité aux informations, permettant de développer des opinions propres [6].

Lorsqu'elles sont rendues accessibles et réutilisables, les données publiques offrent aux agents de nouvelles manières d'innover et de collaborer. De nombreux services ont vu le jour, s'appuyant principalement sur les données ouvertes ; analyse du trafic cycliste à Paris, de la pollution de l'air dans les écoles, localisation des points de recyclage, des toilettes publiques etc. Certaines applications sont aussi fondées sur l'open data telles que *Pokémon Go* qui utilisent les données d'*OpenStreetMap*, *Yuka*, ou plus récemment *Vite Ma Dose* qui analyse

4. Goëta, "Rapport sur l'Open Data".
5. Data.gov.be, *Quels sont les avantages de l'Open Data ?*
6. Ibid.

les données de rendez-vous de vaccination contre le COVID-19 des plateformes Doctolib, Keldoc, Maiia, Ordoclic, MaPharma, AvecMonDoc, Clikodoc, MeSoigner et Bimedoc. [7]

Les avantages de l'open data sont donc certains, cependant, malgré cela et l'obligation juridique en place depuis 5 ans, seul 1 territoire sur 10 qui est concerné par la loi, publie effectivement des données. On peut alors se demander pourquoi certains publient et d'autres non. Le phénomène peut-il être mesuré, si oui par quels facteurs ? C'est à cette question que nous tenterons de répondre au cours de cette analyse. Ce travail fait notamment suite à la thèse de Samuel Goëta, cofondateur de Datactivist chez qui j'ai réalisé mon stage de fin d'études, qui a étudié **l'émergence et la mise en oeuvre des politiques d'open data**, au moyen d'une enquête sociologique menée sur 7 collectivités locales et institutions françaises.

Étant un phénomène relativement récent si l'on considère la loi pour une République Numérique de 2016, la littérature sur le sujet est assez limitée. Néanmoins, l'observatoire open data des territoires réalise et met à disposition sur son site une analyse graphique de la situation d'ouverture de données dans les territoires de France. Elle permet d'établir plusieurs constats ; il semble y avoir un seuil du nombre d'habitants à 100.000 au dessus duquel 53.8% des communes et EPCI ouvrent des données, contre environ 8% en dessous de celui-ci (avec plus de 3.500 habitants pour les communes). On voit aussi que 41.4% des territoires qui publient des données le font sur *data.gouv.fr* tandis que 28.5% le font sur leur propre portail open data.

Le principal apport de cette étude par rapport aux travaux réalisés, réside dans la recherche de facteurs explicatifs. Alors que l'OODT représente le nombre de publications en fonction de la population, de la plateforme utilisée, des thèmes publiés, ou encore de la mise à jour des jeux publiés, nous inclurons dans cette analyse des variables diverses, censées qualifier la situation politique, économique et sociale des territoires. De plus, grâce aux modèles de comptage, nous pourrons établir des prévisions qui seront probablement utiles, notamment pour Datactivist, pour prospecter auprès des organismes identifiés.

7. Goëta, "Rapport sur l'Open Data".

Nous effectuerons l'analyse sur les territoires de France suivants :

- les **collectivités**, c'est-à-dire les régions, départements et communes ;

- les **intercommunalités** (aussi appelées "EPCI"), c'est-à-dire les communautés urbaines (CU), les communautés de communes (CC), les communautés d'agglomérations (CA) ainsi que les métropoles.

Pour cette analyse nous procéderons en trois parties ; dans une première section nous présenterons l'**environnement économique** du sujet et justifierons le **choix des facteurs** retenus pour le traiter, puis dans une deuxième partie nous présenterons la **méthodologie économétrique** que nous utiliserons pour répondre à la problématique. Finalement, nous diviserons la dernière section d'**application** économétrique en 3 sous parties ; présentation de la base et du processus d'élaboration, analyse exploratoire pour comprendre les données et émettre de premières hypothèses, puis modélisations dans le but de répondre à la problématique de l'étude.

Bien que nous récoltions des données pour tous les types de collectivités locales présentés ci-dessus, nous décidons de centrer la partie application seulement sur les **départements** de France qui ont, eux, des données complètes, et qui offrent assez d'observations pour mener à bien une analyse (contrairement aux 18 régions par exemple).

6 Partie 1 : environnement économique

En 2019, la France se plaçait en deuxième position de l'*OURdata Index*, un classement mis en place par l'OCDE pour évaluer l'efficacité des politiques d'ouverture des données gouvernementales. L'indice français est passé de 0.85 en 2018 à 0.9 en 2019, soit 0.03 point de moins que la Corée du Sud - championne open data, notamment grâce aux progrès dans la formation et l'accompagnement des administrations [8]. Nous pouvons aisément supposer que l'obligation imposée par la loi de 2016 contribue fortement à cette place, en poussant les territoires et administrations à ouvrir leurs données. [9]

C'est précisément cet indicateur du nombre de données ouvertes, que nous cherchons à expliquer à travers cette étude. Ainsi, dans cette première partie nous regarderons en détail cette variable, puis nous étudierons les déterminants qui peuvent l'influencer, en les distinguant selon leur nature : économique, politique, démographique et géographique.

6.1 Expliquer le nombre de publications open data

Avant de présenter la variable dépendante de notre étude, regardons la définition d'une "donnée ouverte" fixée par l'*Open Knowledge Foundation* en 2005. On parle d'open data lorsque les principes suivants sont respectés ;

- **disponibilité et accès** : l'accessibilité des données doit être garantie via notamment le téléchargement possible sur internet et la possibilité de modifier les données ;

- **réutilisation et redistribution** : elles doivent être exploitables et permettre des comparaisons entre différentes zones ; c'est ce qu'on appelle l'**interopérabilité** ;

- **participation universelle** : tout le monde doit pouvoir utiliser et redistribuer ces données, c'est-à-dire qu'aucune discrimination ne doit être appliquée lors de leur ouverture [10].

Pour mesurer la maturité open data des territoires, nous considérons le **nombre de**

8. Garrone, "La France de nouveau sur le podium de l'open data en 2019".
9. L'**indice** repose effectivement sur les 3 critères suivants : la disponibilité des données, leur accessibilité et le soutien du gouvernement pour leur réutilisation.
10. Lepine, "Open Data définition : qu'est-ce que c'est ? À quoi ça sert ?"

publications sur un portail open data, recensé par l'OODT depuis 2017. Accessible sur plusieurs plateformes différentes, on trouve sur datagouv la variable avec la description suivante :« ce jeu de données référence les acteurs territoriaux qui, en France, produisent et publient des données publiques ouvertes (au minimum 1 jeu de données publié sur un site web, une plateforme dédiée ou une plateforme mutualisée, y compris sur data.gouv.fr) » [11].

On y trouve ainsi la liste de toutes les collectivités territoriales et leurs établissements pratiquant l'open data, avec pour chaque organisation le nombre de jeux publiés sur datagouv et/ou le nombre de jeux publiés sur un portail local. Éclairons ces quelques termes. Pour rappel nous entendons par "**collectivités**" les régions, départements et communes, et par "**intercommunalités**", "**EPCI**" ou encore "**établissements**" les CA, CU, CC et métropoles.

La publication de données peut s'effectuer en suivant des méthodologies différentes ; voilà les 5 étapes définies par l'*Open Data Canvas* [12], plateforme soutenue par Datactivist et qui donne un cadre à l'ouverture de données en proposant différentes ressources et méthodologies :

- **Diagnostic** de la collectivité pour connaître sa maturité open data, ses objectifs et ses motivations à l'ouverture ;
- **Identification** des données à publier ainsi que des services à contacter pour les obtenir ;
- **Mise en qualité** des données par la standardisation, le nettoyage et l'ajout d'informations manquantes le cas échéant ;
- **Publication** des jeux de données sur data.gouv.fr ou tout autre portail open data propre à la collectivité ;
- **Valorisation** des données publiées par le biais de datavisualisations.

Nous voyons ainsi que la publication peut s'effectuer à la fois sur la plateforme des données publiques françaises *datagouv*, et/ou sur un portail open data local, à l'initiative de l'organisation. Au 14 août 2021, datagouv comptabilise 39.040 jeux de données, 210.074 ressources, 2.740 réutilisations, 80.338 utilisateurs, 2.928 organisations et 9.426 discussions. La

11. OpenDataFrance, *Données de la carte de l'observatoire open data des territoires*.
12. Voir le site internet : https://opendatacanvas.org/

plateforme du gouvernement offre la possibilité à tous les agents de publier et de référencer des données, permettant ainsi la réutilisation par tous. Dans les cas où les collectivités publient sur un portail local c'est-à-dire autre que datagouv, 60% d'entre elles utilisent **OpenDataSoft** [13], qui propose une solution de partage et de réutilisation de données payante.

Néanmoins, lorsqu'une organisation publie sur son propre portail **et** sur datagouv il peut y avoir des problèmes de moissonnage ; certaines collectivités découpent leurs fichiers publiés sur un portail local en sous-jeux sur datagouv, ou alors republient sous leur nom d'organisation des données collectées par l'INSEE, faussant alors les statistiques. C'est par exemple le cas pour la ville d'Agen où l'on voit 5 jeux de données produits et publiés par la mairie sur leur portail local, et 68 sur datagouv. Le recensement du nombre de données sur une plateforme locale tient effectivement compte du producteur ; ainsi le nombre de publications est filtré pour ne comptabiliser que celles produites par l'organisation concernée - ce qui n'est pas possible sur datagouv donc est, a fortiori, moins fiable.

C'est pourquoi pour mesurer l'ouverture de données d'une organisation (collectivité ou EPCI) nous garderons le nombre de publications sur la **plateforme locale** en priorité, et le nombre de publications sur **datagouv** dans le cas où l'organisation n'a pas de portail propre. Nous faisons donc le choix de ne pas sommer ces 2 variables initialement distinctes.

Les données sont mises à jour régulièrement par *OpenDataFrance*, nous analyserons donc le nombre de publications open data des territoires en 2021. Voyons à présent quelles variables, susceptibles d'influencer le nombre de données ouvertes, nous retenons pour notre analyse.

6.2 Les déterminants de l'ouverture de données

Phénomène relativement récent, l'open data n'a pas fait l'objet de beaucoup d'études pour en comprendre les facteurs. En revanche, certaines analyses ont été menées pour mesurer les **bénéfices** aux Etats-Unis, au Royaume-Uni, au Danemark etc. On peut retrouver quelques-uns de ces travaux sur le site de la Banque Mondiale.

13. Voir le site internet : www.opendatasoft.fr

Cependant, effectuant mon stage dans un organisme spécialisé dans l'open data, j'ai pu demandé aux uns et aux autres, selon leur expertise et leur expérience, les **facteurs** susceptibles d'influencer Y - en plus des variables économiques sélectionnées instinctivement pour décrire la situation du territoire.

Les déterminants peuvent être regroupés en 4 familles : économiques, politiques, démographiques et géographiques. Commençons par ce premier groupe de variables.

6.2.1 Les facteurs économiques

Pour qualifier l'économie des territoires nous avons récolté des données sur le **PIB par habitant**, le **taux de chômage**, le **nombre de créations d'entreprises**, le **nombre de nuitées recensées dans des hôtels de tourisme**, la **médiane du niveau de vie**, la **décomposition de l'économie** et les **dépenses totales par habitant**. De l'indicateur macroéconomique le plus courant, au plus spécifique pour ce cas d'analyse, ces 7 facteurs ont pour objectif de mesurer la santé économique des organisations, donc leur capacité à innover.

• Le **produit intérieur brut par habitant** correspond à la valeur du PIB divisée par le nombre d'habitants de la zone géographique[14]. Ainsi divisé par la population, cet indicateur permet de comparer des territoires aux tailles divergentes en les plaçant sur une même échelle puisque le PIB est reporté au nombre d'habitants. Selon l'office fédéral de la statistique Suisse, la croissance du PIB par habitant dépend de deux facteurs principaux liés au travail des individus :

- la productivité du travail (c'est à dire " le rapport entre le PIB et le nombre d'heures effectivement travaillées sur le territoire économique considéré") ;

- le nombre moyen d'heures effectives de travail par habitant sur une année.

Nous attendons une relation positive au nombre de publications, étant donné qu'un PIB par habitant élevé traduit une plus grande richesse, donc une capacité à investir et se développer - notamment dans des politiques open data.

14. Blanpain, "L'espérance de vie par niveau de vie : chez les hommes, 13 ans d'écart entre les plus aisés et les plus modestes".

• De la même manière, nous mesurons dans cette étude le niveau d'activité d'un territoire par son **taux de chômage** ; celui-ci correspond au pourcentage de chômeurs dans la population active, d'après l'INSEE [15].

Un taux de chômage faible est la preuve d'une dynamique du territoire et d'un cercle vertueux en découlant ; les institutions peuvent concentrer l'investissement sur de nouveaux projets ayant pour but d'attirer plus d'individus et de renforcer l'économie par le développement de nouvelles activités à partir des données ouvertes [16].

• Il en va de même pour le **nombre de créations d'entreprises**, le **nombre de nuitées recensées dans des hôtels de tourisme** et la **médiane du niveau de vie** ; un valeur élevée traduit une économie en bonne santé, qui favorise le développement ou la reprise de projets open data.

• Nous cherchons ensuite à caractériser la production des territoires. Pour cela nous avons récolté des données sur la part des secteurs dans la valeur ajoutée, pour rappel les catégories sont les suivantes [17] :

- le **secteur primaire** qui regroupe toutes les activités agricoles ;

- le **secteur secondaire** qui regroupe les activités impliquant une transformation des matières premières ;

- le **secteur tertiaire** qui regroupe les services marchands (transports, activités financières, commerce, information-communication...) et non marchands (administration publique, enseignement, santé, social) [18].

Le poids de chaque secteur a beaucoup évolué, depuis la Révolution Industrielle dans la deuxième moitié du 18e jusqu'à l'apparition de l'informatique et des systèmes de télécommunication dans les années 1980. Ainsi en 2017 en France, 75.9% des emplois se trouvent dans le secteur tertiaire, d'après l'enquête Emploi menée par l'INSEE.

Les collectivités territoriales que nous étudions ici ont organisé leurs économies de diffé-

15. INSEE, *Définition du taux de chômage*.
16. Chignard, "Une brève histoire de l'Open Data".
17. OpenDataFrance, *Données de la carte de l'observatoire open data des territoires*.
18. INSEE, *Définition du secteur tertiaire / Tertiaire*.

rentes manières ; ainsi la part respective de chaque secteur diverge d'un territoire à l'autre, selon le nombre d'emplois existants dans chacun d'eux.

Pour n'écarter aucune possibilité de facteurs explicatifs plus ou moins pertinents, nous décidons de récolter toutes les variables sur la répartition des emplois dans l'économie : nous avons ainsi la part du secteur primaire dans la VA, du secteur secondaire, du tertiaire marchand et du tertiaire non marchand. Cependant, en plus de la corrélation forte qu'il existe entre ces variables il s'agit de *dummies*, donc en laissant les trois secteurs il y aura de la redondance. Conscients de cela nous décidons toutefois de recueillir les données pour ces 4 facteurs explicatifs, mais sélectionnerons celui qui caractérise le plus Y grâce à l'analyse exploratoire.

• Enfin, bien que la volonté de l'organisation soit un facteur important pour initier ou continuer une démarche open data, celui du budget l'est tout autant, car il peut à la fois favoriser l'émergence de telles politiques et les freiner s'il est faible. En outre, nous mesurons les moyens financiers des territoires par les **dépenses totales** qui regroupent les dépenses de fonctionnement, d'investissement et les remboursements d'emprunts. Ainsi, l'agrégat enregistre l'ensemble des dépenses réelles pour une année, hors gestion active de la dette et quelle que soit la nature de la dépense. [19]

Par souci de comparaison entre les territoires de tailles divergentes, nous divisons les dépenses par la population pour obtenir les dépenses totales par habitant. Comme expliqué ci-dessus, la relation que nous attendons entre le nombre de publications et les dépenses est positive.

Toutes les variables économiques sont issues des statistiques locales de l'INSEE, hormis les dépenses totales qui proviennent des comptes consolidés des communes, diffusés par l'observatoire des finances et de la gestion publique locales (OFGL).

19. OFGL, *Méthodologie des agrégats financiers - dépenses totales.*

6.2.2 Les facteurs politiques

Outre les facteurs caractérisant l'économie d'un territoire, sa couleur politique mesurée par le parti du chef de l'exécutif peut expliquer l'ouverture ou non des données. Dans les valeurs mêmes des partis on retrouve les notions de partage, de mise en commun, ou au contraire de conservatisme. Regardons les valeurs des deux partis traditionnels.

Dans le but de réduire les inégalités, la **gauche** propose un niveau de service public important avec une forte intervention de l'état. Luttant contre le capitalisme, elle prône ainsi la solidarité, le partage et la mise en commun pour que chacun ait les mêmes chances [20].

A contrario, pour la **droite** l'individu est le moteur de la société. Ce parti tend donc à minimiser l'intervention de l'état dans l'économie, il favorise le sens du travail et la méritocratie. Alors que la gauche est davantage progressiste, la droite est davantage conservatrice [21].

À travers ces définitions, on comprend aisément que la gauche, visant le progrès et la mise en commun, peut davantage se tourner vers l'open data que la droite, fondée sur la réussite individuelle. Nous aurons l'occasion de vérifier cette hypothèse au cours de l'analyse, grâce à la variable explicative de la couleur politique des territoires.

Pour sa récolte nous avons procédé en 2 étapes ; d'abord nous récupérons le **chef de l'exécutif** grâce au répertoire national des élus (RNE), puis retrouvons le ou les **partis politiques** auxquels il adhère qui sont répertoriés sur *wikidata*, le wikipédia des données [22]. Puisqu'une démarche open data nécessite du temps pour être déployée, nous ne récupérons pas les chefs de l'exécutif en place cette année, mais plutôt ceux recensés au 5 juillet 2019, archives du RNE les plus anciennes disponibles sur datagouv.

Outre les nom et prénom des chefs récupérés grâce au RNE, nous sélectionnons leur **date de naissance** ainsi que leur **catégorie socio-professionnelle**, avec pour but de noter un éventuel impact sur la décision d'ouvrir les données, ou de continuer une démarche open data déjà mise en place sur le territoire concerné. Ainsi, nous calculerons l'âge du chef à partir de sa date de naissance, en supposant qu'un dirigeant plus jeune se tournera davantage vers le

20. JUNIOR, "Droite et gauche : histoire d'un clivage politique".
21. MOYNOT, "La droite et la gauche expliquées à ma fille".
22. Pour en savoir plus sur **Wikidata**, voir le site internet de l'outil.

numérique et les données publiques. De même, nous étudierons empiriquement l'effet de la profession et la catégorie sociale du chef sur une politique d'ouverture de données.

6.2.3 Les facteurs démographiques

Dans une étude publiée par le Programme Société Numérique en 2019 intitulée "*Initiatives autour de l'inclusion numérique des personnes âgées*", on peut lire :« L'âge reste, avec le niveau de diplôme, le principal facteur qui entre en jeu pour déterminer si une personne utilise ou non Internet et les outils numériques. » Effectivement, parmi les personnes de 60 à 74 ans qui résident en France en 2019, 15.4% ne disposent d'aucun accès à Internet depuis leur domicile, et ce taux est de 53.2% chez les plus de 75 ans [23].

Ainsi, en plus des variables économiques et politiques, nous incluons dans notre étude 4 variables censées caractériser la population de la collectivité ou de l'intercommunalité. La première correspond au **nombre d'habitants** lui-même, la deuxième à la **part des plus de 65 ans** dans la population, la troisième à la **part des diplômés d'un BAC+5 ou plus dans la population non scolarisée de 15 ans ou plus** et la dernière aux **effectifs d'étudiants inscrits dans l'enseignement supérieur**.

Par rapport à la variable à expliquer, nous attendons les relations suivantes :

- **positive** entre Y et la population : au seuil de 3500 habitants l'open data devient une obligation, et nous supposons que la hausse de la population s'accompagne d'un dynamisme, favorable à la mise en place de tels projets.

- **négative** entre Y et la part des plus de 65 ans : si la part des personnes âgées dans la population est élevée alors le nombre de publications sera faible voire nul, puisque cela correspondra moins aux besoins de la population du territoire.

- **positive** entre Y et la part des diplômés : un plus grand nombre de diplômés dans la population peut refléter un plus grand savoir sur le territoire et donc une utilité encore plus grande à la publication de données pour leur réutilisation.

- **positive** entre Y et le nombre d'étudiants : cette relation fonctionne comme celle de

23. Numérique, "Initiatives autour de l'inclusion numérique des personnes âgées".

la part des plus de 65 ans avec le nombre de publications. Davantage d'étudiants, notamment en formations techniques, inciteraient les administrations à initier ou améliorer des projets OD.

Pour cette troisième variable nous disposons de 2 mesures ; la première, récoltée telle quelle depuis le site de l'**enseignement supérieur** correspond au nombre d'étudiants, que nous avons divisé par la population dans une deuxième variable pour avoir la part d'étudiants. De cette manière, nous n'écartons pas la possible présence de seuils du nombre d'étudiants en-dessous ou au-dessus desquels le nombre de publications est plus faible ou plus fort. Aussi, comme pour certaines variables économiques, nous décidons de garder ces 2 mesures dans un premier temps pour englober le plus d'informations, puis nous sélectionnerons la variable la plus pertinente au vu de l'analyse exploratoire. De même, nous supposons dès à présent une corrélation négative forte entre la part des plus de 65 ans et du nombre d'étudiants, ici encore nous sélectionnerons les variables avant les modélisations.

Les données de la part des plus de 65 ans et de la part des diplômés ont été récoltées à partir des statistiques locales de l'INSEE, portant sur l'année 2018. Le nombre d'étudiants issu des données de l'enseignement supérieur recense les inscrits pour l'année 2018-2019.

6.2.4 Les facteurs géographiques

"Le manque d'infrastructures numériques reste d'actualité selon une grande partie des experts. Les métropoles tendent à concentrer les moyens matériels et humains. La couverture progresse, le plan Très Haut-débit est lancé, mais des efforts sont encore à faire pour que les habitants des territoires ruraux puissent accéder aux mêmes services et usages que ceux des villes." C'est ce que l'on peut lire dans une étude portant sur l'impact des usages du numérique sur le développement territorial, menée en 2018 par le **Réseau rural**, notamment en charge de travailler sur la transition numérique des territoires ruraux.

Nous avons vu jusqu'ici comment le nombre de publications peut dépendre de l'activité économique, des moyens financiers et de choix politiques. Nous étudions désormais un autre aspect d'une démarche open data : les moyens techniques pour la mettre en oeuvre. Alors que les territoires ruraux sont généralement moins bien équipés en infrastructures numériques que

les territoires urbains, l'ouverture de données peut être fortement impactée. D'une part pour mettre en oeuvre la démarche, mais aussi a posteriori, dans l'utilité des données ouvertes pour les acteurs du territoire qui n'auraient pas accès aux informations et ne seraient donc pas en mesure de consulter ou réutiliser les jeux ouverts.

La définition de la ruralité a beaucoup évolué ces dernières années, mais depuis 2020 celle-ci n'est plus définie en creux (tout ce qui n'est pas urbain est rural), mais en "plein" à partir de la grille communale de densité qui contient 4 modalités :

- dense
- densité intermédiaire
- peu dense
- très peu dense

Le rural regroupe alors toutes les communes peu denses et très peu denses, ce qui englobe 32.8% de la population, alors qu'avant cette définition seule 4.5% de la population était considérée comme vivant en zone rurale [24].

Cependant, même améliorée cette définition a fait l'objet de critiques, citons par exemple celle du géographe Martin Vanier qui reproche de ne pas prendre en compte les interdépendances qui existent entre les différents espaces. Ainsi, l'INSEE s'est attelé à trouver une nouvelle définition du rural, publiée dans son étude "La France et ses territoires" le 29 avril 2021 [25]. Le raisonnement initié avec la grille de densité est complété avec l'analyse des aires d'attraction des villes, ce qui permet de prendre en compte les relations existantes entre les territoires. Les catégories sont désormais au nombre de 6 et distinguent l'influence forte ou faible d'un pôle, sur la base de l'indicateur "domicile-travail". On retrouve les catégories suivantes :

- urbain dense
- urbain densité intermédiaire
- rural sous forte influence d'un pôle
- rural sous faible influence d'un pôle

24. Olga, "Qu'est-ce que le « rural » ? Analyse des zonages de l'Insee en vigueur depuis 2020".
25. Ibid.

- rural autonome peu dense

- rural autonome très peu dense

Les données ainsi fournies par l'INSEE avec ces deux définitions sont disponibles uniquement pour les communes de France, il est cependant possible de les calculer pour les autres niveaux géographiques. Ainsi, afin de recueillir ces informations aux niveaux régionaux et départementaux nous utiliserons 3 méthodes différentes.

• En partant de la nouvelle définition du rural qui est constituée de 6 modalités, nous élèverons les données aux niveaux géographiques supérieurs en **calculant le mode**, c'est-à-dire la modalité apparaissant le plus, en groupant les communes par régions ou départements.

• En partant cette fois de la grille de densité (2020) composée de 4 modalités, nous appliquerons la **méthode d'agrégation proposée par l'INSEE** pour élever les données communales aux niveaux supérieurs. Illustrée en annexe n°1, la règle de décision s'applique de la manière suivante :

- si la part de la population vivant en communes très denses est supérieure à 50%, alors la maille [26] est considérée comme étant très dense (catégorie **1**)

- si la part de la population vivant en communes très denses et denses est supérieure à 50% et que la maille n'est pas de catégorie 1, alors la maille est considérée comme étant dense (catégorie **2**)

- si la part de la population vivant en communes très peu denses est supérieure à 50%, alors la maille est considérée comme étant très peu dense (catégorie **4**)

- si la part de la population vivant en communes peu denses et très peu denses est supérieure à 50% et que la maille n'est pas de catégorie 4, alors la maille est considérée comme étant peu dense (catégorie **3**) [27]

• De la même manière, cette méthode d'agrégation a fait l'objet de critiques, notamment par **Olivier Bouba-Olga**, chercheur en sciences sociales qui estime cette définition encore perfectible. Pour lui, la limite de cette règle de décision est« de ne pas tenir compte de la part de la population urbaine et rurale de chaque entité, mais de transformer cette part en une

26. La **maille** désigne ici tout niveau géographique supra communal, à savoir EPCI, départements ou régions.

27. INSEE, *Méthode d'agrégation pour obtenir la grille de densité à un niveau géographique supra communal.*

variable binaire (urbain ou rural) en fonction de l'orientation dominante » [28]. Il propose alors une autre alternative que nous utiliserons comme troisième méthode pour obtenir la ruralité au niveau des départements et des régions ; elle consiste à calculer la part de la population qui réside en commune rurale (c'est-à-dire peu denses ou très peu denses). Elle permet d'obtenir alors, non pas des catégories, mais un pourcentage allant de 0 à 100% et par conséquent plus précis que les modalités.

Pour conclure sur ce facteur de ruralité, nous aurons pour les communes 2 variables issues des définitions de 2020 et 2021 avec respectivement 4 et 6 classes. Puis, nous obtiendrons pour les départements et les régions 3 variables indiquant le niveau de ruralité, via les 3 méthodes énoncées ci-dessus. La sélection se ferra, ici encore, lors de la partie exploratoire des données. Nous faisons le choix de ne pas calculer cette variable pour les intercommunalités qui regroupent parfois des communes de plusieurs départements, ce qui rendrait les calculs trop complexes.

Finalement, une autre variable géographique que nous inclurons comme déterminant potentiel de l'ouverture de données est celle des **flux principaux de migration résidentielle**. Ces derniers correspondent au nombre de personnes de un an ou plus ayant déménagé d'un territoire vers un autre, en considérant uniquement le flux le plus important de chaque territoire [29]. Les valeurs sont celles du code officiel géographique (COG) du département de destination. Par cette variable, nous voulions voir si certains départements avaient un impact sur le nombre de publications, traduisant un pôle attractif où des politiques open data sont développées et pourraient influencer d'autres zones.

6.3 Récapitulatif des variables utilisées pour l'analyse

Récapitulons les informations des variables choisies dans un tableau, pour plus de clarté.

28. Olga, "Qu'est-ce que le « rural » ? Analyse des zonages de l'Insee en vigueur depuis 2020".
29. INSEE, *Documentation fichier détail : Migrations résidentielles.*

TABLE 1 – Résumé des variables explicatives retenues pour l'analyse

Variable	Mesure	Domaine	Source	Année
PIB par habitant	€ /hab	Économique	*INSEE statistiques locales*	2018
Taux de chômage annuel moyen	%	Économique	*INSEE statistiques locales*	2020
Nombre de création d'entreprises		Économique	*INSEE statistiques locales*	2020
Nb de nuitées ds hotels de tourisme	milliers	Économique	*INSEE statistiques locales*	2019
Médiane du niveau de vie	€	Économique	*INSEE statistiques locales*	2018
Part de l'agriculture ds la VA	%	Économique	*INSEE statistiques locales*	2018
Part de l'industrie ds la VA	%	Économique	*INSEE statistiques locales*	2018
Part du tertiaire marchand ds la VA	%	Économique	*INSEE statistiques locales*	2018
Part du tertiaire non marchand ds la VA	%	Économique	*INSEE statistiques locales*	2018
Dépenses totales par habitant	€ /hab	Économique	*OFGL*	2019
Parti politique du chef de l'exécutif	3 modalités	Politique	*Wikidata*	2019
Age du chef de l'exécutif	année	Politique	*RNE datagouv*	2019
CSP du chef de l'exécutif	string	Politique	*RNE datagouv*	2019
Nombre d'habitants		Démographique	*OFGL*	2019
Part des 65 ans ou plus	%	Démographique	*INSEE statistiques locales*	2018
Part des diplômés d'un BAC+5 ou plus dans la pop. non scolarisée de 15 ans ou +	%	Démographique	*INSEE statistiques locales*	2018
Effectifs d'étudiants inscrits dans l'enseignement supérieur	nb d'étudiants	Démographique	*Enseignement supérieur*	2018-2019
Effectifs d'étudiants inscrits dans l'enseignement supérieur	% de la pop	Démographique	*Enseignement supérieur*	2018-2019
Niveau de densité	4 modalités	Géographique	*INSEE*	2021
Niveau de ruralité	6 modalités	Géographique	*INSEE*	2021
Part de la population vivant en zone rurale	% de la pop	Géographique	*INSEE*	2021
Flux principal de migration résidentielle	n° dep	Géographique	*INSEE statistiques locales*	2018

Le tableau n°1 reprend les principales informations des facteurs explicatifs retenus pour l'analyse et justifiés au cours de cette section. Il y a 21 variables caractérisant 4 domaines des territoires, issues de 6 sources différentes (7 avec les données de l'observatoire open data des territoires pour Y). Il existe une **différence de temporalité** dans les données récoltées ; on voit sur le tableau que certaines concernent les années 2018 ou 2019 tandis que d'autres portent sur 2020 ou 2021, soit un écart de 3 ans au maximum. Pour chacune des variables, nous avons cherché les données les plus actuelles possible (excepté celles sur le chef de l'exécutif pour les raisons évoquées précédemment), tout en étant limités par la disponibilité de certaines.

Ainsi, les années recensées dans le tableau sont les plus récentes que nous ayons trouvées pour analyser le nombre de publications open data de 2021. Cependant, comme nous l'avons expliqué au cours de la partie, une démarche open data peut demander plusieurs années pour mûrir et être déployée, c'est pourquoi avoir des variables explicatives de *t-1* à *t-3* n'est pas aberrant. En ce qui concerne le niveau de ruralité, les données portent sur l'année 2021 puisque c'est au cours du mois d'avril que l'INSEE a mis au point cette nouvelle définition du rural. Ici aussi, le fait que les informations ne soient pas des années précédentes comme les autres facteurs explicatifs n'est pas gênant, puisque la ruralité d'un territoire n'est pas une variable amenée à évoluer au cours du temps ; seulement lorsque les définitions changent. Enfin, un dernier point concernant le décalage temporel entre certaines variables exogènes. L'étendue de la période étant relativement stable ; avant la crise du coronavirus et effets de la crise des Subprimes largement estompés, et étant donné la nature des variables (agrégats économiques et non indices financiers par exemple), nous faisons l'hypothèse que les variations des données d'une année sur l'autre ne sont pas significatives au point de biaiser l'analyse car les facteurs ne seraient pas de la même année. Nous retrouvons par exemple en annexe n°2, l'évolution du taux de chômage dans la région Île-de-France de 2006 à 2020 ; on constate alors que l'écart entre les années 2019 et 2020 est très faible (de 0.1%).

Les 21 variables présentes dans le tableau ne sont pas toutes disponibles pour les 5 niveaux géographiques (régions, départements, communes et EPCI). Nous retrouvons leur disponibilité en table n°2 ; il y a 22 facteurs explicatifs pour les régions, 17 pour les départe-

ments, 14 pour les communes et 10 pour les EPCI.

TABLE 2 – Disponibilités des données pour chaque niveau géographique

	Régions	**Départements**	**Commes**	**EPCI**
PIB par habitant	✓			
Taux de chômage annuel moyen	✓	✓		
Nombre de création d'entreprises	✓	✓	✓	✓
Nb de nuitées ds hotels de tourisme	✓	✓		
Médiane du niveau de vie	✓	✓		✓
Part de l'agriculture ds la VA	✓			
Part de l'industrie	✓			
Part du tertiaire marchand	✓			
Part du tertiaire non marchand	✓			
Dépenses totales par habitant	✓	✓	✓	✓
Parti politique du chef de l'exécutif	✓	✓	✓	✓
Age du chef de l'exécutif	✓	✓	✓	✓
CSP du chef de l'exécutif	✓	✓	✓	✓
Nombre d'habitants	✓	✓	✓	✓
Part des 65 ans ou plus	✓	✓	✓	✓
Part des diplômés d'un BAC+5 ou +	✓	✓	✓	✓
Effectifs d'étudiants inscrits dans l'enseignement supérieur (volume)	✓	✓	✓	
Effectifs d'étudiants inscrits dans l'enseignement supérieur (%)	✓	✓	✓	
Niveau de densité	✓	✓	✓	
Niveau de ruralité	✓	✓	✓	
Part de la pop vivant en zone rurale	✓	✓	✓	
Flux principal de migr. résidentielle	✓	✓	✓	✓

Maintenant que nous connaissons les variables et le sujet de l'analyse, regardons dans une nouvelle partie quelle méthodologie nous allons adopter pour le traiter.

7 Partie 2 : méthodologie économétrique

Pour cette partie de méthodologie économétrique nous procéderons en 2 étapes : nous détaillerons dans un premier temps les possibles estimations et préciserons celles que nous choisirons d'appliquer pour la partie modélisation, puis nous développerons les différents tests et analyses que nous utiliserons en partie d'exploration des données.

7.1 Les modélisations

N'étudiant pas le nombre de publications dans le temps mais à une date précise (2021) nous travaillons avec des **données en coupe** aussi appelées "données transversales". La variable à expliquer est de nature quantitative **discrète** donc nous sommes dans le cas de modèles de comptage. Ces derniers sont employés lorsque la variable dépendante résulte d'un processus de comptage ; ici, le nombre de publications est une valeur forcément positive et entière.

Formulés par John Nelder et Robert Wedderburn en 1972, les modèles linéaires généralisés (GLM) englobent le modèle linéaire général, le modèle log-linéaire, la régression logistique et la régression de Poisson. Ils sont composés d'une variable dépendante Y à laquelle est associée une loi de probabilité, de variables explicatives X_i et d'un **lien** décrivant la relation fonctionnelle entre la combinaison linéaire des variables explicatives et l'espérance mathématique de la variable Y. [30]

Y peut être distribué selon plusieurs lois : *Bernouilli, binomiale, Poisson* ou loi *normale*. Lorsque la distribution suit une loi de Poisson comme c'est le cas dans notre analyse puisque Y est une variable de comptage, le lien entre la composante aléatoire et la composante déterministe correspond au logarithme de l'espérance. Ainsi, une modélisation linéaire classique ne serait pas adaptée, puisqu'elle suppose la normalité de la variable à expliquer. De même, compte tenu de cette distribution, la variance des résidus n'est pas constante mais proportionnelle aux comptages moyens prédits par le modèle. En outre, utiliser un modèle linéaire classique pour traiter des phénomènes de ce type peut entraîner une estimation biaisée de

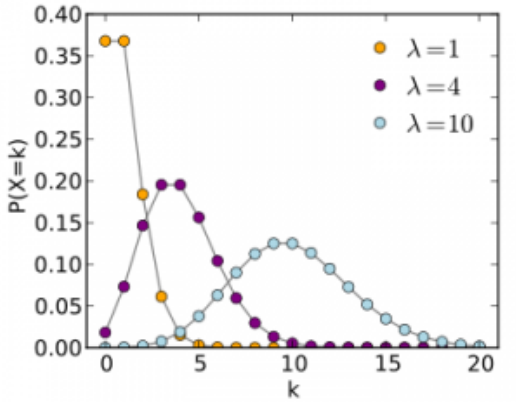
30. CNAM, *Modèles linéaires généralisés*.

l'erreur standard des paramètres, et peut conduire à des prédictions négatives qui n'auraient pas de sens avec une telle variable dépendante. [31]

La loi de Poisson est un cas particulier de la loi binomiale où la fonction de probabilité est la suivante :

$$Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

Le paramètre λ défini donc la distribution, il est égal à l'espérance et la variance de la série, plus il augmente et plus la distribution se rapproche de la loi normale. Cette dernière apparaît donc comme un cas particulier de la loi de Poisson, lorsque le paramètre λ est égal à 20 ou 30. Nous pouvons observer différentes lois de Poisson ci-dessous, en variant le **paramètre de distribution**.



Ainsi, le paramètre λ dépend des variables explicatives via une forme log-linéaire, elle-même définie par l'équation suivante : [32]

$$log(\mu_y) = \sum_{j=1}^{p} \beta_j X_{ij}$$

De la même manière que des estimations par les moindres carrés ordinaires (MCO), les modèles de régression de Poisson doivent respecter plusieurs conditions pour être valides :

- les réponses doivent être **indépendantes**

- les réponses doivent être distribuées selon une **loi de Poisson** de paramètre λ

31. Vedova, *GLM sur données de comptage (régression de Poisson) avec R.*
32. Cours de « Variables qualitatives 2 » de Mme Travers.

- il ne doit pas exister de **surdispersion**

Ce premier critère signifie que les observations du phénomène ne doivent pas être corrélées entre elles. La deuxième hypothèse est généralement supposée mais peut être vérifiée en représentant graphiquement les comptages observés et leur distribution théorique sous une loi de poisson (en calculant donc au préalablement le paramètre lambda comme la moyenne de la série Y). Enfin, cette troisième hypothèse est forte et nécessite une attention particulière. En théorie, avec la loi de Poisson la variance est égale à la moyenne de Y, soit : E(Y) = Var(Y) = λ. Or, en pratique on trouve parfois une variance supérieure à la moyenne : on parle alors de **surdispersion**. Ce phénomène peut conduire à la sous-estimation de l'erreur standard des paramètres du modèle, donc à une p-value faible et par conséquent des conclusions erronées sur les liens entre les variables explicatives et Y.

Plusieurs méthodes existent pour vérifier la dispersion de la série :

- en **comparant la variance et la moyenne** ; si la variance est supérieure à cette dernière alors il y a un problème ;

- en **calculant le rapport entre la déviance résiduelle de l'estimation et le degré de liberté** : si ce ratio est supérieur à 1, alors il y a un problème de sur-dispersion.

Cette surdispersion peut être le résultat de plusieurs choses :

- dépendance entre les réponses

- absence d'une variable explicative importante

- sur-représentation des valeurs zéro dans l'échantillon

Une première méthode pour prendre en compte ce problème est d'introduire un paramètre de dispersion de manière à avoir Var(Y) = $\phi E(Y) = \phi\lambda$. Ce dernier conduit à l'augmentation de l'erreur standard des paramètres du modèle ; nous sommes alors dans le cas de structures dites de "**quasi Poisson**". Une seconde méthode pour prendre en compte la dispersion est l'estimation d'un modèle **binomial négatif**, pour lequel on suppose $Var(Y) = \lambda(1 + \alpha\lambda) = E(Y) + \alpha E(Y)^2$. [33]

33. Cours de « Variables qualitatives 2 » de Mme Travers.

Il est alors possible d'effectuer un test statistique basé sur le ratio de vraisemblance pour connaître le modèle le plus approprié, c'est-à-dire celui qui prend en compte la sur-dispersion ou non. Ce test est applicable sous R grâce à la fonction *odTest()* du package "**pscl**".

Dans certains cas d'analyse, le nombre de zéro est important et nécessite d'être pris en compte pour mesurer au mieux le phénomène. Les estimations qui prennent en compte cela sont appelées "**modèles à inflation de zéro**". Il est possible d'estimer un modèle de Poisson à inflation de zéro (ZIP) ou bien un modèle binomial négatif à inflation de zéro (ZINB). Ces modélisations ont l'avantage de comporter 2 parties, pour mieux mesurer et expliquer Y :

- une partie de **loi binaire** qui explique l'évènement ou le non-évènement ;

- une partie de **loi de Poisson** ou de **loi binomiale négative** qui explique quantitativement Y.

Avec ces nombreuses estimations, il convient de choisir la plus adaptée aux données, pour cela nous recourons au test de *Vuong* qui permet la comparaison des modèles d'une même loi, estimés de manière classique ou en inflation de zéro. L'hypothèse nulle correspond à l'équivalence des modèles, tandis que l'hypothèse alternative est acceptée lorsque le modèle n°1 est supérieur au modèle n°2.

Par ailleurs, lorsque le phénomène n'est pas observé (Y=0) mais que cela tient d'une décision consciente et volontaire (attraper des poissons à la ligne *versus* acheter des carottes bio qui serait alors voulu), il est plus adapté d'appliquer un modèle appelé "**double hurdle**" qui prend en compte dans sa deuxième composante, une loi (poisson ou binomiale négative) **censurée** où Y est strictement supérieur à 0. Ainsi, une première partie du modèle explique pourquoi on observe le phénomène (partie binaire) et la deuxième, lorsque Y>0 explique ce comptage.

Dans notre cas d'étude par exemple, l'ouverture de données est un choix conscient de chaque collectivité qui décide ou non de rendre publiques et accessibles certaines informations. Pour certaines communes qui sont loin de ces sujets et qui n'en ont pas connaissance, le choix n'est peut-être pas pleinement conscient. Cependant, comme précisé précédemment nous

effectuerons l'analyse sur les départements qui, par leur taille, connaissent les enjeux liés à l'open data, étant, qui plus est, tous concernés par la loi pour une République Numérique.

Enfin, nous comparerons ces différentes estimations en nous référant, en plus du test de Vuong, à des mesures telles que le critère d'Akaike (noté **AIC**) : le meilleur modèle est alors celui qui minimise ce critère basé sur la vraisemblance des modèles.

7.2 Tests et analyse exploratoire

Dans cette section nous allons expliquer les tests que nous utiliserons dans la partie suivante portant sur l'analyse concrète des données.

• Avant toutes choses, nous chercherons et traiterons les données atypiques. Ainsi, après identification visuelle des points potentiellement atypiques, nous appliquerons les tests de **Grubbs** et **Rosner**, pour vérifier statistiquement la présence ou non d'outliers. Ils sont tous deux basés sur l'hypothèse nulle d'absence de valeurs atypiques dans la série.

• Aussi, nous regarderons la distribution des variables quantitatives continues pour voir si elles suivent une loi normale. Nous effectuerons pour cela le test de **Shapiro** pour lequel H_0 correspond à une distribution normale.

• De même, pour connaître la dépendance entre les variables qualitatives nous appliquerons des tests de **Khi-deux**, reposant sur l'hypothèse nulle de la dépendance des variables.

• Nous étudierons la **corrélation** entre les variables de type quantitatif, en regardant le coefficient : s'il est compris entre 0.5 et 0.6 la corrélation est moyenne, en revanche elle est forte à partir de 0.6.

Également dans le but d'explorer les données, nous réaliserons un **arbre de régression** (CART) et une **analyse en composantes principales**.

• Les arbres de régression constituent une méthode supervisée de Machine Learning. Un arbre est une classe d'algorithmes non paramétriques qui fonctionnent en partitionnant l'espace d'entités en un certain nombre de régions plus petites avec des valeurs de réponse similaires, à l'aide d'un ensemble de règles de fractionnement. Cette méthode de division a

pour avantage de produire des règles simples qui sont faciles à interpréter et à visualiser avec des diagrammes en arbre. Le fonctionnement des arbres de décision est le suivant : à chaque noeud l'algorithme considère les N variables et cherche celle qui divise le jeu de données de la manière la plus optimale possible, c'est-à-dire avec une hétérogénéité inter-groupe et une homogénéité intra-groupe.

• Dans l'ACP les variables déterminent les axes, elles se présentent sous forme de flèches dans un cercle unitaire de corrélations ; plus elles sont importantes dans l'explication d'un axe, plus elles seront positionnées proche de cet axe et aux extrémités. Le but d'une ACP est de faire ressortir les variables les plus importantes qui définissent des "variables latentes", aussi appelées 'composantes principales'. Ces dernières représentent une combinaison linéaire des variables explicatives initiales. Une ACP peut être très utile pour connaître les relations qui existent entre les variables et les rassembler autour d'axes fictifs.

• Enfin, puisque nous avons de nombreuses variables, nous procéderons à une **sélection de variables** avant les modélisations. Nous appliquerons ainsi les 3 méthodes que sont la procédure ascendante (*backward*), descendante (*forward*) et bidirectionnelle (*both*). Elles visent toutes trois à minimiser le critère AIC ; elles évaluent ainsi ce critère à chaque fois qu'une variable est ajoutée ou retirée du modèle.

Maintenant que nous avons présenté la méthodologie ainsi que les tests que nous appliquerons pour répondre à la problématique, nous pouvons les mettre en pratique dans une troisième partie.

8 Partie 3 : présentation des données, application

Nous allons diviser cette partie application de la méthodologie expliquée, en 3 parties. Dans un premier temps nous présenterons la base en expliquant le processus de construction, dans un deuxième temps nous mènerons une analyse préliminaire des variables dans le but d'explorer les relations existantes, mettre en exergue des sous-populations, émettre des hypothèses et éventuellement construire de nouvelles variables. Enfin, dans une dernière partie nous chercherons le meilleur modèle économétrique qui réponde à la problématique : **quels facteurs influencent l'ouverture de données des territoires ?**. Toutes les manipulations de cette étude, de la récolte des données à la modélisation, ont été réalisées sous R studio.

8.1 Construction de la base de données

Le processus de récolte des données pour construire la base pour l'analyse a été un travail fastidieux, avec de nombreux casse-têtes, des ajouts, des changements nécessitant régulièrement de revenir en arrière, modifier le script et réitérer. Au total, la construction de la base s'est étalée sur près de 4 mois - car comme expliqué dans les remerciements, j'ai eu la chance de réaliser cette analyse en partie pendant mon stage, constituant pour Datactivist un projet de R&D.

Le fichier duquel nous récoltons notre variable à expliquer "nombre de données ouvertes" est celui de l'OODT qui recense uniquement les collectivités et EPCI **pratiquant l'open data** ; il n'y a donc pas d'observation pour les organisations qui n'ont pas encore ouvert de données. Ainsi, nous ne partons pas de ce fichier mais plutôt de celui construit et publié sur datagouv pendant mon stage (comme expliqué dans le rapport), qui, lui, recense toutes les organisations de France. De plus, les fichiers contiennent pour chaque organisation le COG (excepté pour les EPCI qui n'en ont pas) et le numéro de SIREN, qui permettent des jointures plus faciles pour les nombreuses données à ajouter à cette base initiale.

Nous disposons des observations suivantes, pour un total de 37.537 :

- 17 régions
- 100 départements
- 34930 communes
- 1260 EPCI

Initialement, nous avions exporté toutes les données depuis leur source et les importions sous R pour les mettre ensemble, ce qui revenait à exporter depuis 7 sites différents. Cependant, dans un souci de reproductibilité et de simplicité de l'analyse, nous avons par la suite décidé de limiter les exports au maximum, tout en privilégiant les imports via des liens ou des commandes, directement depuis R Studio. Par conséquent, en usant d'astuces, seules les statistiques locales de l'INSEE ont dû être exportées, ainsi que les **corrections** pour la variable à expliquer.

En effet, bien qu'administrées par OpenDataFrance, nous nous étions rendus compte en commençant quelques statistiques descriptives sur les métropoles, que les données n'étaient pas à jour : certaines métropoles n'étaient pas recensées alors qu'elles pratiquaient l'open data. Alors, sachant que nous visions le nombre de publications open data des territoires en privilégiant celles sur un portail local, nous avons vérifié à la main la fiabilité des données disponibles pour les régions, les départements ainsi que les métropoles de France. Nous vérifiions ainsi les organisations pour lesquelles aucune publication n'était déclarée, en cherchant sur internet un éventuel portail open data local, et dans le cas où il n'y en avait pas, vérifiions aussi l'existence de l'organisation sur datagouv. Sur les 57 territoires déclarés ne pratiquant pas l'open data, nous avons pu corriger 12 erreurs pour ceux qui dans les faits la pratiquaient. Cela était début juillet, quelques jours avant qu'ils ne mettent à jour leurs données et que l'on retrouve le bon nombre de publications pour les organisations... Cependant, il reste 2 corrections toujours valables car non ajoutées dans leurs données ; que nous appliquons en important ce court fichier CSV et en ajoutant les 2 lignes au jeu de l'OODT.

Ainsi, en partant des fichiers datagouv, nous avons ajouté les variables petit à petit en utilisant leurs liens internet. Nous nous sommes néanmoins vite heurtés à des difficultés ; des doublons, des numéros de SIREN ayant changé donc pas ajournés de la même manière sur

tous les sites, des données mal renseignées...

Un **exemple** pour illustrer cela ; nous récupérons le chef de l'exécutif via le RNE, un fichier produit par le *Ministère de l'Intérieur*. Or, en l'ajoutant aux données communales, on "gagne" 2 observations ; la première est liée au fait que le maire de Montguers (Drôme) ait 2 fiches renseignées avec des dates de naissance et CSP différentes, et la seconde au fait que 2 maires soient renseignés pour la commune de Ponsan-Soubiran (Gers) avec la même date d'élection et le même nom de famille. Dans le premier cas, une simple recherche internet permet de garder l'observation avec les bonnes valeurs, mais le deuxième est plus compliqué car les informations sont contradictoires même sur les sites internet. Finalement, à partir de plusieurs sources nous constatons que l'un aurait succédé à l'autre. Ce cas particulier montre que même les données officielles peuvent être imparfaites, demandant du temps pour retrouver la source des problèmes et les corriger.

Une fois les données des chefs récupérées, le but était d'obtenir le parti politique de ces derniers, en utilisant **wikidata**, une base de connaissances éditée de manière collaborative. Pour retrouver des informations sur cette plateforme, il faut passer par le service des requêtes qui se font avec le langage SPARQL. Le fonctionnement de wikidata est le suivant ; chaque donnée peut être entrée soit comme *item* donc son identifiant commencera par la lettre Q (par exemple "France" : *Q142*), soit comme *propriété* et commencera par la lettre P (par exemple "pays de citoyenneté" : *P27*). Pour obtenir la couleur politique des territoires, nous étions initialement passés par l'outil **OpenRefine** qui permet de concilier des données de différentes sources, et notamment de wikidata, mais nécessitait des traitements additionnels en dehors de R studio. Ainsi, toujours dans l'optique de minimiser les manipulations, nous avons choisi d'utiliser les requêtes SPARQL de l'outil, en les entrant directement sur R, grâce au package *WikidataQueryServiceR*.

Pour ce faire, nous avons cherché dans un premier temps la requête qui donne pour chaque commune (ou régions, départements, EPCI) le chef de l'exécutif et son parti politique. Après de longues recherches pour comprendre l'outil et ses subtilités, la requête pour l'obtenir est la suivante :

```
SELECT DISTINCT ?coll ?collLabel ?chef_exec ?chef_execLabel ?partiLabel
WHERE
{
  ?coll wdt:P31 wd:Q36784.
  ?coll wdt:P6 ?chef_exec.
  OPTIONAL { ?chef_exec wdt:P102 ?parti. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr,en". }
}
ORDER BY ?collLabel
```

Cependant, les données entrées sur wikidata ne sont pas nécessairement à jour et complètes, donc les informations récupérées par cette requête ne sont pas toujours exactes. C'est pour cette raison que nous avons décidé de récupérer les chefs de l'exécutif non pas par *wikidata* mais via le *RNE* qui ne contient (normalement) pas d'erreurs.

Dans un second temps, nous avons donc cherché à récupérer une liste des partis politiques associés aux chefs de l'exécutif de nos 4 niveaux géographiques. Pour cela nous sommes partis de l'*item* "être humain", puis avons appliqué plusieurs filtres pour limiter le temps d'exécution de la requête : nous avons sélectionné les politiciens français actuels, puis avons récupéré la couleur politique. La requête finale est la suivante :

```
SELECT DISTINCT ?chefLabel ?partiLabel
WHERE {
  ?chef wdt:P31 wd:Q5 .    #tous les êtres humains
  ?chef wdt:P106 wd:Q82955 .    #on limite aux politiciens
  ?chef wdt:P27 wd:Q142  .    #on limite aux pers françaises
  ?chef wdt:P102 ?parti .     #on recup le parti
  ?chef wdt:P569 ?date .   #on recup la date de naissance
  FILTER(YEAR(?date) > 1920).    #on trie pour ne garder que les politiciens
                                 # "actuels" (moins de 100 ans)
  SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }
}
ORDER BY ?chefLabel
```

En entrant cette requête sous R nous obtenons une liste des politiciens et de leurs **partis politiques**, que nous ajoutons aux données de l'analyse en joignant les jeux par les noms et prénoms des chefs (mis préalablement en majuscules et sans accents de manière à optimiser le *match*). Toutefois, lorsqu'un politicien a a été rattaché à plusieurs partis, ils sont tous renseignés sur wikidata et donc récupérés dans notre base ; cela nécessitera un traitement spécifique avant l'analyse exploratoire des variables.

Le reste de l'élaboration de la base a été plus facile dans l'ensemble, seule la traduction de la règle de décision pour obtenir le **niveau de densité** au niveau supra-communal en code R a été laborieuse.

Finalement, nous obtenons 4 bases aux dimensions différentes : les régions (21x27), les départements (117x23), les communes (35078x21) et les intercommunalités (1313x18). Le nombre d'observations est plus important qu'avant l'ajout des variables, puisqu'elles ont été dédoublonnées lorsqu'un chef de l'exécutif avait plusieurs partis politiques. Nous traiterons cela avant de commencer l'analyse des variables.

8.2 Analyse exploratoire

Pour rappel nous étudions la variable quantitative du **nombre de publications open data** des territoires en 2021. Nous disposons de 4 niveaux géographiques distincts qui ne peuvent pas être traités collectivement, étant donné les fortes disparités entre ces derniers. La table n°3 ci-dessous illustre ce fait, on y voit l'étendue des variables quantitatives communes aux 4 niveaux, leur moyenne ainsi que leur médiane.

On observe que la distribution de la variable à expliquer diffère grandement d'un type d'organisation à un autre. La médiane des communes et EPCI est à 0, ce qui traduit une inflation en 0 pour ces organisations-là, alors que pour les régions la médiane est de 69 et la moyenne de 115 publications. De même, les nombres de créations d'entreprises sont hétérogènes au sein des territoires : avec une médiane de 3 par an pour les villes et 51 438 pour les régions. Cela semble logique puisque pour les facteurs explicatifs, les **régions** correspondent à l'agrégation des communes et des départements, et les **départements** correspondent à l'agrégation des communes. Ce phénomène est valable pour les déterminants de Y mais pas pour le nombre de publications lui-même, puisque celles-ci sont comptabilisées en considérant les territoires comme des organisations à part entière. Quoi qu'il en soit, une analyse tous niveaux confondus n'est pas possible. Aussi, en choisissant le meilleur rapport données disponibles / données complètes, nous décidons de mener l'analyse sur les 94 départements de France métropolitaine - les territoires d'outre-mer ayant des caractéristiques différentes.

TABLE 3 – Différences d'échelle entre les niveaux géographiques

		minimum	**maximum**	**moyenne**	**médiane**
nombre de publications	*régions*	2	303	114.6	69
	départements	0	356	32.86	5
	communes	0	374	0.17	0
	EPCI	0	631	4.92	0
nombre de créations d'entreprises	*régions*	22 158	251 781	89 300	51 438
	départements	521	76 851	9 468	6 076
	communes	0	76 851	28.42	3
	EPCI	16	175 339	1 000.6	183
dépenses totales par habitant	*régions*	388.5	581.8	483.8	496.1
	départements	819.3	3789.9	1153.6	1135.6
	communes	175.1	77188.5	1267.1	994.5
	EPCI	8.6	4195.8	655.3	574.8
part des 65 ans ou +	*régions*	14.8	23.7	19.71	20.4
	départements	11.9	30.1	21.86	21.8
	communes	0	100	22.31	21.2
	EPCI	9.6	39.6	23	22.3

Nous allons diviser cette section en trois sous-parties ; dans un premier temps nous réaliserons les manipulations essentielles avant toute analyse et présenterons les données, puis nous effectuerons une analyse exploratoire avant de modéliser les relations entre les facteurs explicatifs et Y dans une dernière partie.

8.2.1 Traitement et présentation de la base

Comme évoqué précédemment, les données nécessitent un traitement pré analyse, notamment pour les champs suivants :

• la **couleur politique** : nous avons pour l'instant l'historique de tous les partis auxquels a appartenu le président du département, le maximum étant de 5 pour le chef de

l'exécutif en place en Charente-Maritime en 2019.

- la **CSP du chef** : les niveaux de précision des données sont variables d'un département à un autre ; il s'agit parfois d'un métier, parfois de la catégorie.

Pour ce premier cas nous harmonisons les partis politiques en les regroupant dans les catégories plus générales "Droite", "Gauche" et "sans étiquette". Ainsi, nous cherchons le bord politique des 17 modalités de départ, afin de les regrouper dans ces 3 catégories. En faisant cette agrégation, nous obtenons pour chaque département une valeur - les partis politiques multiples pour un chef correspondant aux anciens noms des mouvements ; il nous suffit alors de supprimer les doublons grâce à la commande *unique()*.

Pour le second cas, nous utilisons la nomenclature de l'INSEE sur les professions et catégories socioprofessionnelles. Nous retrouvons pour chaque valeur actuelle, la CSP agrégée au premier niveau, que nous remplaçons dans les données. Après harmonisation, nous passons de 33 valeurs à 8, correspondant aux 8 premières CSP (sur 9).

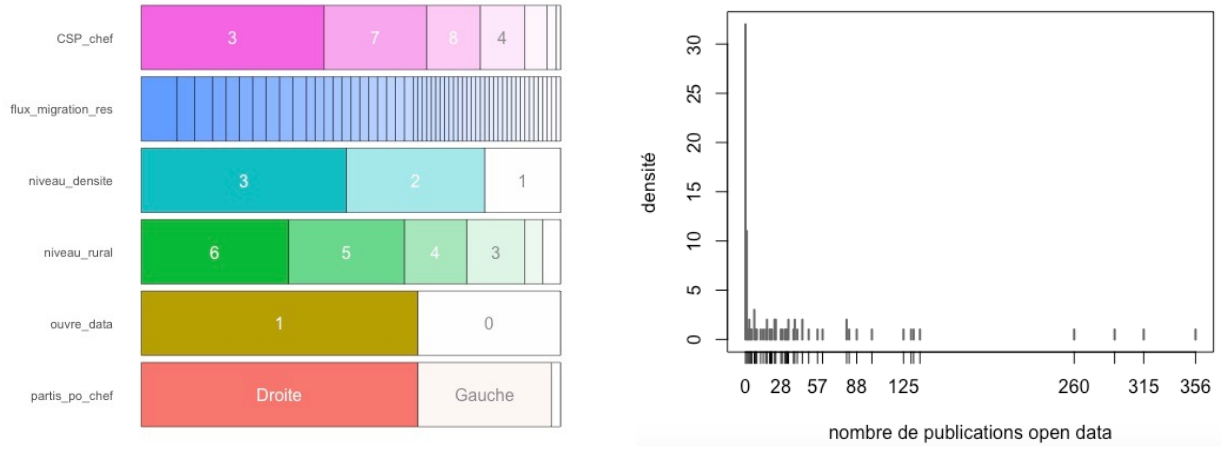
Suite au formatage des données notre base est prête à l'emploi, elle contient 94 observations et 23 variables, dont 17 facteurs explicatifs ; 12 sont de nature quantitative et 5 de nature qualitative. Nous retrouvons en annexe n°3 un **dictionnaire des variables**, utile principalement pour connaître les modalités des variables qualitatives catégorielles.

Une précision avant de commencer l'analyse ; la construction de la base se faisant en utilisant des liens vers les données en ligne ajournées régulièrement (c'est le cas de Y), il se peut que les valeurs changent légèrement et donc, qu'en exécutant le script, les résultats soient différents de ceux présentés ici - obtenus à partir des données récupérées en juillet.

8.2.2 Analyse non supervisée

L'analyse non supervisée correspond à l'étude des données sans étiquettes, c'est-à-dire sans une variable dépendante d'un côté et des variables indépendantes de l'autre, mais toutes au même niveau. Nous commençons par quelques statistiques descriptives, pour regarder la distribution de notre variable à expliquer ainsi que la répartition des variables qualitatives.

FIGURE 1 – Répartition des variables catégorielles et distribution de Y

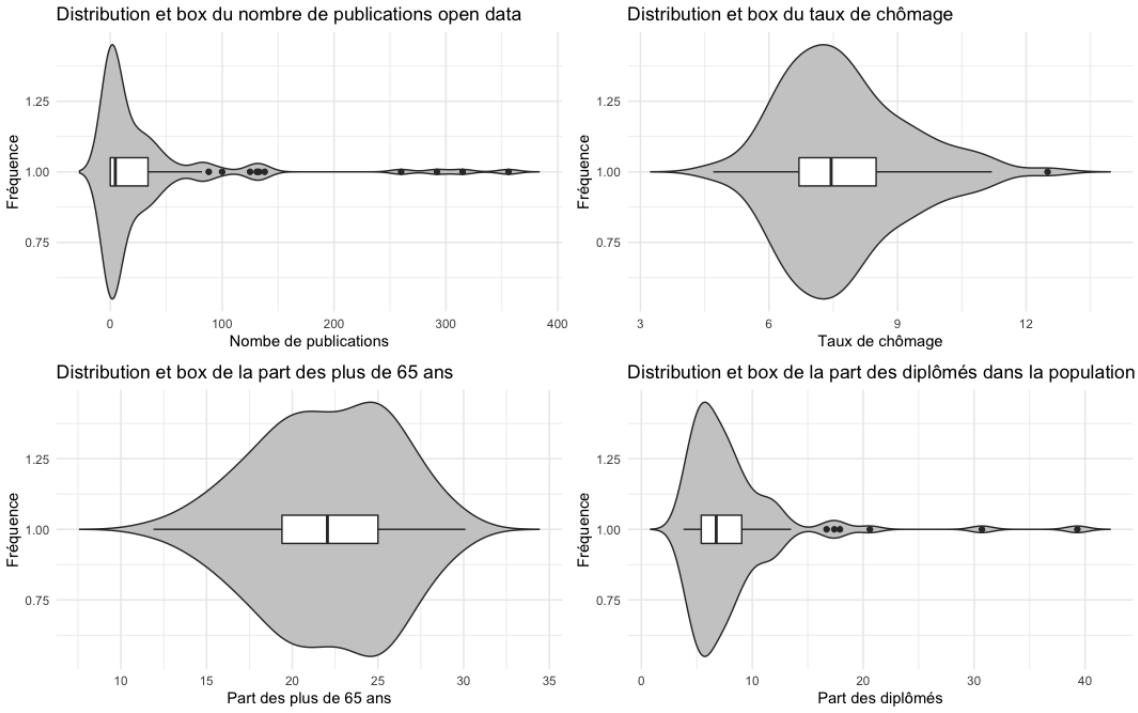


La figure ci-dessus montre la **répartition** des variables qualitatives ; nous voyons alors que la majorité des présidents de département appartiennent à la CSP n°3, c'est-à-dire aux cadres et professions intellectuelles supérieures. La variable des **flux de migration résidentielle** semble peu pertinente étant donné le nombre de modalités distinctes (55) et le peu d'observations dans chacune d'elle (maximum 8 par catégorie) ; nous l'écarterons donc pour l'analyse supervisée et les modélisations. Nous notons également que près de 50% des départements sont peu denses (catégorie n°3 de '*niveau_densite*'), ou selon la nouvelle définition de l'INSEE ; un tiers sont considérés comme ruraux autonomes très peu denses (catégorie n°6 de '*niveau_rural*'). Enfin, on remarque que deux tiers des départements sont à droite politiquement parlant (en considérant les chefs de 2019, soit avant les nouvelles élections).

Nous voyons aussi, sur la figure n°1, la **distribution** de la variable à expliquer. Ainsi, le nombre de publications varie de 0, qui concerne 32 départements de France, à 356, qui correspond à la Seine-Saint-Denis. On observe visuellement cette forte étendue de la série ; nous voyons clairement qu'elle ne suit pas une loi normale ce qui est lié à la nature même de Y qui correspond à un comptage ne pouvant être négatif.

Voyons à présent si certaines données sont atypiques, et si elles nécessitent d'être traitées en conséquent. Pour cela nous regardons les *violin plots* qui offrent la même information que des boîtes à moustaches, mais avec la courbe de distribution en plus.

FIGURE 2 – Violin plots des 4 premières variables quantitatives



Les violin plots de la figure 2 ainsi que ceux présents en annexe n°4 montrent qu'il existe certains points atypiques pour les variables quantitatives, dûs aux valeurs fortement supérieures au 3è quantile. La variable à expliquer (nombre de publications open data par département) contient elle-même 10 outliers. En effet, comme nous avons pu le voir précédemment, sa distribution est très étalée et l'écart d'ouverture de données entre les départements est assez notable. Le taux de chômage quant à lui ne possède qu'un outlier qui correspond au département des Pyrénées-Orientales, considéré comme une zone de *fragilité sociale* en termes de chômage. [34]

Au total, sur les 13 variables quantitatives (avec Y) 10 ont des points potentiellement atypiques, seuls la part des plus de 65 ans, le pourcentage de population vivant en zone rurale et l'âge du président du département n'ont pas d'outliers. Nous vérifions ce constat visuel en appliquant les tests statistiques de Grubbs pour le taux de chômage et de Rosner pour les autres.

34. GARNIER, "Pourquoi notre taux de chômage ne diminue pas ? Découvrez le paradoxe occitan".

Au vu des résultats de ces tests disponibles en annexe n°5, nous voyons que sont considérées atypiques les valeurs suivantes :

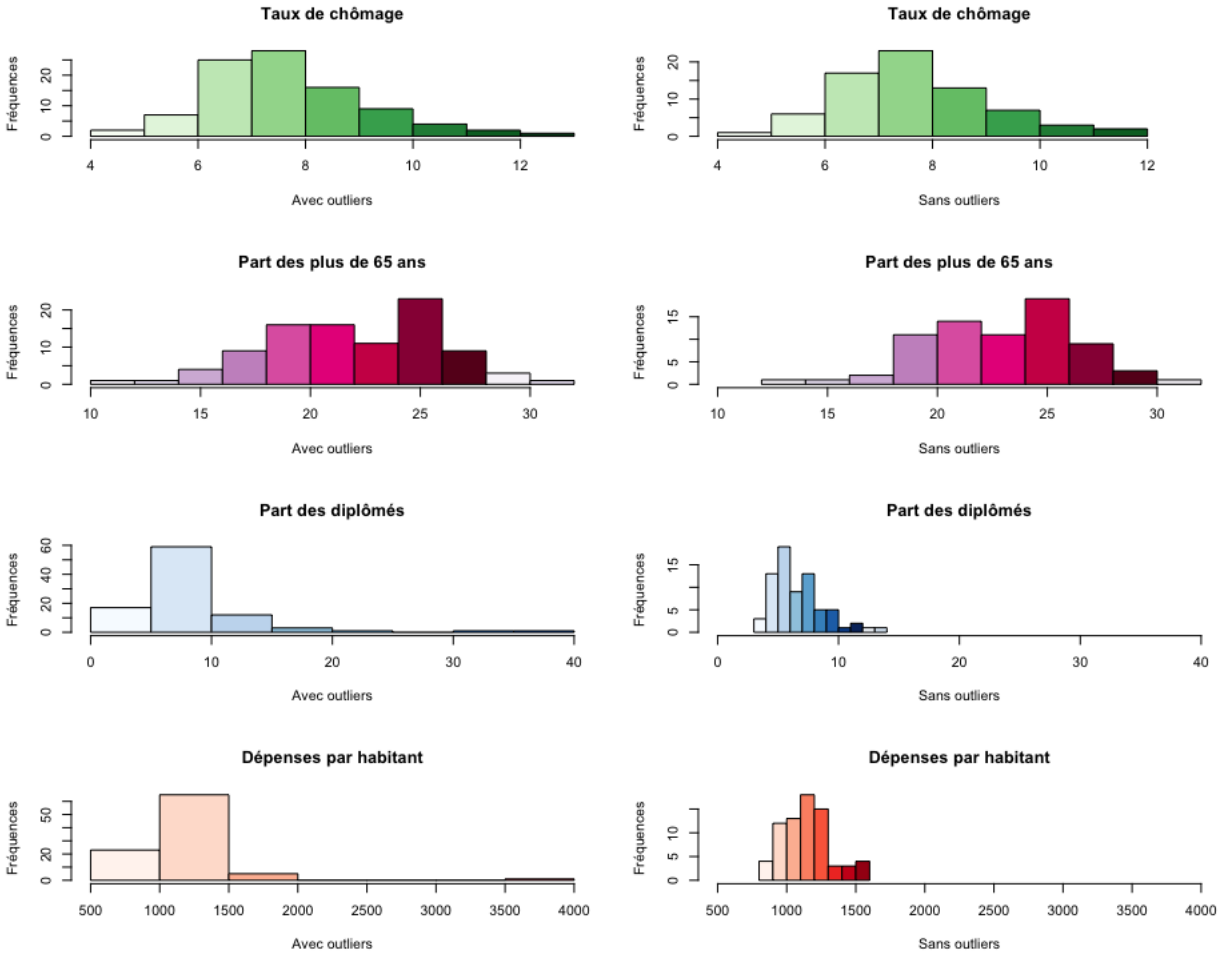
- taux de chômage $\geq$ 12.5%
- part des diplômés $\geq$ 16.7%
- dépenses par habitant $\geq$ 1 770.4 €
- part des étudiants $\geq$ 16.5% de la population
- nombre d'étudiants $\geq$ 98 938
- population $\geq$ 2 639 070 habitants
- niveau de vie $\geq$ 26 600 €
- nombre de créations d'entreprises $\geq$ 32 480
- nombre de nuitées d'hotels $\geq$ 6 724

Outre la variable à expliquer qui contient 10 outliers, le test statistique révèle 12 valeurs atypiques. Nous retirons les observations concernées dans une nouvelle base, nous passons donc de 94 à 72 ; nous avons perdu 22 observations. Le nombre de publications maximal est désormais de 80 et concerne le département du Loir-et-Cher. Nous comparons maintenant les distributions des variables avec et sans outliers.

Comme visible sur la figure n°3 et en annexe n°6, l'étendue des variables sans outliers (colonne de droite) est évidemment réduite par rapport aux distributions avec outliers (colonne de gauche). Pour la part des diplômés par exemple, le taux le plus élevé dans la base initiale était de 39.3% correspondant à Paris, et est maintenant de 13.5% correspondant à l'Essonne (Île-de-France). De même, les dépenses totales des départements s'étendaient jusqu'à 3789.95€ par habitant pour Paris, et vont désormais jusqu'à 1559.5€ pour la Creuse (Nouvelle-Aquitaine).

En outre, l'asymétrie à droite est réduite pour toutes les variables, bien que certaines demeurent loin de la loi normale. Lorsque l'on applique le test de **Shapiro** pour étudier statistiquement leur normalité, on s'aperçoit en effet que seulement 5 des 13 variables quantitatives ont une distribution normale.

FIGURE 3 – Distribution des 4 premières variables avec et sans points atypiques



Examinons à présent les **relations** entre les variables quantitatives, avant de regarder les **dépendances** entre les qualitatives. Nous faisons cette analyse et les suivantes sur les données débarrassées des points atypiques, soit aux dimensions (72x23).

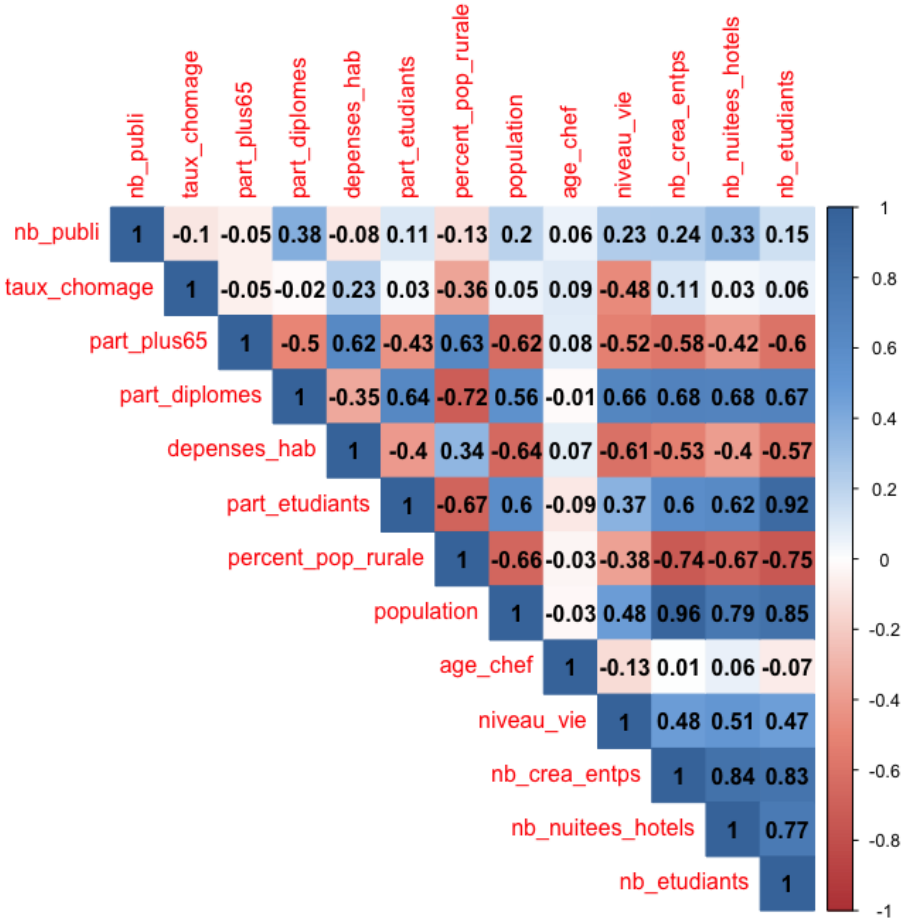
Nous trouvons en figure n°4 la **matrice** qui reprend visuellement les corrélations qui existent entre les 13 variables numériques. Nous avons fait le choix d'inclure Y dans la matrice, de manière à regarder quelles variables lui sont le plus corrélées, et donc lesquelles sont plus susceptibles de l'expliquer. Nous voyons d'emblée que des corrélations fortes existent entre certains facteurs explicatifs. Nous relevons **10 corrélations moyennes** (coefficient compris entre 0.5 et 0.6), et **24 corrélations fortes** (coefficient strictement supérieur à 0.6), et identifions des groupes de variables fortement liées entre elles :

• part des plus de 65 ans, part des diplômés, part d'étudiants, part de population vivant en zone rurale, population, nombre de créations d'entreprises, nombre de nuitées en hôtels de tourisme et nombre d'étudiants - avec presque toujours des coefficients supérieurs à 0.6 ;

• dans une moindre mesure les variables du niveau de vie et des dépenses par habitant sont corrélées à celles du groupe identifié, avec des coefficients entre 0.4 et 0.6.

Finalement, nous constatons que les relations entre les facteurs explicatifs et Y ne sont pas très fortes : la plus élevée est de 0.38 et concerne la **part des diplômés**. Toutefois, on s'aperçoit que toutes les variables du groupe identifié sont les plus fortement corrélées à Y. Par conséquent, bien qu'il faille choisir celles qui seront intégrées aux modèles pour ne pas fausser les estimations, nous pouvons supposer dès à présent qu'elles joueront un rôle important dans l'explication du phénomène d'ouverture des données.

FIGURE 4 – Corrélations liant les variables quantitatives



Nous pouvons à présent étudier les dépendances entre les variables qualitatives.

TABLE 4 – P-values des tests de Pearson sur les variables qualitatives

	Ouvre données	**Niveau ruralité**	**Niveau densité**	**Flux migration**	**Parti politique**	**CSP du chef**
Ouvre données		0.0865	0.1527	0.5058	0.5591	0.7605
Niveau ruralité			0.0055	0.0035	0.5994	0.8197
Niveau densité				0.4846	0.3149	0.7043
Flux migration					0.7299	0.6327
Parti politique						0.0370

Les **tests de dépendance** du Khi-2 appliqués aux variables qualitatives montrent ici aussi 2 groupements de variables liées entre elles :

- le niveau de ruralité, le niveau de densité et les flux de migration résidentielle, toutes trois étant des facteurs géographiques ;
- la CSP du chef et son parti politique.

Nous savions, en prenant 2 définitions différentes pour la même variable d'urbanisation des départements, qu'elles seraient dépendantes l'une de l'autre. Nous comptons cependant sur l'analyse supervisée et la sélection de variables pour savoir laquelle des deux retenir. Enfin, en regardant la dépendance des variables avec le nombre de publications converti en variable binaire (**ouvre_data** oui ou non), on voit que la plus susceptible d'expliquer Y est celle qui indique le niveau de ruralité, avec la nouvelle définition de l'INSEE de 2021.

Avant de terminer cette analyse non supervisée, nous résumons les informations dans un tableau. Nous notons alors que selon les relations que nous attendions des variables explicatives à Y, seulement 2 sont contraires à la littérature ou aux hypothèses émises en partie économique : il s'agit de l'âge du chef et des dépenses totales par habitant. Par ailleurs, leurs coefficients de corrélation sont très proches de 0, traduisant quoi qu'il en soit un faible lien avec le nombre de jeux ouverts.

TABLE 5 – Résumé de l'analyse non supervisée des variables quantitatives

	Nombre d'outliers	**Distrib. normale**	**Corrélation attendue***	**Vraie corr. avec Y**
Nombre de publications	**10**	**X**	/	/
Part des 65 ans ou +	0	✓	-	-0.05
Part de la pop. rurale	0	✓	-	-0.13
Âge du président du département	0	X	-	0.06
Nombre d'habitants	1	X	+	0.2
Taux de chômage annuel moyen	1	✓	-	-0.1
Dépenses totales par habitant	2	✓	+	-0.08
Part d'étudiants (%)	2	X	+	0.11
Nb de nuitées ds hotels de tourisme	4	X	+	0.33
Médiane du niveau de vie	4	✓	+	0.23
Nombre de création d'entreprises	5	X	+	0.24
Part des diplômés d'un BAC+5 ou +	6	X	+	0.38
Nombre d'étudiants	6	X	+	0.15

**Nature de la relation attendue avec Y (positive ou négative).*

TABLE 6 – Résumé de l'analyse non supervisée des variables qualitatives sans outliers

	Modalité dominante	**% base****	**Dépendance avec Y**
Y binaire : ouvre données	**Ouverture**	**64%**	/
Niveau de ruralité	Rural autonome très peu dense	43%	✓
Niveau de densité	Peu dense	58%	X
Flux de migration	Haute-Garonne	11%	X
Parti politique	Droite	67%	X
CSP du chef de l'exécutif	Cadres et prof. int. sup.	42%	X

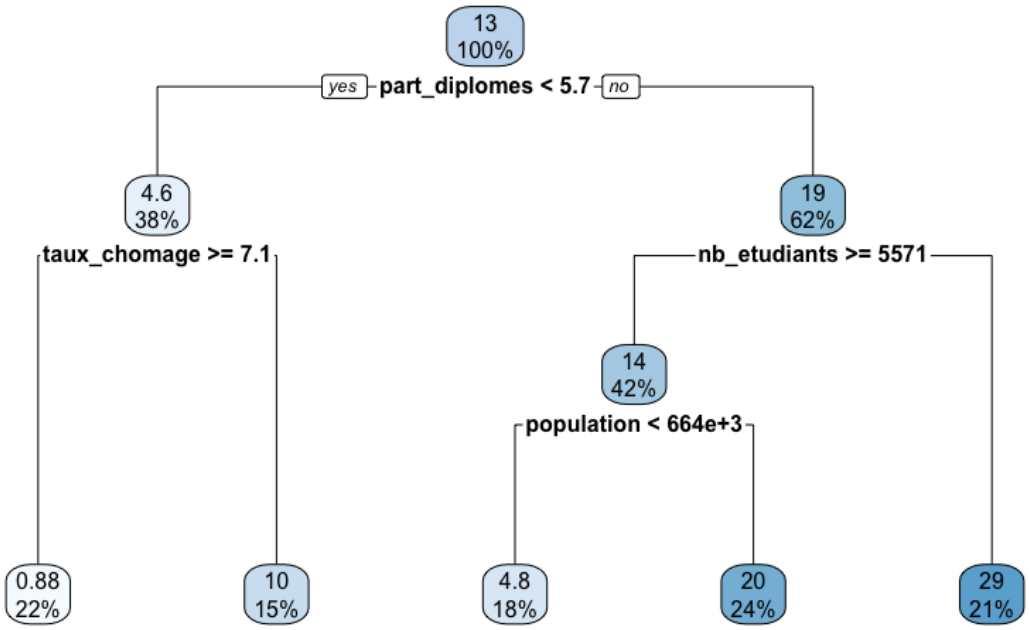
***Part de l'échantillon que représente la modalité dominante.*

8.2.3 Analyse supervisée

Dans une seconde partie, regardons plus en détail grâce aux datavisualisations, les relations qui lient les variables explicatives à Y. Pour ce faire nous étudierons les relations de Y avec les variables quantitatives dans un premier temps, puis avec les variables qualitatives.

1. Variables quantitatives

FIGURE 5 – Arbre de régression sur les données sans outliers



L'**arbre** ci-dessus découpe les observations en groupes qui se veulent les plus hétérogènes entre eux mais les plus homogènes en leur sein. D'après le résumé de cette régression, les variables les plus importantes sont les suivantes :

- part des diplômés ;
- nombre de créations d'entreprises ;
- part des étudiants ;
- population ;
- pourcentage de population vivant en zone rurale ;
- nombre d'étudiants.

En outre, il s'agit du **groupement de variables** (6 variables présentes sur 8) identifié à partir de la matrice de corrélation dans la section précédente, ce qui confirme notre hypothèse selon laquelle ces déterminants jouent un rôle important dans l'explication du nombre de données ouvertes.

Intéressons-nous à présent à l'arbre lui-même. En partant de l'échantillon complet (72 observations) avec une moyenne de 13 jeux ouverts par département, on voit que la première division s'effectue avec la **part des diplômés d'un BAC+5 ou plus dans la population**. Il apparaît ainsi que les départements ayant moins de 5.7% de diplômés ont une ouverture moyenne de 4 à 5 jeux, alors que ceux qui ont 5.7% ou plus ouvrent en moyenne 19 jeux, soit 4 fois plus. Ces 2 sous-groupes représentent respectivement 38 et 62% de l'échantillon. Parmi les départements de cette première catégorie, ceux dont le taux de chômage égale ou excède 7.1% ne publient pas de données ou très peu (1 jeu), tandis que ceux qui ont moins de 7.1% de chômage ouvrent 10 jeux en moyenne. Finalement, la sous-population où l'ouverture de données est la plus importante correspond aux départements avec au moins 5.7% de diplômés, et moins de 5571 étudiants ; ils publient alors 29 jeux en moyenne et représentent 21% des 72 départements débarrassés des valeurs atypiques.

Nous avons alors 5 sous-populations identifiées, qui représentent plus ou moins 20% de l'échantillon total. Sachant que la variable de la part des diplômés est la plus importante et celle qui divise au mieux l'échantillon initial, nous décidons de souligner cela en créant une variable binaire "*part_diplomes_5.65*" qui prend la valeur "1" lorsque cette part est au moins égale à 5.65%, et prend la valeur "0" dans le cas contraire.

Dans un même but de comprendre et mesurer les interactions existantes entre les facteurs explicatifs, nous réalisons à présent une **analyse en composantes principales**. Visible en annexe n°7, l'inertie associée aux composantes est de 66.2% pour les deux premières et 74.8% en considérant la troisième, ce qui signifie que les 3 axes fictifs créés récupèrent trois quarts de l'information contenue dans les variables quantitatives. Définissons à présent, en regardant la contribution ainsi que la corrélation des variables aux 3 dimensions, la nature des axes créés.

TABLE 7 – Contributions et corrélations des variables aux dimensions 1, 2 et 3

	Variables	**Contribution**	**Corrélation**
Axe n°1	*Nombre de créations d'entreprises*	12.21%	0.88
	Population	12.17%	0.87
	Nombre d'étudiants	12.07%	0.87
	Part des diplômés	11.97%	0.87
	Pourcentage population rurale	11.08%	-0.83
	Nombre de nuitées dans hôtels	10.21%	0.8
	Part des plus de 65 ans	9.55%	-0.77
	Part des étudiants	8.8%	0.74
Axe n°2	*Taux de chômage*	45.6%	0.87
	Médiane du niveau de vie	22.89%	-0.62
	Dépenses par habitant	10.4%	0.42
Axe n°3	*Âge du chef*	71.54%	0.86

Comme nous le voyons dans le tableau ci-dessus, la dimension n°1 est exactement composée du groupe de variables identifié en regardant la matrice de corrélations ; nous constatons par ailleurs que les contributions des 8 variables qui la composent sont quasiment identiques allant de 8.8 à 12.21%. De même, leurs corrélations à l'axe sont très fortes puisque les coefficients (en valeur absolue) sont proches de 1 ; les variables se situent donc à l'extrémité de l'axe, près du cercle de corrélation. A contrario, pour l'axe 2 la contribution des trois variables est plutôt inégale ; c'est principalement le taux de chômage qui est représenté, puis la médiane du niveau de vie à 22.89% et les dépenses pas habitant à 10.4%. Enfin, seule la variable de l'âge du chef de l'exécutif contribue significativement à l'axe n°3. En outre, nous écartons cette troisième dimension de l'ACP qui n'offre pas d'interprétation des variables, avec, de surcroît, une perte d'information de 29%.

Au vu des informations contenues dans la table n°7, nous pouvons définir les axes de l'ACP de telle manière :

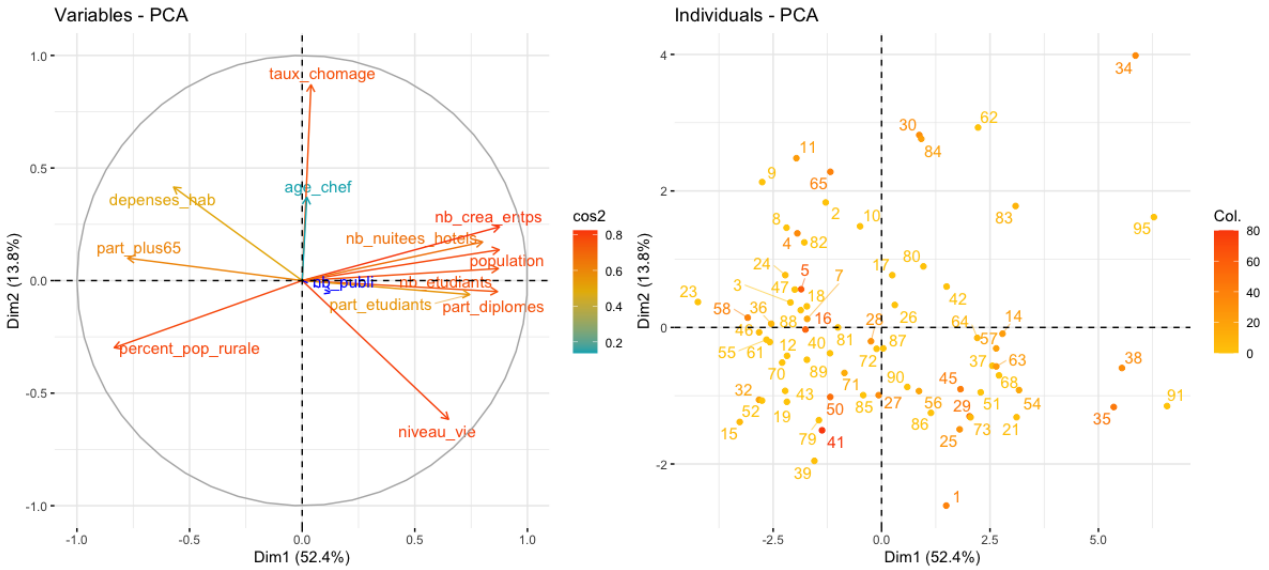
- la première dimension correspond au **dynamisme** du département reflété par une

population nombreuse, jeune (beaucoup d'étudiants et peu de personnes de plus de 65 ans), diplômée, urbaine, et par une économie attractive : créations d'entreprises et tourisme, mesuré par les nuitées recensées dans les hôtels de tourisme.

- la deuxième dimension peut être définie par la **pauvreté** du département ; en effet selon les variables qui la composent, elle correspond à un fort taux de chômage, un faible niveau de vie médian donc un faible revenu disponible, ainsi que des dépenses par habitant élevées qui peuvent être la conséquence du faible pouvoir d'achat des habitants (par exemple des dépenses en allocations ou en subventions aux entreprises).

Nous regardons à présent la projection des variables et des départements sur le plan composé des 2 dimensions que nous venons de définir.

FIGURE 6 – Projection des variables et des départements sur le plan factoriel



Nous retrouvons visuellement le groupe de variables le long de l'axe n°1, avec le pourcentage de population rurale à l'opposé car comme nous l'avons vu sa corrélation avec la première composante est négative. On distingue la variable du nombre de publications proche du barycentre, traduisant une dépendance faible aux axes créés. Cependant, on voit qu'elle se situe dans le quart inférieur droit du cercle, ainsi le nombre de publications serait plus important pour les départements dynamiques et plus aisés. En outre, les hypothèses émises en première partie seraient vérifiées : une dynamique initiée par une activité prospère placerait le dépar-

tement dans un cercle vertueux, favorable à la mis en place de projets open data pour ouvrir petit à petit ses données, ou continuer une démarche existante.

Sur le graphique de droite de la figure n°6 sont projetés les départements identifiables par leur COG. La coloration des points correspond au nombre de jeux ouverts. En raisonnant par la définition des axes et la position de Y sur le cercle de corrélation, nous devrions trouver plus de publications (points rouges) dans le quart inférieur droit du graphique. Cependant, les 4 départements qui publient le plus de données (respectivement 80, 61, 57 et 50 jeux) et qui correspondent aux départements 41, 5, 16 et 50, se situent dans la partie gauche du graphique, traduisant un faible dynamisme. Aussi, nous pouvons comparer le positionnement des départements ouvrant le même nombre de données :

- **29 et 35** (correspondant au Finistère et à l'Ille-et-Vilaine) ont publié tous deux 45 jeux. On remarque qu'ils se situent dans le quart inférieur droit, c'est-à-dire relativement dynamiques et aisés.

- **1 et 65** (correspondant à Ain et aux Hautes-Pyrénées) ont publié 39 jeux. Ce premier se situe dans le quart inférieur droit mais le second est à son parfait opposé dans le quart supérieur gauche, traduisant de la pauvreté et un faible dynamisme ; ils ont pourtant le même nombre de données ouvertes.

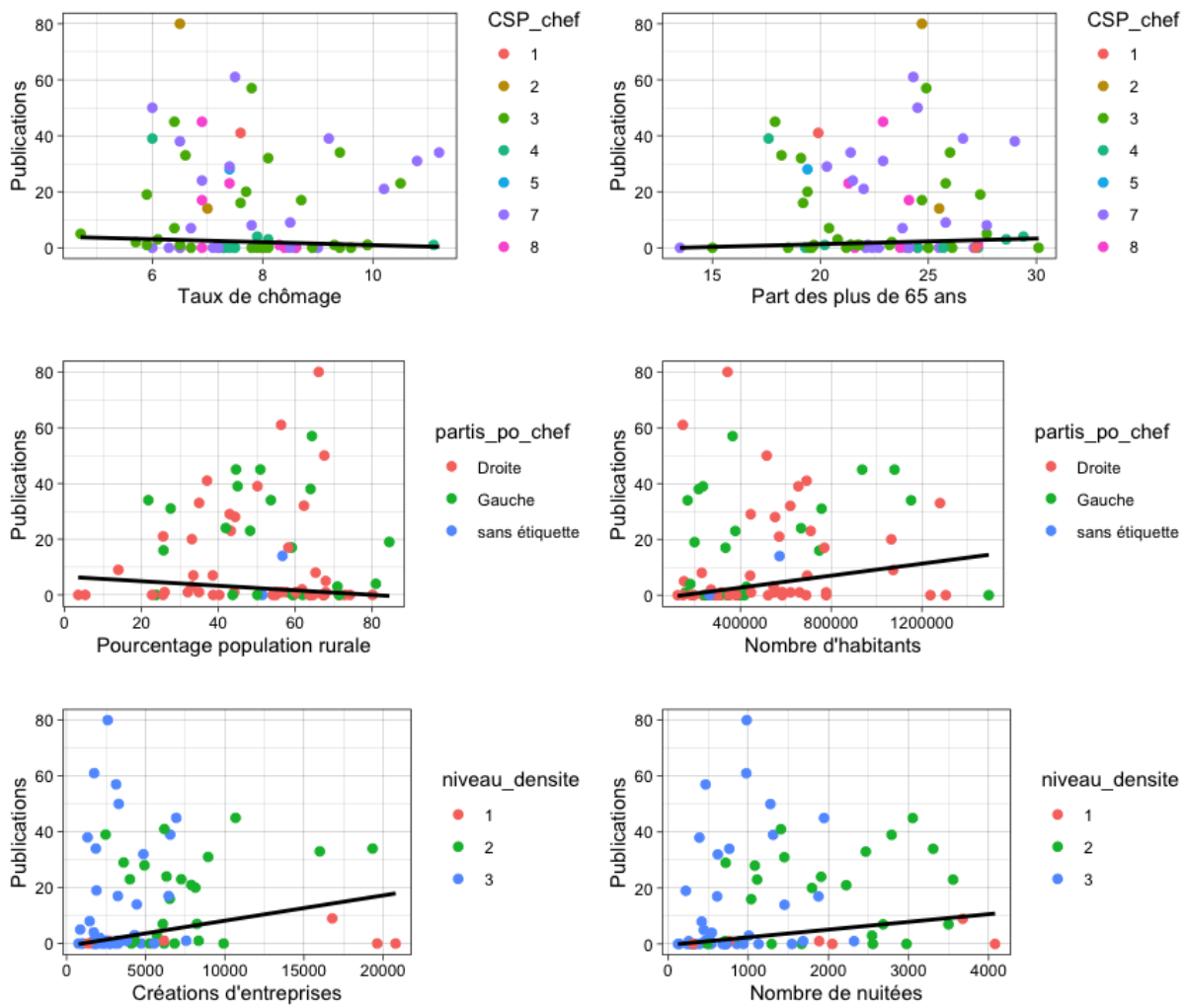
- **4 et 34** (correspondant aux Alpes-de-Haute-Provence et au Hérault) ont publié 34 jeux. La situation est encore différente ici puisque le Hérault se trouve tout en haut à droite du plan et donc a, a priori, une économie dynamique, mais est considéré comme un département 'pauvre' à cause d'un taux de chômage élevé (11.2%). D'un autre côté, les Alpes-de-Haute-Provence sont définies comme pauvres et peu dynamiques, par la place sur le graphique, mais publient elles aussi 34 jeux de données.

Cette projection des départements sur le plan factoriel semble montrer un certain côté "*aléatoire*" du nombre de publications open data. Des départements ayant la même maturité open data ont des caractéristiques socio-économiques très différentes (sauf dans le cas étudié du Finistère et de l'Ille-et-Vilaine). De plus, d'un simple coup d'oeil on ne peut identifier sur le graphique des individus, des groupements de départements situés au même endroit et

ayant approximativement le même nombre de publications (points avec la même coloration). Cela peut être lié à 2 choses : soit le phénomène dépend de variables non prises en compte dans l'analyse, soit il est trop récent et dépend de facteurs non mesurables (expérience des administrations, besoins des habitants...). Quoi qu'il en soit, nous tirerons de vraies conclusions grâce aux modélisations ; celles-ci permettront par exemple de voir quelle part de Y les variables permettent d'expliquer. De l'ACP nous gardons les axes que nous ajoutons comme facteurs explicatifs à la base, de manière à tester leur pertinence lors des estimations.

Continuons notre démarche d'exploration des données en représentant Y en fonction des variables quantitatives grâce à des nuages de points où nous colorons les observations à partir de certaines variables catégorielles.

FIGURE 7 – Nuages de points entre Y et 6 variables quantitatives



À partir de la figure n°7 nous voyons que les relations entre le nombre de publications open data et les variables quantitatives ne sont pas aisément identifiables. Ayant dans cette base débarrassée des outliers 26 départements ne pratiquant pas l'open data, il est difficile de conclure sur la nature de la relation avec la variable quantitative représentée. Nous pouvons néanmoins affirmer qu'elle n'est pas linéaire, et donc qu'une telle régression ne conviendrait pas à nos données. Par ailleurs, la coloration des points avec 3 variables catégorielles (CSP du chef, couleur politique et niveau de densité), nous donne une information supplémentaire. La CSP du chef de l'exécutif contenant trop de modalités, on ne constate pas de groupements particuliers. En revanche, il semble que la couleur politique et le niveau de densité divisent bien les variables quantitatives puisque des groupes se distinguent correctement, mais moins clairement pour Y car ils se distinguent davantage sur l'axe horizontal que vertical. Nous verrons statistiquement la pertinence des variables catégorielles dans l'explication de Y, grâce aux modèles.

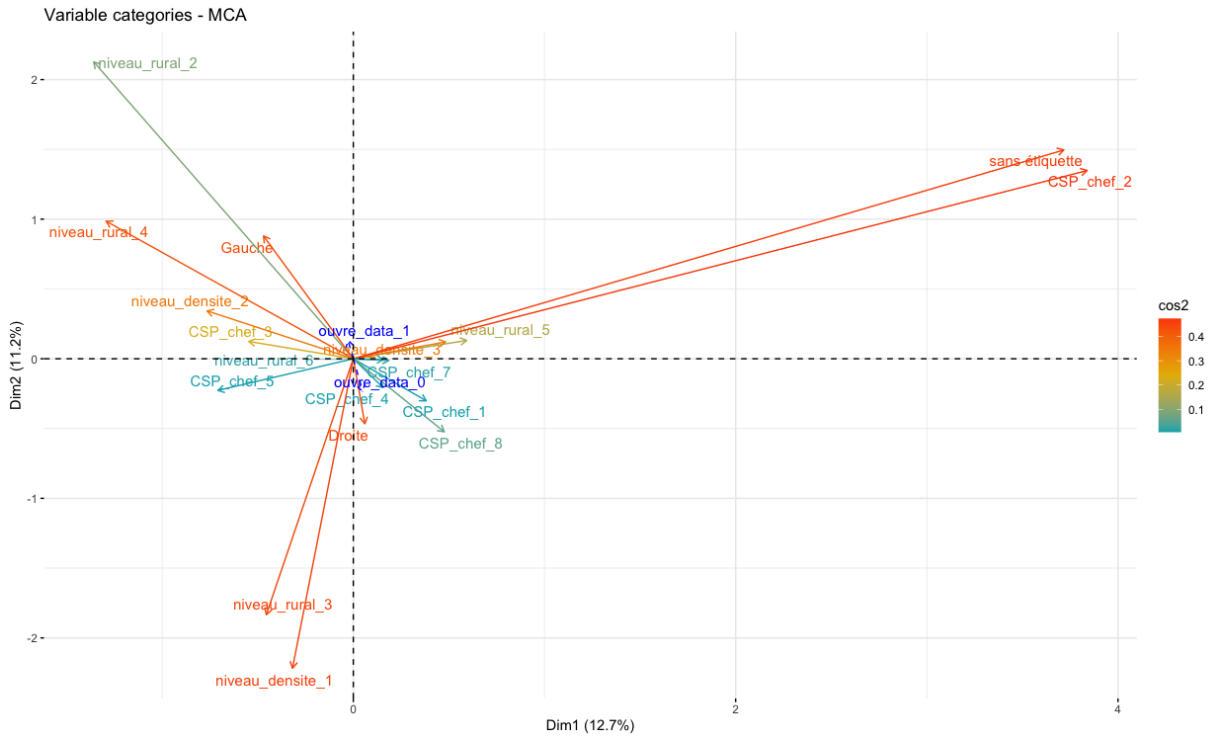
Passons désormais à l'analyse des variables qualitatives, croisées de différentes manières avec le nombre de publications open data.

2. Variables qualitatives

Nous effectuons dans un premier temps une **analyse des correspondances multiples** qui, comme l'ACP, a pour objectif de mettre en exergue les relations entre les variables en les rassemblant autour d'axes fictifs.

La figure n°8 présente l'ACM appliquée sur nos données sans outliers. On constate que les modalités des variables qualitatives sont reparties sur tout le plan factoriel. Cependant, on remarque d'emblée que les modalités de la variable à expliquer *binaire* se situent toutes deux proches du barycentre, traduisant une faible dépendance aux axes fictifs de l'ACM. Il est donc difficile de conclure sur les relations entre Y et les variables puisque ses modalités ne sont pas positionnées clairement dans un quart du graphique. Étant donné que nous nous intéressons dans cette partie aux relations entre les variables et le phénomène d'ouverture de données, et pas seulement entre les variables explicatives, nous ne continuons pas l'ACM puisque la variable à expliquer se situe trop proche de l'origine du graphique.

FIGURE 8 – Projection des variables sur le plan factoriel



Nous regardons par ailleurs les interactions entre les données de type qualitatif et Y, en croisant les variables. Ainsi, nous retrouvons en figure n°9 des diagrammes qui montrent le nombre de jeux ouverts médian par modalité de variables. Nous examinons simultanément ces graphiques et la table n°8 qui indique quant à elle le nombre d'observations par modalité. Ainsi nous vérifions la représentativité des modalités avant de tirer toute conclusion.

À première vue, il apparaît sur les diagrammes croisés que le nombre de publications médian est plus élevé pour les départements de densité intermédiaire, et pour ceux dont le président est de gauche et appartient à la CSP 'Artisans, commerçants et chefs d'entreprise'. Seulement, lorsque l'on regarde la représentativité des modalités grâce au tableau n°8, nous nous apercevons que le nombre de publications médian de certaines modalités ne peut être interprété, étant donné le peu d'observations qui les composent. C'est le cas des modalités suivantes qui n'en contiennent que 1 ou 2 :

- **modalité n°2 du niveau de ruralité**, c'est-à-dire "urbain densité intermédiaire" ;

- **modalité n°3 de la couleur politique**, c'est-à-dire "sans étiquette" ;

- **modalités n°1, 2 et 5 de la CSP du chef**, c'est-à-dire "agriculteurs exploitants", "artisans, commerçants et chefs d'entreprise", et "employés".

En outre, en considérant cela et le déséquilibre entre les autres modalités, il semblerait que l'ouverture de données soit, en médiane, plus présente dans les départements denses et à gauche. Ainsi, 50% des départements **denses** publient moins de 21 jeux de données et 50% publient plus de 21 jeux, alors que parmi les 42 départements peu denses, 50% ne pratiquent pas l'open data (Y=0) et 50% publient plus d'un jeu de données. De même, 11 des 22 départements de **gauche** publient plus de 18 jeux et 11 en publient moins de 18, pour les 48 départements de gauche, la moitié publie plus de 7 jeux et l'autre moins de 7.

FIGURE 9 – Diagrammes croisés ; nombre de publications moyen par modalité de variable

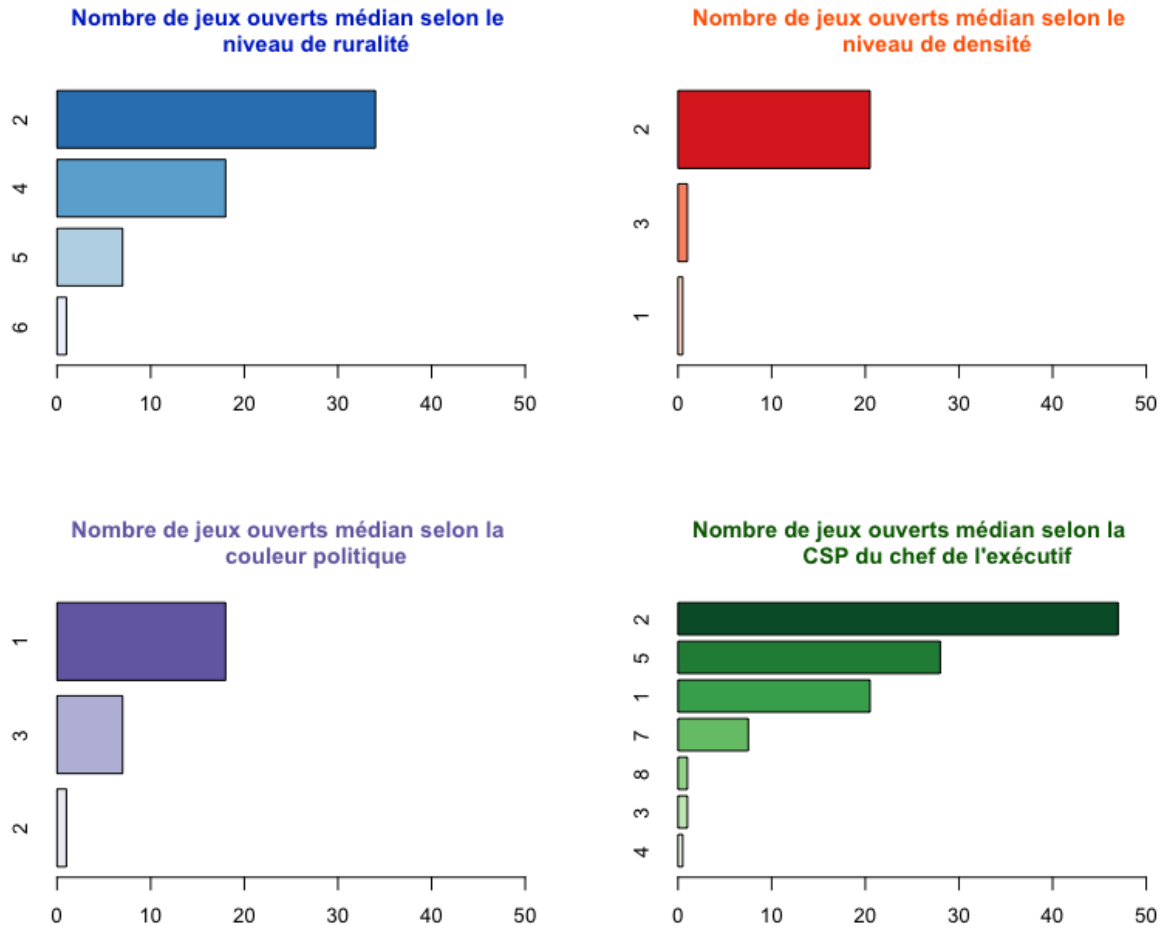


TABLE 8 – Répartition des observations dans les modalités des variables catégorielles

	N° mod.	**Modalités**	**Nb d'obs.**
Niveau de ruralité	2	*Urbain densité intermédiaire*	1
	3	*Rural sous forte influence d'un pôle*	8
	4	*Rural sous faible influence d'un pôle*	10
	5	*Rural autonome peu dense*	22
	6	*Rural autonome très peu dense*	31
Niveau de densité	1	*Très dense*	6
	2	*Dense*	24
	3	*Peu dense*	42
Couleur politique	1	*Gauche*	22
	2	*Droite*	48
	3	*Sans étiquette*	2
CSP du chef	1	*Agriculteurs exploitants*	2
	2	*Artisans, com. et chefs d'entps.*	2
	3	*Cadres et prof. int. sup.*	30
	4	*Professions Intermédiaires*	8
	5	*Employés*	1
	7	*Retraités*	22
	8	*Autres pers. sans act. pro.*	7

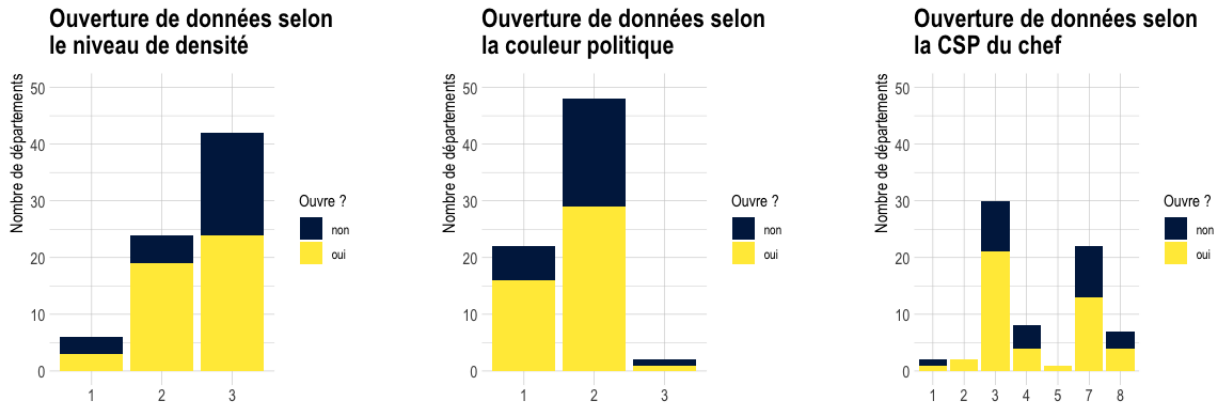
Enfin, nous terminons cette partie analytique des variables qualitatives en visualisant la variable à expliquer en **mode binaire** (ouvre des données : oui / non) en fonction de ces dernières.

Les graphiques à barres empilées visibles en figure n°10 montrent le nombre de départements contenus dans chaque modalité des variables, nous pouvons alors comparer la proportion de départements ouvrant ou n'ouvrant pas de données dans une même catégorie. Par exemple, nous constatons que le nombre de départements denses (niveau de ruralité n°2) qui ouvrent des données est proportionnellement supérieur au nombre de départements

denses qui n'ouvrent pas de données. Il en va de même pour les départements politiquement à gauche et ceux dont le président est cadre ou exerce une "*profession intellectuelle supérieure*".

Alors que les diagrammes croisés représentaient le nombre de publications médian (Y quantitatif) pouvant être biaisé par la sous ou la sur représentativité de certaines modalités, les graphiques ci-dessous présentent le nombre de départements ouvrant ou n'ouvrant pas de données par modalités. Il n'y a alors aucun biais d'échantillon puisque l'on compare ici le nombre de départements au sein d'une modalité et non au sein d'une variable. Toutefois, les constats posés sur les variables équilibrées à partir des premiers graphiques sont confirmés par cette deuxième série de graphiques croisés.

FIGURE 10 – Nombre de départements ouvrant des données par catégorie de variables



Nous avons, tout au long de cette analyse supervisée, pu faire de nombreux constats et émettre quelques hypothèses que nous résumons à présent dans un tableau, avant de les vérifier par les modélisations économétriques.

Les variables quantitatives listées sur les 2 premières lignes du tableau correspondent aux divisions des deux extrémités de l'arbre de régression : aussi doivent-elles être considérées simultanément. Les départements qui ont un nombre de publications moyen le plus élevé sont ceux ayant au moins 5.65% de diplômés d'un BAC+5 ou plus dans leur population **et** ayant moins de 5571 étudiants inscrits dans l'enseignement supérieur. De même, les variables du "dynamisme" et de la "pauvreté" se rapportent aux axes de l'ACP, définis à partir

des variables quantitatives qui les composent. Le "**dynamisme**" englobe ainsi les 8 variables identifiées avec la matrice de corrélation, à savoir *'part des plus de 65 ans', 'part des diplômés', 'part d'étudiants', 'part de population vivant en zone rurale', 'population', 'nombre de créations d'entreprises', 'nombre de nuitées en hôtels de tourisme'* et *'nombre d'étudiants'*. La "**pauvreté**" quant à elle, est définie à presque 50% par un taux de chômage élevé, puis par une médiane de niveau de vie basse et, dans une moindre mesure, des dépenses par habitant fortes. Enfin, les 3 dernières variables mentionnées dans le tableau font référence aux constats que nous avons émis grâce aux différents graphiques croisant les variables qualitatives et Y.

TABLE 9 – Identification des sous-populations ouvrant le plus ou le moins de données

OUVERTURE FORTE		**OUVERTURE FAIBLE**	
Variable	**Valeur**	**Variable**	**Valeur**
Part des diplômés	$\geqslant 5.65\%$	*Part des diplômés*	$< 5.65\%$
Nombre d'étudiants	< 5571	*Taux de chômage*	$\geqslant 7.1\%$
Dynamisme	élevé	*Dynamisme*	faible
Pauvreté	faible	*Pauvreté*	élevée
Niveau de densité	dense	*Niveau de densité*	peu dense
Couleur politique	gauche	*Couleur politique*	droite
CSP du chef	Cadres et prof. int. sup.	*CSP du chef*	Prof. intermédiaires*

**Sous réserve de la représentativité de cette modalité où l'on trouve 8 départements, soit 11% de l'échantillon.*

Pour rappel nous avons créé 3 variables à partir de cette analyse exploratoire :

- le **dynamisme** du département, première composante de l'ACP ;
- la **pauvreté** du département, deuxième composante de l'ACP ;
- '**part_diplomes_5.65**' construite à partir de la part des diplômés, définie par l'arbre comme la variable la plus importante et qui divise au mieux l'échantillon total.

8.3 Modélisations

Dans cette dernière partie nous allons chercher le meilleur modèle pour répondre à la problématique, c'est-à-dire celui qui valide les hypothèses fondamentales propres à la méthode utilisée, et qui a de bonnes performances économétriques.

Nous traiterons cette partie en deux temps : recherche du meilleur modèle et sélection, puis interprétation des résultats.

8.3.1 Recherche du meilleur modèle

Comme expliqué en partie n°2 (méthodologie économétrique), le phénomène étudié étant de nature quantitative discrète, nous décidons d'appliquer des modèles linéaires généralisés (GLM) pour identifier les déterminants de l'ouverture de données des territoires.

Les variables socio-économiques que nous avons récoltées étant fortement corrélées les unes aux autres, nous procédons dans un premier temps à la sélection de variables que nous inclurons dans une première régression par la **loi de Poisson**.

TABLE 10 – Sélection de variables pour modéliser le nombre de publications

Backward	**Forward**	**Stepwise**
• Variable binaire '*part diplômés*' $\geqslant 5.65\%$ • CSP du chef • Niveau de densité • Couleur politique	• Variable binaire '*part diplômés*' $\geqslant 5.65\%$ • CSP du chef • Nombre d'étudiants • Part d'étudiants • Créations d'entreprises • Population	• Variable binaire '*part diplômés*' $\geqslant 5.65\%$ • CSP du chef • Niveau de densité • Couleur politique

Présents en table 10, les résultats des méthodes de sélection ascendantes, descendantes et *stepwise* (qui combine ces deux premières) montrent que les 3 méthodes sélectionnent la variable binaire créée à partir de l'arbre de régression, montrant sa pertinence élevée. Aussi, les méthodes '*backward*' et '*stepwise*' retiennent toutes deux le niveau de densité du

département, la CSP du chef de l'exécutif et sa couleur politique. Toutefois, c'est la modalité 3 de la couleur politique qui est significative (sans étiquette), or elle ne regroupe que 2 observations. Ainsi, nous incluons dans un premier temps cette variable aux modèles, mais resterons très prudents sur son interprétation.

Enfin, la méthode descendante dite "*forward*" sélectionne en plus de la binaire sur la part des diplômés et de la CSP du chef, la part et le nombre d'étudiants, les créations d'entreprises et la population. Cependant, nous savons grâce à l'analyse exploratoire que ces 4 variables sont fortement corrélées (avec un coefficient de corrélation supérieur à 0.6), donc nous ne pouvons les inclure toutes les 4 dans un même modèle. Par conséquent nous choisissons d'inclure celle dont la significativité est la plus importante, c'est-à-dire le nombre de créations d'entreprises.

En outre, nous estimons un premier modèle composé de la **variable binaire sur la part des diplômés**, du **niveau de densité**, du **nombre de créations d'entreprises**, de la **CSP** ainsi que du **parti politique** du chef de l'exécutif.

Nous estimons ce premier modèle qui prend en compte des variables caractérisant les départements entre 2018 et 2021, pour expliquer le nombre de publications open data de 2021. Pour rappel, nous avons fait l'hypothèse que ces quelques décalages dans l'année de collecte des facteurs ne sont pas problématiques pour expliquer le phénomène d'ouverture de données. Nous observons donc les résultats de cette première estimation dans le tableau n°11.

À partir de cette table, nous voyons avant toute chose que la grande majorité des coefficients estimés pour les variables et leurs modalités sont significatifs, ce qui peut être le signe d'une **surdispersion**. En outre, nous décidons de vérifier cela avant de tirer quelconque conclusion. Nous évaluons la dispersion en comparant la moyenne et la variance de la série à expliquer : le nombre de publications qui s'étend de 0 à 80.

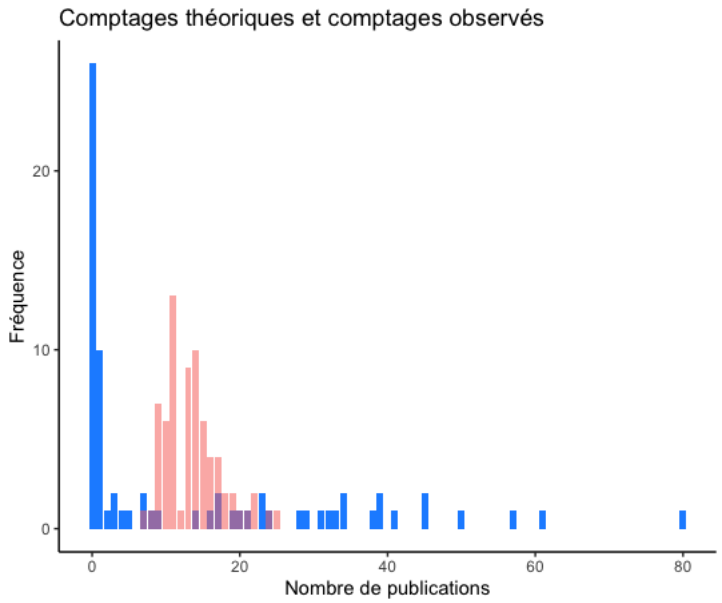
TABLE 11 – Estimation d'un premier GLM sur la base sans outliers

Variables	**Coef**	**Err.Std**	**p-value**
Constante	-0.837	0.390	0.032 *
Part des diplômés $\geqslant 5.65\% - 1$	1.668	0.0114	0.000 ***
Part d'étudiants	-0.072	0.024	0.003 **
Niveau de densité - 2	2.451	0.330	0.000 ***
Niveau de densité - 3	3.073	0.346	0.000 ***
Créations d'entreprises	0.000	0.000	0.000 ***
CSP du chef - 2	0.497	0.203	0.014 *
CSP du chef - 3	-1.258	0.178	0.000 ***
CSP du chef - 4	-1.836	0.227	0.000 ***
CSP du chef - 5	0.092	0.252	0.715
CSP du chef - 7	-0.469	0.171	0.006 **
CSP du chef - 8	-1.273	0.203	0.000 ***
Parti politique du chef - gauche	0.464	0.076	0.000 ***
Parti politique du chef - SE	-1.353	0.275	0.000 ***

Nous trouvons en figure n°11 les valeurs théoriques de Y en rouge et les valeurs observées en bleu. Les valeurs théoriques des comptages ont été simulées selon une distribution de Poisson de paramètre $\lambda = \mathrm{E(Y)} = 13.44$. On remarque alors des distributions très différentes ; la variable du nombre de publications ne suit pas une distribution de Poisson de paramètre $\lambda = 13.44$. On observe aussi une certaine inflation en zéro dans les comptages observés ; nous allons donc estimer les modèles adéquats, après vérification statistique des constats visuels, émis à partir de ce graphique n°11.

Les calculs révèlent que la variance de la série Y est de 341.77 soit 25 fois plus que la moyenne, ce qui confirme amplement l'hypothèse de surdispersion. De plus, en calculant le ratio de dispersion nous constatons qu'il est de 14.59, ce qui confirme une fois de plus le fait que la condition de la dispersion homogène pour la loi de Poisson n'est pas acceptée.

FIGURE 11 – Diagnostic de la surdispersion



Pour résoudre ce problème nous estimons dans un deuxième temps un **modèle Binomial Négatif**. Nous vérifions ensuite la distribution du nombre de données ouvertes, et nous remarquons que 26 départements sur 72 ne pratiquent pas l'open data, ce qui représente 26% de la base. Nous décidons donc de considérer cette **inflation en zéro** en appliquant les modèles adaptés : c'est-à-dire les modèles à inflation zéro pour la loi de Poisson ou pour la loi Binomiale Négative. Néanmoins, la deuxième composante de ces modèles peut aussi prendre les valeurs Y=0, traduisant alors un phénomène indépendant de la volonté des départements. Or ce n'est pas le cas dans cette analyse, c'est pourquoi nous estimons de la même manière des **modèles "double hurdle"** où la loi de la deuxième composante est censurée (Y strictement supérieur à 0). Nous supposons que ce type de modèle sera plus adapté aux données, mais effectuons aussi les modèles à inflation de zéro, pour pouvoir comparer les résultats.

Étant donné que la première partie de ces modèles est axée sur la variable binaire : ouvrir ou ne pas ouvrir de données, les variables pour l'expliquer peuvent être différentes de celles pour expliquer le **nombre** de publications. Par conséquent, nous effectuons une nouvelle sélection de variables par les méthodes pas à pas, afin d'identifier les facteurs de l'ouverture de données. Nous retrouvons les résultats des sélections en table 12.

TABLE 12 – Sélection de variables pour modéliser le fait d'ouvrir ses données (Y binaire)

Backward	**Forward**	**Stepwise**
• Variable binaire '*part diplômés*' $\geqslant 5.65\%$ • CSP du chef • Niveau de densité • Couleur politique • Population • Créations d'entreprises • Part des étudiants • Nombre d'étudiants • Nuitées en hôtels • Niveau ruralité • Part des diplômés • Âge du chef	• Variable binaire '*part diplômés*' $\geqslant 5.65\%$ • CSP du chef • Niveau de densité • Couleur politique	• Variable binaire '*part diplômés*' $\geqslant 5.65\%$ • CSP du chef • Niveau de densité • Couleur politique • Population • Créations d'entreprises • Part des étudiants • Nombre d'étudiants • Nuitées en hôtels • Niveau ruralité • Part des diplômés • Âge du chef • Dynamisme • Pauvreté

Les méthodes sélectionnent toutes trois les mêmes 5 variables que lorsque nous avions appliqué les méthodes **backward** et **bidirectionnelle** pour expliquer le nombre de publications (Y quantitatif) en table 10. Seulement ici pour expliquer Y en mode binaire, davantage de variables sont retenues par ces mêmes méthodes : il s'agit principalement des variables du groupement observé avec la matrice de corrélation, qui sont fortement corrélées entre elles. En outre, parmi ce groupement de variables nous décidons de ne garder que celle de la part des diplômés dichotomisée, de la CSP et du parti politique du chef, du niveau de densité ainsi que de l'âge du chef qui n'est pas corrélé aux autres variables. Ainsi, par rapport à la partie de **comptage** dans le modèle, nous avons les mêmes facteurs explicatifs mis à part l'âge du chef que nous ajoutons ici, dans la partie **binomiale** du modèle.

TABLE 13 – Résumé des estimations GLM

	Modèle	**Variables (Y quanti)**	**Variables (Y quali)**	**AIC**
Loi Poisson				
- modèle 1	**classique**	• *CSP chef* • *Parti politique* • *Niveau de densité* • *Part des diplômés >5.65%* • *Créations d'entreprises*	/	1164.77
- modèle 2	**ZI**	• *CSP chef* • *Parti politique* • *Niveau de densité* • *Part des diplômés >5.65%*	• *Part diplômés >5.65%*	779.26
- modèle 3	**hurdle**	• *CSP chef* • *Parti politique* • *Niveau de densité* • *Part des diplômés >5.65%*	• *Part diplômés >5.65%*	779.24
Loi Binom. Nég.				
- modèle 4	**classique**	• *CSP chef* • *Parti politique* • *Niveau de densité* • *Part des diplômés >5.65%*	/	461.33
- modèle 5	**ZI**	• *Niveau de densité* • *Part des diplômés >5.65%*	1	459.04
- modèle 6	**hurdle**	• *Niveau de densité* • *Part des diplômés >5.65%*	• *Part diplômés >5.65%*	454.78

Nous retrouvons en table n°13 le résumé des différents modèles estimés ; avec la loi de Poisson ou la loi Binomiale Négative, de manière classique, en inflation à zéro (ZI) ou en double hurdle. Pour chacune des estimations, nous avons retiré au fur et à mesure les variables qui n'étaient pas significatives. Ainsi par exemple, pour le modèle n°5, *i.e.* de la loi binomiale négative estimée en tenant compte de l'inflation en zéro, aucun facteur n'est significatif dans la partie binomiale du modèle. Par ailleurs, nous constatons que seule la variable binaire de la part des diplômés est en mesure d'expliquer le fait d'ouvrir des données ou non dans

les autres estimations. Aussi, en regardant la valeur du critère d'Akaike, il semblerait que le meilleur modèle soit le sixième puisqu'il le minimise. Toutefois, avant de conclure sur le meilleur modèle, nous appliquons les tests du log de vraisemblance et de *Vuong* pour voir statistiquement lequel est le plus adapté aux données. De même, nous regardons pour les modèles estimés à partir de la loi binomiale négative, la significativité du logarithme de θ, qui indique s'il est pertinent d'utiliser cette loi plus que la loi de Poisson.

TABLE 14 – Comparaison des modèles estimés

Test	**Modèle 1**	**Modèle 2**	**Statistique**	**p-value**	**Conclusion**
odTest	Poisson	Bin. Négative	705.44	0.000	Bin Négative
Vuong	Poisson	ZIP	-3.427	0.000	ZIP
Vuong	Bin. Négative	ZINB	2.458	0.007 (raw)	Bin. Négative
			-0.481	0.315 (AIC)	ZINB
			-3.828	0.000 (BIC)	ZINB
Vuong	Poisson	HP	-3.394	0.000	HP
Vuong	Bin. Négative	HNB	0.638	0.262 (raw)	Bin. Négative
			-0.768	0.221 (AIC)	HNB
			-2.369	0.009 (BIC)	HNB
Vuong	ZIP	HP	-0.021	0.491	HP
Vuong	ZINB	HNB	-1.009	0.156	HNB

Nous observons en tableau n°14 les résultats des tests de comparaison effectués sur les modèles estimés. Nous avons recensé les différentes p-value des tests lorsque, selon le critère considéré, la conclusion n'était pas la même. Le modèle mentionné en colonne "conclusion" correspond au meilleur des 2 modèles testés conjointement.

Nous constatons alors que la loi binomiale négative est plus adaptée que la loi de Poisson sur une estimation classique, comme nous pouvons le voir grâce au test concernant la sur-dispersion (odTest). Ensuite, nous voyons que les estimations "classiques", que ce soit pour la loi de Poisson ou la binomiale négative, sont toujours moins bonnes que les estimations

par modèle à inflation de zéro ou bien par double hurdle. En revanche, pour ces 2 types de lois, les tests révèlent qu'il n'y a pas de différence significative entre les modèles ZI et les *hurdles* (H). Par conséquent, nous nous intéressons à la significativité du paramètre θ dans les modèles estimés avec la loi binomiale négative. Ce dernier n'est significatif au seuil de 5% ni dans le modèle ZINB, ni dans le HNB, par conséquent il n'est pas pertinent d'utiliser cette loi. C'est donc la **loi de Poisson** qui correspond le plus à nos données, bien que les valeurs du critère d'Akaïke pour ce type de modèle soient plus importantes, comme nous avons pu le voir en table n°13. Au sein des estimations utilisant la loi de Poisson (gris clair dans la table 14), les modèles hurdles et ZI sont supérieurs au modèle classique, mais il n'existe pas de différence significative entre ces 2 derniers (nous voyons effectivement que la p-value est supérieure à 0.05). De plus, nous avions vu en tableau n°13 que les variables significatives sont les mêmes, et que les critères AIC sont quasiment égaux, étant respectivement de 779.26 et 779.24.

Lorsque nous regardons les valeurs des coefficients estimés par ces 2 modèles, nous remarquons qu'ils sont très proches pour la partie de comptage, en revanche le coefficient de la variable *Part des diplômés > 5.65%* qui permet d'expliquer le fait d'ouvrir ses données ou non (partie binomiale du modèle), est lui négatif dans une estimation et positif dans l'autre. Or, nous avons vu que les *double hurdles* prennent en compte le fait que la réalisation de l'évènement (Y=1) est la conséquence d'une décision prise volontairement, donc semblent plus appropriés à notre contexte d'étude. De plus, le coefficient de la variable pour la partie binaire y est positif, ce qui correspond à la relation théorique espérée entre ces variables.

Nous gardons donc le modèle "double hurdle" estimé avec la loi de Poisson comme modèle final, que nous interprétons dans une dernière partie.

8.3.2 Interprétation des résultats

TABLE 15 – Modèle final : double hurdle en loi de Poisson sur la base sans outliers

	Variables	Coef	e ^{coef}	Err.Std	p-value
Y binaire	*Constante*	1.539	4.662	0.385	0.000 ***
	Part des diplômés $\geqslant 5.65\% - 1$	0.848	2.334	0.120	0.000 ***
	Niveau de densité - 2	1.327	3.798	0.331	0.000 ***
	Niveau de densité - 3	1.878	6.538	0.335	0.000 ***
	CSP du chef - 2	0.117	1.124	0.213	0.581
	CSP du chef - 3	-1.351	0.259	0.181	0.000 ***
	CSP du chef - 4	-1.713	0.180	0.234	0.000 ***
	CSP du chef - 5	-0.381	0.683	0.245	0.120
	CSP du chef - 7	-0.551	0.576	0.172	0.001 **
	CSP du chef - 8	-1.231	0.292	0.207	0.000 ***
	Parti politique du chef - gauche	0.447	1.563	0.073	0.000 ***
	Parti politique du chef - SE	-0.895	0.408	0.314	0.004 **
Y quantitatif	*Constante*	-0.375	0.688	0.392	0.339
	Part des diplômés $\geqslant 5.65\% - 1$	1.628	5.091	0.531	0.002 **

*Niveaux de significativité : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Nous calculons pour ce modèle final le pseudo R^2 sous R : celui-ci est de 0.429, ce qui est relativement correct. Ayant retiré préalablement les variables qui n'étaient pas significatives dans l'explication de Y, toutes sont désormais significatives.

Nous interprétons à présent les coefficients des 2 parties du modèle l'une après l'autre, en commençant par la partie qui explique le fait d'ouvrir ou de ne pas ouvrir de données. Pour interpréter les coefficients, nous procédons à une transformation pour avoir les effets réels sur la variable dépendante, telle que : **effet** $= e^{coef}$.

Le coefficient estimé étant positif, la part des diplômés supérieure à 5.65% a un impact significatif positif sur la probabilité d'ouvrir des données. Plus précisément, un département

ayant au moins 5.65% de diplômés dans sa population a une probabilité d'ouvrir ses données 5.09 fois supérieure à un département ayant moins de 5.65% de diplômés.

Nous n'interprétons ici que les modalités contenant suffisamment de modalités. Effectivement, les modalités 1, 2 et 5 de la CSP du chef contiennent moins de 3 observations. Sachant que la catégorie n°1 doit être la catégorie de référence, nous ne préférons pas tirer de conclusions sur cette variable puisque les comparaisons en termes d'ouverture de données se ferraient alors par rapport à une sous-population composée de seulement 2 départements. En revanche, pour les autres variables et à l'aide de la partie de **comptage** du modèle, on peut dire que :

- un département **dense** a un nombre de publications multiplié par 3.77 par rapport à un département très dense (modalité de référence de cette variable)

- un département **peu dense** a un nombre de publications multiplié par 6.54 par rapport à un département très dense

- un département politiquement **à gauche** a 1.56 fois plus de publications qu'un département de droite (nous n'interprétons pas la modalité "sans étiquette" qui ne contient que 2 observations)

- un département ayant **5.65% de diplômés ou plus** a 2.33 fois de publications qu'un département ayant moins de 5.65% de diplômés

Par conséquent, les effets empiriques des variables sur Y sont pour la plupart cohérents avec la théorie. Cependant, pour le niveau de densité par exemple nous attendions une relation négative avec Y : ici un département peu dense a a priori plus de publications open data qu'un département très dense. Ici encore, ce résultat peut être relativisé par le fait qu'il n'y a que 6 observations dans la catégorie de référence, étant donné que les départements considérés comme atypiques, et donc retirés de la base étaient ceux à forte densité.

Nous voyons par ailleurs que la nature des relations est vérifiée pour les variables de la part des diplômés ainsi que de la couleur politique.

9 Conclusion

Étant donné la place grandissante de la **data** en ce XXIè siècle et la nécessité d'être rendue compréhensible et accessible par tous, cette étude visait à identifier les déterminants de l'ouverture de données des territoires français.

Ayant récolté des données pour différents niveaux géographiques (collectivités et EPCI), nous avons par la suite décidé de cibler l'analyse à l'ouverture de données des 94 départements de France métropolitaine. Ainsi, nous avons cherché à expliquer le **nombre de publications *open data*** de ces collectivités, par différents facteurs pouvant être regroupés en 4 grandes classes ; économiques, politiques, démographiques et géographiques. En divisant la partie application en une analyse exploratoire poussée puis en des modélisations, nous avons pu tirer de nombreuses conclusions permettant de répondre, en partie, à la problématique de ce travail.

Quels facteurs influencent l'ouverture de données des départements ?

L'exploration des données dans un premier temps a permis de révéler la nature des relations liant les variables, les liens de causalité existant entre les facteurs explicatifs et la variable cible, ou encore les sous-populations de départements avec des probabilités plus grandes d'ouvrir des données etc. En outre, nous avons vu comment la **part des diplômés** pouvait impacter l'ouverture de données ; lorsque celle-ci est supérieure ou égale à 5.65% de la population le nombre de publications open data est en moyenne de 19, tandis qu'il n'est que de 5 pour les départements ayant moins de 5.65% de diplômés. De même, le groupe où l'ouverture de données est la plus importante est celui avec au moins 5.65% de diplômés **ET** moins de 5571 étudiants inscrits dans l'enseignement supérieur. A contrario, on retrouve une faible ouverture de données pour les départements ayant moins de 5.65% de diplômés et 7.1% de chômage ou plus.

De la même manière, nous avons pu étudier les liens entre les variables en réalisant une ACP dans le but de grouper les déterminants selon leurs corrélations ; ils ont été ainsi regroupés autour de 2 composantes principales : le **dynamisme** et le **niveau de richesse**/d'activité du département. De cette façon, nous avons vu que les départements susceptibles d'ouvrir

des données sont ceux qui ont un contexte socio-économique favorable caractérisé par une activité soutenue. Cependant, en comparant les départements ayant des caractéristiques communes, nous nous sommes rendu compte que le nombre de publications n'était pas toujours en adéquation avec le contexte socio-économique du territoire. Nous avons ainsi pu relever quelques départements qui, par leur économie en bonne santé, sont en mesure d'ouvrir mais n'ont pour l'instant initié aucune démarche.

Cette analyse en composantes principales a donc mis en lumière une partie **non mesurée** du phénomène ; l'ouverture de données dépendrait aussi de facteurs non récoltés dans cette étude, ou non récoltables s'ils ne peuvent pas être quantifiés.

Par ailleurs, l'exploration des variables de type qualitatif a montré un nombre plus important de publications dans les départements **politiquement à gauche**, et **denses** par rapport aux départements peu denses. Par conséquent, outre les facteurs économiques, l'analyse exploratoire a montré la pertinence des variables politiques et géographiques pour expliquer et mesurer l'ouverture de données.

Après cette phase de visualisation des données, nous avons estimés différents **modèles de comptage** pour expliquer au mieux Y. Celui-ci peut être décomposé en 2 phénomènes : le fait d'ouvrir ses données (variable binaire) et le nombre de données ouvertes (variable quantitative). Nous avons donc appliqué un modèle **double hurdle** qui a l'avantage de mesurer ces 2 évènements à la fois, avec des facteurs explicatifs différents pour chaque partie. En outre, un département qui a au moins 5.65% de diplômés dans sa population a une probabilité d'ouvrir ses données 5.09 fois supérieure à un département ayant moins de 5.65% de diplômés. Un département **dense** a un nombre de publications multiplié par 3.77, et un département **peu dense** a un nombre de publications multiplié par 6.54 par rapport à un département très dense. De même, un département politiquement **à gauche** a 1.56 fois plus de publications qu'un département de droite. Enfin, un département ayant **5.65% de diplômés ou plus** a 2.33 fois de publications qu'un département ayant moins de 5.65% de diplômés.

Nous avons ainsi pu conclure sur les effets des variables indépendantes sur le nombre de publications open data. Par rapport aux relations attendues entre les facteurs et la variable à expliquer les liens sont cohérents dans l'ensemble, excepté pour le niveau de densité où nous attendions plus de publications pour les départements très denses. Néanmoins, nous avons pu relativiser ce constat en montrant la sous-représentativité de la modalité de référence de cette variable. Nous retrouvons ainsi en annexe n°8, la répartition des départements par niveau de densité, sur la base de données complète et sur celle débarrassée des valeurs atypiques.

De surcroît, le nombre restreint d'observations dans de nombreuses catégories nous a freinés dans les interprétations et donc les conclusions concrètes du problème. S'ajoute à cela la difficulté présentée précédemment, à savoir lorsque des caractéristiques identifiant les départements les plus susceptibles d'ouvrir sont mises en avant, mais que le nombre de publications reste trop aléatoire pour être certain de l'effet d'une variable. Malgré ces difficultés, l'étude a été très intéressante car elle a permis de révéler, entre autres, que les départements sont parfois **en capacité** économique de mener à bien un projet d'ouverture de données, mais ne le font pas.

En outre, proposer un accompagnement aux départements concernés peut être une opportunité de collaboration, puisque ces derniers seront en mesure d'aller au bout de la démarche grâce à une activité soutenue ou un dynamisme économique en place. Cela concerne par exemple le Territoire de Belfort, Essonne, le Haut-Rhin, Marne, les Landes, ou encore le Tarn-et-Garonne qui n'ont pour le moment initié aucune démarche d'ouverture de données.

De la même manière, il pourrait être pertinent de proposer un accompagnement d'ouverture plus importante pour les départements n'ayant actuellement que 1 jeu ouvert, et étant en mesure d'ouvrir plus. Nous pouvons nous demander par ailleurs, s'il s'agit des débuts d'une démarche open data, ou si les départements en question ne souhaitent publier qu'un jeu de données, de manière à respecter la **loi Lemaire** au strict minimum. Pour cela nous vérifions les portails open data des quelques territoires caractérisés par un bon dynamisme et ne publiant actuellement qu'une base de données, pour voir si les nombres de publications sont les mêmes. Il s'agit des départements de Vienne, Ariège et Indre-et-Loire.

Après vérification, nous constatons que le département d'Indre-et-Loire compte au 23 août 6 publications, Vienne a toujours 1 jeu ouvert et pour Ariège, en allant voir leur page dédiée sur datagouv, nous ne voyons aucune publication open data. Cela montre le manque de fiabilité et l'instabilité des données, rendant l'analyse plus complexe.

Quoiqu'il en soit, ce travail a permis de montrer que le nombre de publications open data ne dépend pas d'une seule variable mais est explicable par un contexte, favorable ou défavorable à la mise en place de politiques open data, et mesuré par les différentes variables de cette étude.

10 Bibliographie

Articles

[2] N. Blanpain. “L’espérance de vie par niveau de vie : chez les hommes, 13 ans d’écart entre les plus aisés et les plus modestes”. In : *INSEE Première* (Février 2018).

[3] S. Chignard. “Une brève histoire de l’Open Data”. In : *Paris Innovation Review* (Mars 2013).

[6] P. Garnier. “Pourquoi notre taux de chômage ne diminue pas ? Découvrez le paradoxe occitan”. In : *Made In Perpignan* (Novembre 2019).

[7] A. Garrone. “La France de nouveau sur le podium de l’open data en 2019”. In : *Le blog d’Etalab* (Novembre 2020).

[8] S. Goëta. “Rapport sur l’Open Data”. In : (Décembre 2019).

[13] S. junior. “Droite et gauche : histoire d’un clivage politique”. In : (Décembre 2013).

[14] B. Lepine. “Open Data définition : qu’est-ce que c’est ? À quoi ça sert ?” In : *LeBigData.fr* (Novembre 2019).

[15] P. Moynot. “La droite et la gauche expliquées à ma fille”. In : *Le Monde* (Mars 2012).

[16] P. S. Numérique. “Initiatives autour de l’inclusion numérique des personnes âgées”. In : *Société Numérique* (Novembre 2019).

[18] O. B. Olga. “Qu’est-ce que le « rural » ? Analyse des zonages de l’Insee en vigueur depuis 2020”. In : *Géo confluences* (Février 2021).

Sites internet

[1] L.-D. Benyayer et S. Chignard. *La CADA fête ses 40 ans : retour sur les premiers pas de la transparence administrative*. URL : https://donneesouvertes.info/. (accessed : 13.08.2021).

[4] CNAM. *Modèles linéaires généralisés*. URL : https://maths.cnam.fr/IMG/pdf/Presentation_MODGEN_02_2007.pdf. (accessed : 21.05.2021).

[5] Data.gov.be. *Quels sont les avantages de l'Open Data ?* URL : https://data.gov.be/fr/faq/quel-sont-les-avantages-de-lopen-data. (accessed : 13.08.2021).

[9] INSEE. *Définition du secteur tertiaire / Tertiaire*. URL : https://www.insee.fr/fr/metadonnees/definition/c1584. (accessed : 15.08.2021).

[10] INSEE. *Définition du taux de chômage*. URL : https://www.insee.fr/fr/metadonnees/definition/c1687. (accessed : 15.08.2021).

[11] INSEE. *Documentation fichier détail : Migrations résidentielles*. URL : https://www.insee.fr/fr/information/2383290. (accessed : 10.05.2021).

[12] INSEE. *Méthode d'agrégation pour obtenir la grille de densité à un niveau géographique supra communal*. URL : https://www.insee.fr/fr/statistiques/fichier/2114627/methode-agregation.pdf. (accessed : 17.08.2021).

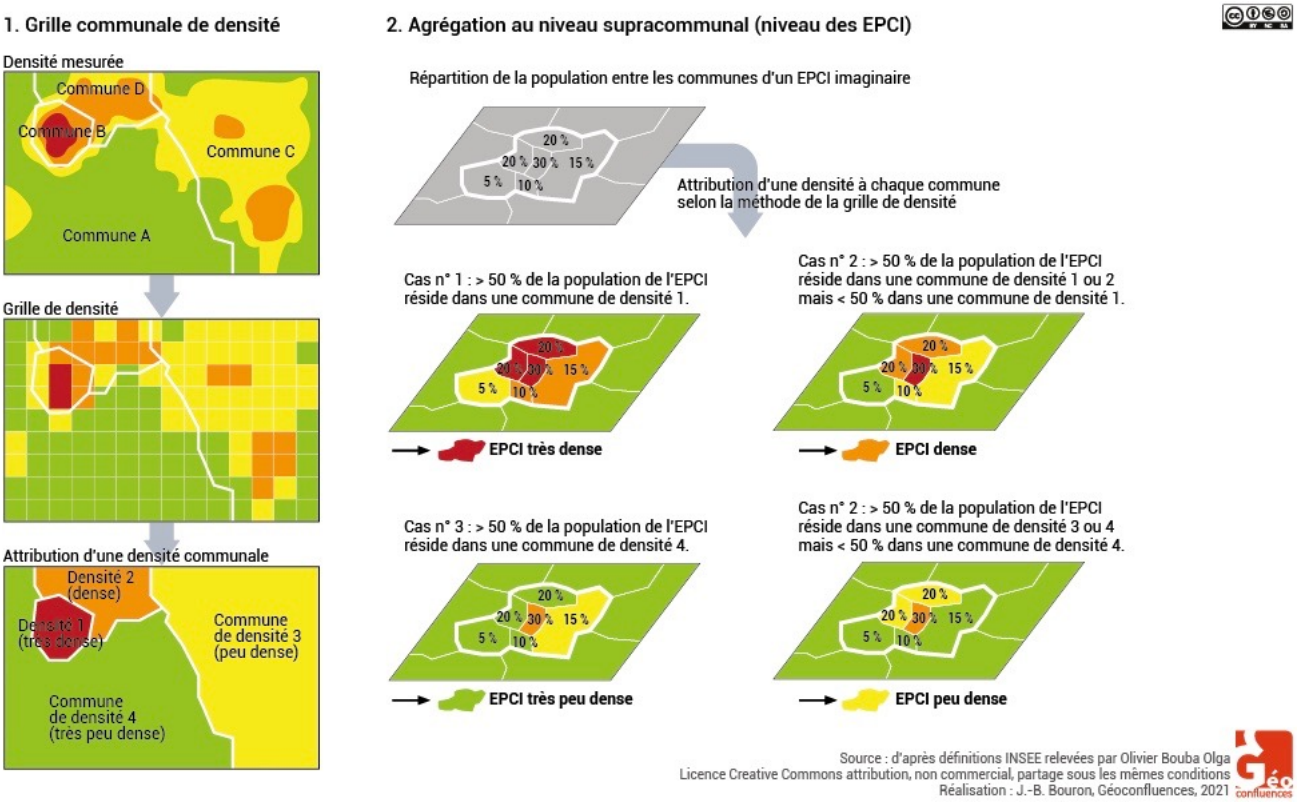
[17] OFGL. *Méthodologie des agrégats financiers - dépenses totales*. URL : https://data.ofgl.fr/pages/methodologie-agregats-financiers/. (accessed : 11.05.2021).

[19] OpenDataFrance. *Données de la carte de l'observatoire open data des territoires*. URL : https://www.data.gouv.fr/fr/datasets/donnees-de-la-carte-de-lobservatoire-open-data-des-territoires/. (accessed : 17.08.2021).

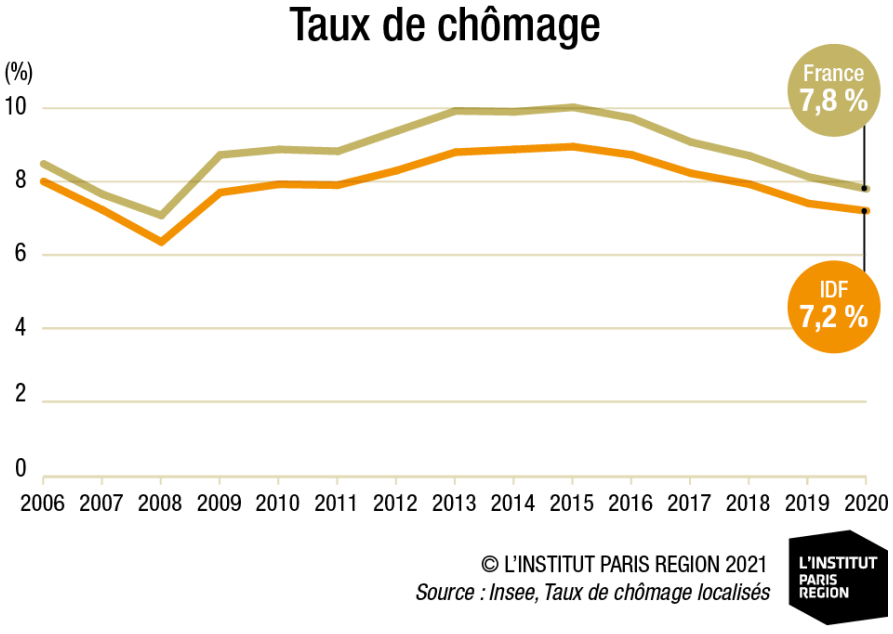
[20] C. D. Vedova. *GLM sur données de comptage (régression de Poisson) avec R*. URL : https://delladata.fr/tutoriel-glm-sur-donnees-de-comptage-regression-de-poisson-avec-r/. (accessed : 21.05.2021).

11 Annexes

Annexe n°1 : Agrégation de la grille communale de densité à l'échelle des EPCI.



Annexe n°2 : Évolution du taux de chômage en Île-de-France de 2006 à 2020.



Annexe n°3 : Dictionnaire des variables.

nom : Nom du département

...	Ain, Aisne, Allier, Alpes-de-Haute-Provence [...]

nb_publi : Nombre de publications open data

...	coefficient numérique

ouvre_data : Département ouvre des données ou non

0	N'ouvre pas
1	Ouvre

code_region : COG de la région

11	Ile-de-France
24	Centre-Val de Loire
27	Bourgogne-Franche-Comté
28	Normandie
32	Hauts-de-France
44	Grand Est
52	Pays de la Loire
53	Bretagne
75	Nouvelle-Aquitaine
76	Occitanie
84	Auvergne-Rhône-Alpes
93	Provence-Alpes-Côte d'Azur

nb_ptf : Nombre de données publiées sur une platforme local

...	coefficient numérique

nb_datagouv : Nombre de données publiées sur datagouv

...	coefficient numérique

population : Nombre d'habitants

...	coefficient numérique

CSP_chef : Catégorie socio-professionnelle du chef de l'exécutif

1	Agriculteurs exploitants
2	Artisans, commerçants et chefs d'entreprise
3	Cadres et professions intellectuelles supérieures
4	Professions Intermédiaires
5	Employés
7	Retraités
8	Autres personnes sans activité professionnelle

age_chef : Âge du chef de l'exécutif

...	coefficient numérique

partis_po_chef : Orientation politique prépondérante

1	Gauche
2	Droite
3	Sans étiquette

part_plus65 : Pourcentage de personnes âgées d'au moins 65 ans

...	coefficient numérique

niveau_vie : Médiane du niveau de vie (en €)

...	coefficient numérique

part_diplomes : Pourcentage de la population diplômée d'un BAC+5 ou plus

...	coefficient numérique

taux_chomage : Taux de chômage

...	coefficient numérique

nb_crea_entps : Nombre d'entreprises créées en une année

...	coefficient numérique

nb_nuitees_hotels : Nombre de nuits recensées en hôtels de tourisme

...	coefficient numérique

flux_migration_res : Flux principal de migration résidentielle

3	Allier
7	Ardèche
8	Ardennes
13	Bouches-du-Rhône
14	Calvados
...	[...]

depenses_hab : Dépenses totales par habitant (en €)

...	coefficient numérique

nb_etudiants : Nombre d'étudiants inscrits dans l'enseignement supérieur

...	coefficient numérique

part_etudiants : Pourcentage d'étudiants inscrits dans l'enseignement supérieur

...	coefficient numérique

niveau_rural : Niveau de ruralité (nouvelle définition de l'INSEE)

1	Urbain dense
2	Urbain densité intermédiaire
3	Rural sous forte influence d'un pôle
4	Rural sous faible influence d'un pôle
5	Rural autonome peu dense
6	Rural autonome très peu dense

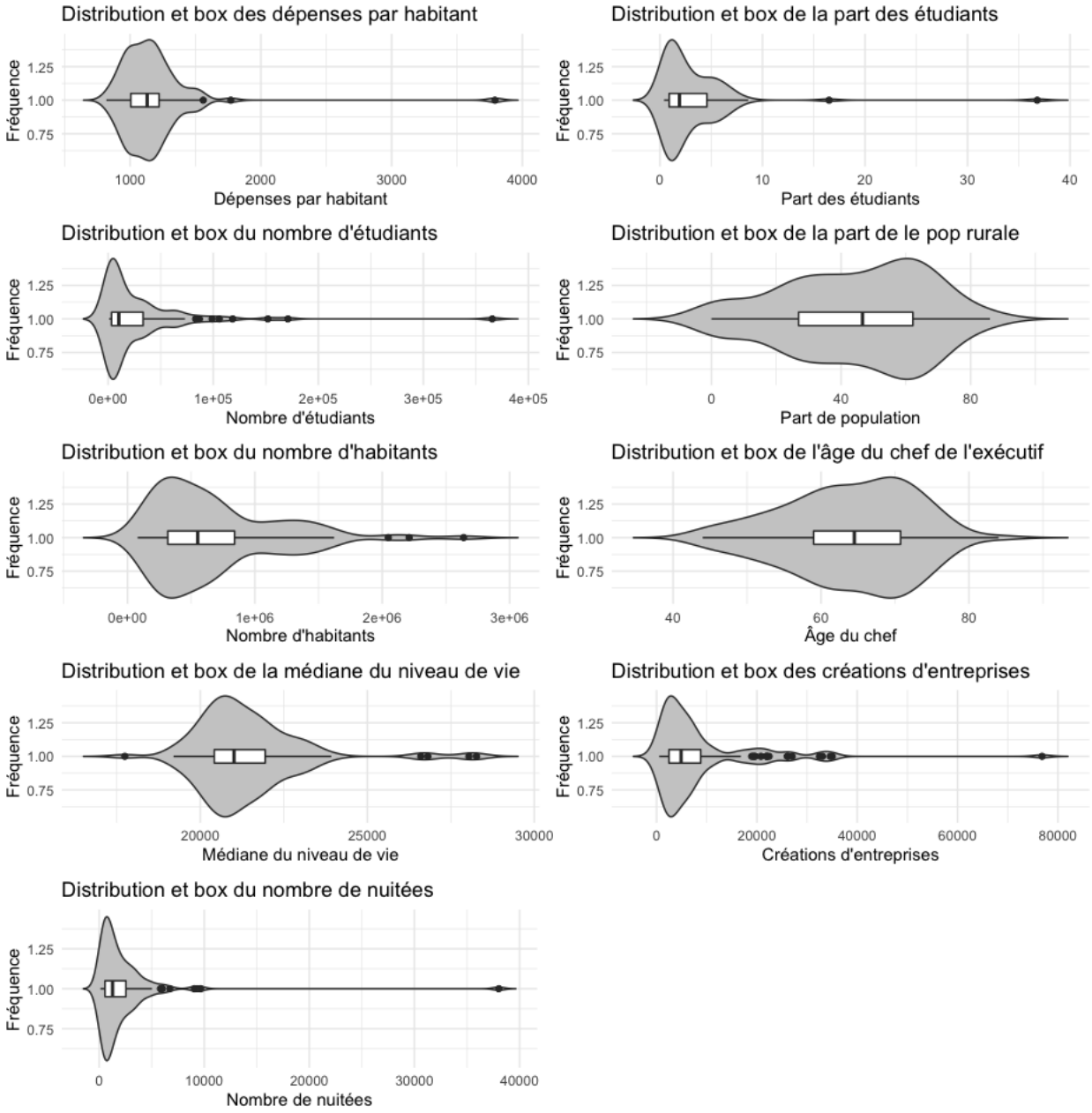
niveau_densite : Niveau de densité

1	Très dense
2	Dense
3	Peu dense

percent_pop_rurale : Pourcentage de population vivant en zone rurale

...	coefficient numérique

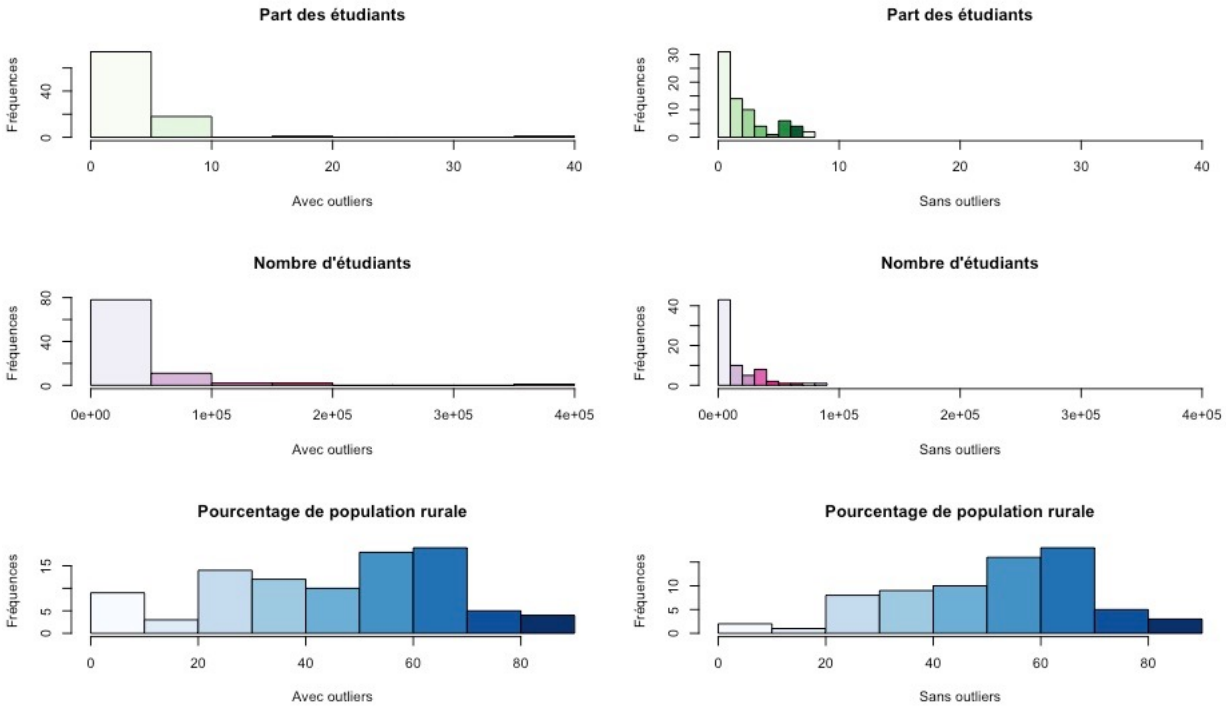
Annexe n°4 : Violins plots des 9 autres variables quantitatives.

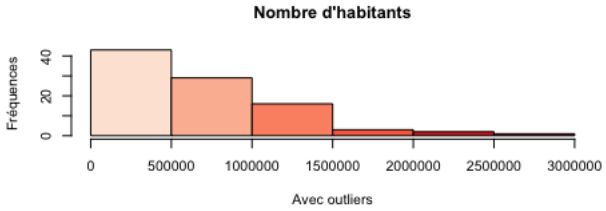


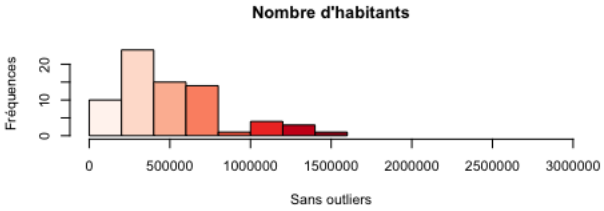
Annexe n°5 : Résultats des tests statistiques pour la détection d'outliers.

	Outliers potentiels	**Test**	**Outliers effectifs**
Nombre de publications	10	Rosner	10
Taux de chômage annuel moyen	1	Grubbs	1
Nombre de création d'entreprises	10	Rosner	5
Nb de nuitées ds hotels de tourisme	6	Rosner	4
Médiane du niveau de vie	5	Rosner	4
Dépenses totales par habitant	3	Rosner	2
Nombre d'habitants	3	Rosner	1
Part des diplômés d'un BAC+5 ou +	6	Rosner	6
Nombre d'étudiants (volume)	8	Rosner	6
Part d'étudiants (%)	2	Rosner	2

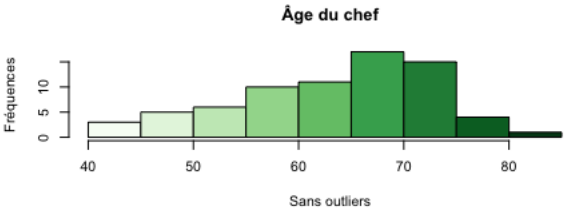
Annexe n°6 : Histogrammes de distributions des 8 autres variables quantitatives.

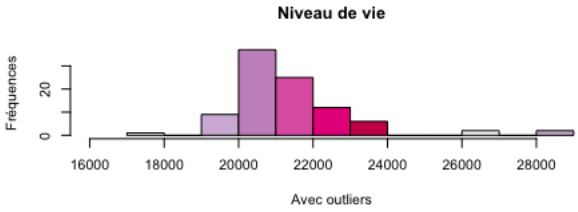


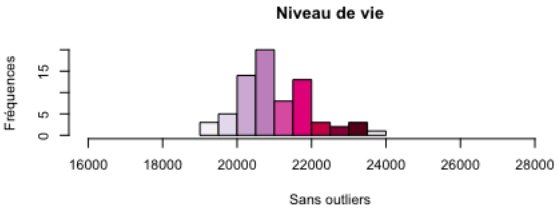

Nombre d'habitants
Fréquences
0
20
40
0
500000
1000000
1500000
2000000
2500000
3000000
Avec outliers

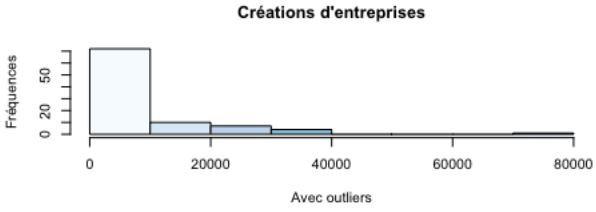

Nombre d'habitants
Fréquences
0
10
20
0
500000
1000000
1500000
2000000
2500000
3000000
Sans outliers

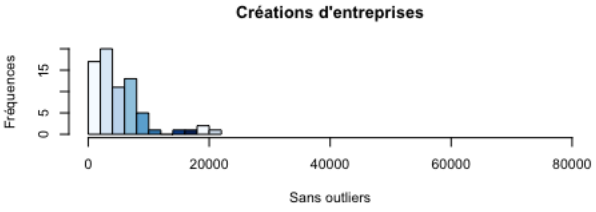

Âge du chef
Fréquences
0
5
15
40
50
60
70
80
Avec outliers

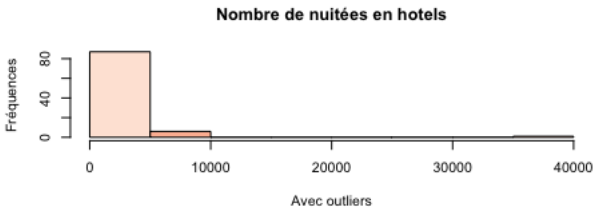

Âge du chef
Fréquences
0
5
10
40
50
60
70
80
Sans outliers

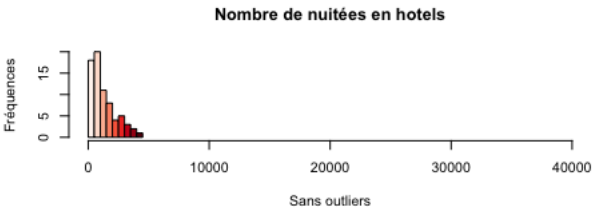

Niveau de vie
Fréquences
0
20
16000
18000
20000
22000
24000
26000
28000
Avec outliers


Niveau de vie
Fréquences
0
5
15
16000
18000
20000
22000
24000
26000
28000
Sans outliers

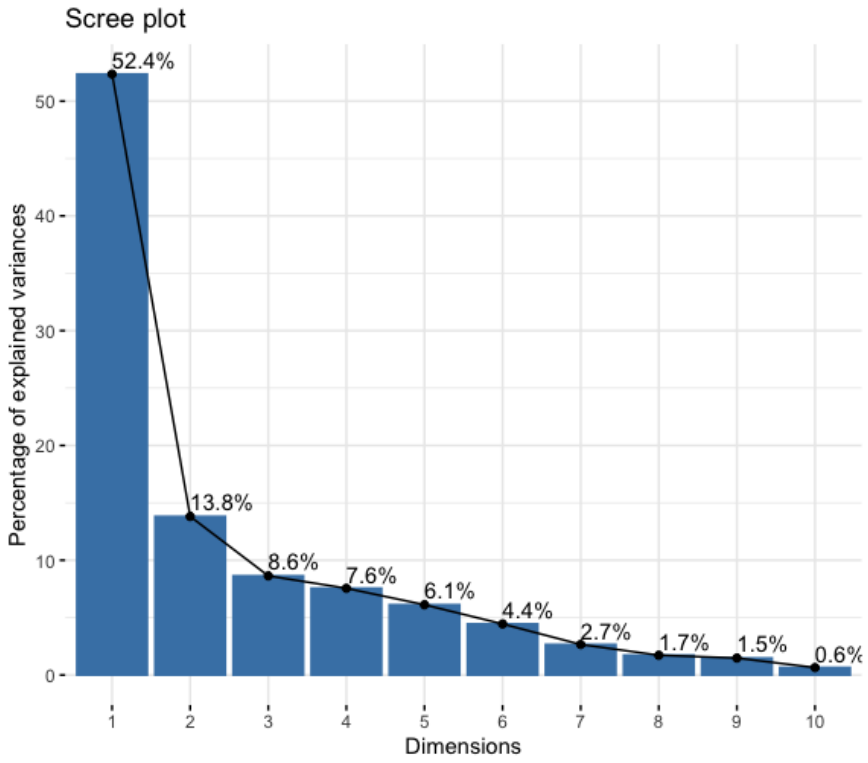

Créations d'entreprises
Fréquences
0
20
50
0
20000
40000
60000
80000
Avec outliers


Créations d'entreprises
Fréquences
0
5
15
0
20000
40000
60000
80000
Sans outliers

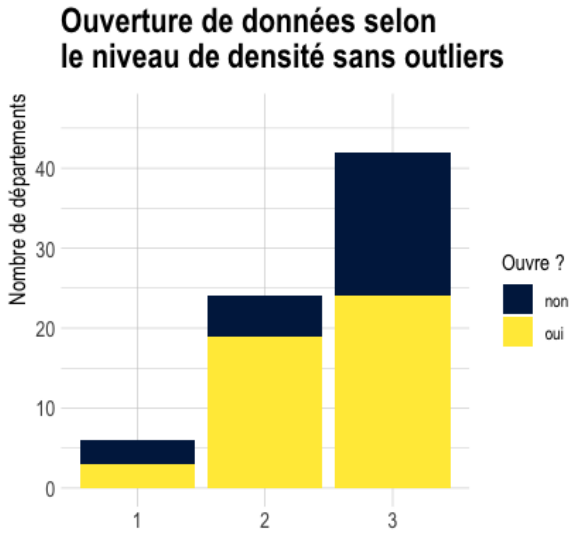

Nombre de nuitées en hotels
Fréquences
0
40
80
0
10000
20000
30000
40000
Avec outliers


Nombre de nuitées en hotels
Fréquences
0
5
15
0
10000
20000
30000
40000
Sans outliers

Annexe n°7 : Inertie associée aux axes de l'ACP sur les données sans outliers.



Annexe n°8 : Répartition des départements par niveau de densité, sans et avec outliers.



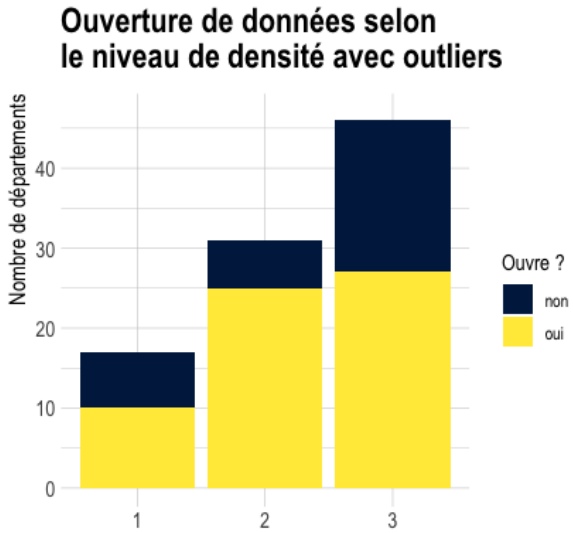


Table des figures

Table des tableaux

Table des matières