Quels facteurs influencent l'ouverture de données des territoires ?

Diane THIERRY

Chaque seconde dans le monde, 1.7Mo de données sont créées. Selon une étude menée par l'International Data Corporation (IDC) en 2018, le volume de données atteindra 175 Zo en 2025, ce qui, stocké sur des disques Blu-ray, permettrait d'atteindre 23 fois la lune. Alors que les données se massifient, leur accessibilité et leur ouverture deviennent un enjeu majeur pour bon nombre de gouvernements. En France depuis 2016, la loi pour une **République Numérique** oblige les administrations et collectivités locales de plus de 3.500 habitants ou de plus de 50 agents à ouvrir leurs données. L'**open data** qui par cette loi devient la règle et non plus l'exception, désigne l'ensemble des données accessibles et ouvertes à tous, qui peuvent être utilisées et partagées librement. De nombreux bienfaits découlent de cette ouverture de données, que ce soit pour les pouvoirs publics, pour l'économie ou pour les citoyens eux-mêmes qui peuvent réutiliser les données et accéder à l'information. Cependant, seul 1 territoire sur 10 qui est concerné par cette loi publie effectivement des données. Nous nous demandons alors : quels facteurs influencent l'ouverture de données des territoires ? C'est à cette problématique connue des organismes accompagnant l'ouverture de données, que nous avons tenté de répondre dans cette analyse. Pour cela, nous avons construit une base composée de données diverses et variées ayant pour but de caractériser les territoires au niveau économique, politique, géographique et démographique. Au total, ce sont 22 variables explicatives récoltées, provenant de 7 sources différentes et portant sur les 4 niveaux territoriaux suivants ; régions, départements, communes et EPCI. Le principal apport de cette étude se trouve dans la recherche qui se veut la plus exhaustive possible des déterminants de l'open data des territoires.

• Méthodologie économétrique

La nature même de la variable à expliquer définit la méthodologie économétrique à utiliser. En effet, la variable que nous cherchons à expliquer ici est de nature quantitative continue ; il s'agit du **nombre de publications open data** qui étant un dénombrement, est nécessairement positif et entier. Les estimations permettant de modéliser ce type de

données sont les régressions linéaires généralisées, il s'agit dans notre cas des '**modèles de comptage'**. Ceux-ci sont adaptés à notre étude, contrairement aux modélisations classiques de type MCO qui supposent la normalité de la variable à expliquer (Y).

Le modèle de comptage le plus basique est celui où l'on suppose que Y suit la loi de *Poisson*, qui implique une variance et une moyenne égales. Or, il arrive parfois que la variance soit supérieure à la moyenne. Une des causes de ce phénomène appelé '**surdispersion**' est la surreprésentation des valeurs zéro dans l'échantillon. Nous savons que 9 collectivités sur 10 ne publient aucun jeu de données, il est donc nécessaire de prendre cela en compte en utilisant d'autres lois de distributions telles que la loi de *quasi poisson* ou la loi *binomiale négative*. Aussi, pour prendre en compte cette inflation en zéro nous sommes en mesure d'appliquer des modèles qui décomposent Y en 2 parties : une relative à la probabilité d'observer l'événement et une autre relative au comptage. Cette dernière, en incluant ou excluant Y=0, considère le non-évènement comme volontaire (modèle « double hurdle ») ou involontaire (modèle « à inflation de zéros »). S'agissant dans notre cas d'une politique consciente menée par les territoires, nous pensons qu'elle sera plus adaptée et l'appliquerons à nos données.

• ***Résumé méthodologique***

→ **MCO** non adaptés.

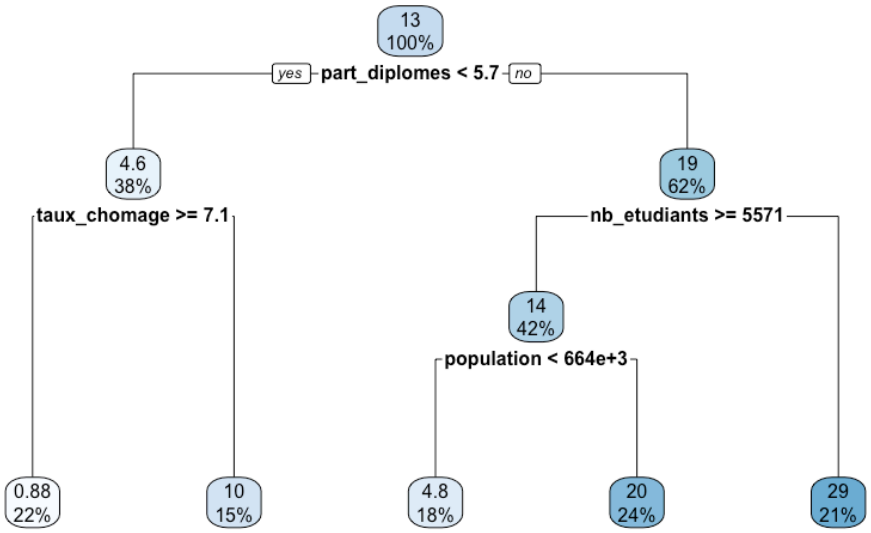
→ **GLM** sur les données → loi de *Poisson*

Sur-dispersion	**Sur-représentation des 0**
loi de *quasi poisson*	modèle à inflation de zéros
loi *binomiale négative*	double hurdle (Y=0 volontaire)

Bien que nous ayons récupéré des données sur les régions, les départements, les communes et les EPCI, celles-ci se trouvent incomplètes pour de nombreuses organisations. C'est pourquoi nous décidons de centrer l'analyse empirique uniquement sur les 93 départements de France métropolitaine.

• Résultats obtenus

La partie d'application pour répondre à la problématique énoncée s'est déroulée en 2 temps ; d'abord nous avons exploré les données visuellement et statistiquement pour identifier les territoires ouvrant le plus de données, puis nous avons validé ou invalidé ces hypothèses à partir de modèles suivant la méthodologie énoncée ci-dessus. Les différentes analyses visuelles ont montré l'importance de la variable de la part des diplômés d'un BAC+5 ou plus dans la population, comme nous le voyons sur l'arbre de régression ci-joint. La sous-population publiant le plus de données correspond aux départements ayant plus de 5.65% de diplômés et moins de 5571 étudiants ; c'est 21% des départements (parmi l'échantillon corrigé des valeurs atypiques) et leur nombre de publications moyen est de 29. À l'opposé, les départements ne publiant pas (Y=0) ou très peu (Y=1), sont ceux qui ont moins de 5.65% de diplômés et plus de 7.1% de chômage. Les autres analyses graphiques ont révélé que l'open data est favorisé par une bonne santé économique (faible taux de chômage, créations d'entreprises, population dense, niveau de vie élevé…), et par un contexte politique favorable (ouverture de données plus grande pour les départements de gauche). Cependant, cette analyse exploratoire a permis aussi de révéler que des départements avec des caractéristiques socio-économiques similaires ont des maturités open data très différentes. Elle a ainsi montré que les facteurs considérés dans cette analyse ne permettent pas d'expliquer pleinement le phénomène d'ouverture de données.



Dans un second temps, l'application de modèles économétriques a permis de confirmer la plupart de ces hypothèses. Après avoir régressé par la loi de *poisson*, nous nous sommes rendu compte que la variance était largement supérieure à la moyenne. De plus, il se trouve que 36% des départements de la base débarrassée des valeurs extrêmes ne pratiquent pas l'open data. Nous avons pris en compte cette surreprésentation des zéros

par les modèles listés ci-dessus. Le plus adapté à nos données est le *double hurle* qui décompose Y en une partie de comptage et une partie binaire, pour expliquer à la fois la quantité de données ouverte et le fait d'ouvrir ou non, en considérant que ce dernier dépend de la volonté des départements. La variable de la part des diplômés supérieure à 5.65% - construite à partir de l'arbre de régression, permet d'expliquer les 2 parties du modèle ainsi estimé. 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 en ayant moins. Les relations entre les variables explicatives et le nombre de publications open data se sont pour la plupart vérifiées grâce aux modélisations, néanmoins il reste quelques incohérences. Par exemple, le nombre de publications semble a priori plus important dans les départements à densité faible par rapport à un département à densité forte, mais cela est dû à la sous-représentativité de la modalité de cette variable. Les autres relations sont quant à elles conformes à ce que nous attendions d'un point de vue théorique.

• Conclusion

Ainsi l'analyse a-t-elle montré quelques facteurs de l'ouverture de données des départements, à savoir la part des diplômés dans la population, la couleur politique et le niveau de densité. L'incohérence de certaines relations et l'inconformité de certains départements ayant pourtant les mêmes caractéristiques socio-économiques peuvent s'expliquer de différentes manières. Premièrement, les données desquelles nous tenons le nombre de publications open data sont parfois imparfaites et peuvent être biaisées par un comptage non représentatif de la maturité réelle des départements. Deuxièmement, les 17 facteurs explicatifs récoltés au niveau départemental ne suffisent pas à expliquer Y et nous voyons qu'il dépend de déterminants non pris en compte dans cette étude. Quoi qu'il en soit, cette étude aura permis de mettre en avant certains facteurs de l'ouverture de données ; montrant principalement l'importance du dynamisme économique. Dès lors, il pourrait être intéressant de compléter cette analyse par l'ajout de nouveaux facteurs explicatifs, ou encore en écartant les publications comptabilisées sur *datagouv*, qui peuvent être faussées par des données republiées ou découpées en sous-jeux.