AGENIS-NEVERS Marc
Publication : 10/05/2017
Version : 1.0

Couverture médiatique internationale de deux quotidiens français, évolution et modélisation

Contexte

Deux grands quotidiens français sont étudiés au travers de leur couverture de l'actualité internationale. Une technique de *web scraping* permet de mesurer chaque année le nombre d'articles publiés mentionnant chacun des 200 pays du monde. Cette note analyse l'évolution de cette mesure au fil des années, et teste plusieurs variables explicatives.

Méthodologie

Choix des titres de presse

Le choix de la source de données pour la couverture médiatique s'est porté sur les titres « Libération » et « Le Figaro » deux grands quotidiens nationaux figurant parmi les 10 principaux (ACPM 2016) et dont la ligne éditoriale se situe respectivement à gauche et à droite. Ces deux journaux sont présents en ligne depuis plus de 10 ans et permettent aux internautes d'effectuer des recherches historiques dans leurs bases de données d'articles. Toutefois, dans le cas du Figaro, le nombre de résultats avant 2009 est très faible, pour une raison inconnue (articles peu archivés, ou rareté des publications sur le site à l'époque ?).

Web scraping

La méthode utilisée pour collecter les données s'inspire du *web scraping* (voir Encadré 1). Cette pratique étant controversée voire illégale dans certains cas, on précise ici qu'aucun contenu n'est récupéré mais uniquement une information sur la quantité de contenu disponible (résultat d'une recherche).

Un script R, installé sur une machine virtuelle Google, récupère le contenu d'une requête de recherche d'articles sur chacun des journaux (exemple : http://www.liberation.fr/recherche/). Une triple boucle « for » requête sur chaque mois, année, et pays individuellement. La page HTML est enregistrée au format texte afin d'isoler la chaîne de caractère mentionnant le nombre de résultats :

```
> scrap[grep("facettes__nombre", scrap)]
[1] "        <div class=\"facettes__nombre\">298 rA©sultats</div>"
```

Le web scraping (parfois appelé harvesting) est une technique d'extraction du contenu de sites Web, via un script ou un programme, dans le but de le transformer pour permettre son utilisation dans un autre contexte, par exemple le référencement. Source (Wikipédia 2017)

Encadré 1: Définition du Web scraping

Autres sources de données

Plusieurs sources de données sont mises à contribution afin de faciliter l'analyse et de fournir des variables explicatives à un modèle :

- Données de population des pays (Wikipédia, Liste des pays par population 2017) : ces données issues de Wikipédia proviennent pour majeure partie de la CIA (document *World Factbook*) ou de sources officielles nationales.
- Données aériennes de flux de passagers depuis la France, toutes destinations: (EUROSTAT 2017)
- Appartenance des pays à la francophonie (Wikipédia, Liste des pays ayant le français pour langue officielle 2017): contribution collective, Wikipédia.
- Appartenance des pays à l'Union Européenne : (Europa.eu s.d.)
- Distances entre capitales de chaque pays données par l'outil GeoDist du CEPII: (Mayer 2011)
- Codes ISO des noms de pays (2 et 3 lettres) pour fusionner les différents jeux de données : (Wikipédia, ISO 3166-1 2013)

Ces diverses sources de données sont fusionnées dans un tableau complet, la plus grosse difficulté étant la concordance entre la liste de pays de chacune des sources, certains litiges internationaux ou interprétations pouvant engendrer des divergences. Certaines analyses excluent donc des pays qui ne figurent pas dans certaines bases ? Par ailleurs les données de passagers aériens sont manquantes pour 109 des 199 pays en 2015 (elles n'ont probablement pas été remontées pour ces pays).

Nettoyage de données

Nous excluons certains pays des tableaux : la France (puisque c'est le centre de l'analyse mais également parce que les résultats de Libération bloquent à 999 et le terme « France » dépasse ce plafond), mais également la Dominique (dont les données d'articles sont perturbées par ce nom très courant). Par ailleurs le web scraping pour Libération donne des résultats aberrants pour le Mali et la Géorgie (il semble qu'il y ait un bug sur leur site avec la lettre « i » à la fin des termes de recherche).

Du fait de la méthode utilisée, certains articles sont comptés en doublon ou plus, ainsi peut-on imaginer qu'un article sur les jeux olympiques citera de nombreux pays et sera compté de multiples fois. Il est toutefois justifiable de pondérer doublement un article qui parle de deux pays distincts, on préfèrera donc parler de nombre de « citations ».

Statistiques descriptives

Evolution du nombre d'articles

L'évolution du nombre de citations de pays étrangers est présentée dans les graphiques ci-dessous, par année en Figure 1.

On notera que le nombre de citations ou d'articles totaux est largement supérieur dans le cas du Figaro (140 000 au maximum) que de Libération (27 000 au maximum). Ce dernier semble en baisse sensible sur les 4 dernières années (peut-être en raison de la baisse globale des tirages de Libération) alors que sur la même période les citations du Figaro sont stables. La courbe vert clair qui se détache

clairement est celle des Etats-Unis, avec deux pics en 2008 et 2012, qui pourraient correspondre aux années des élections d'Obama.

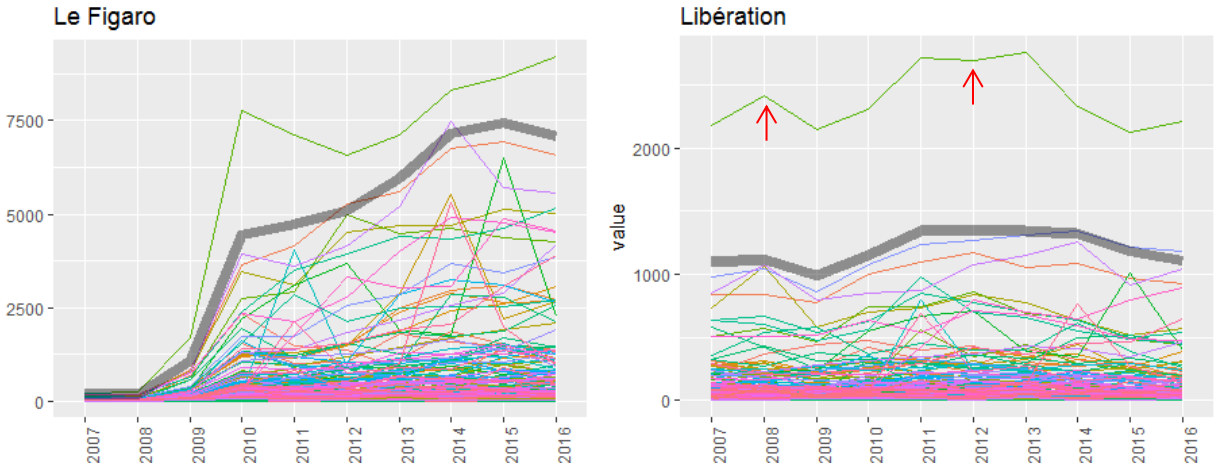


Figure 1: Nombre de citations par an par pays. La ligne noire est le total, échelle séparée.

Classement des pays les plus représentés

On représente de façon similaire le total des citations enregistrées sur 10 ans, par pays, en sélectionnant les 15 premiers Figure 2:

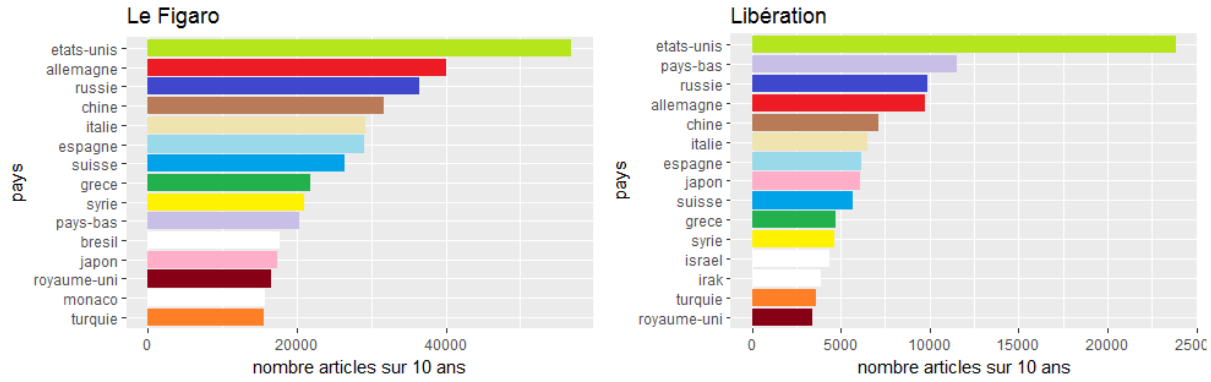


Figure 2: Importance comparée des pays sur 10 ans

On peut noter le rang étonamment élevé des pays-bas chez Libération (2ème rang devant la Russie), et surtout l'importance prépondérante des Etats-Unis (qui est jusqu'à deux fois plus citée que le pays suivant chez Libération). Le classement du « top 15 » est très donc proche entre les deux journaux, à deux exceptions près (Monaco+Brésil pour le Figaro, Irak+Israël pour Libération) ; cette proximité est tout à fait surprenante alors que le positionnement politique est opposé. Une explication pourrait être la part importante d'information « reprise » comme les dépêches AFP, qui serait commune aux deux quotidiens. Les sujets traités semblent donc proches, même si la façon de les traiter pourra différer.

Inégalité de représentation

Afin de quantifier l'inégalité de représentativité des pays d'une façon générale, c'est-à-dire le fait que certains pays représentent une part très grande du total, et d'autres une part infime, on dispose

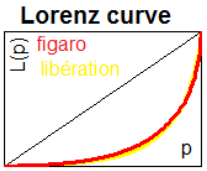
d'une mesure connue en économétrie qui est l'indice de Gini. Il se définit comme suit, pour n pays (ici $n = 199$) et leur nombre de citations y_i :

$$G = \frac{2 * \sum_{i=1}^{n} i \cdot y_i}{n * \sum_{i=1}^{n} y_i} - \frac{n+1}{n}$$

On calcule G pour chaque année (Tableau 1), et il apparaît comme quasi constant dans le temps et homogène entre les deux quotidiens avec une inégalité légèrement plus élevée pour le Figaro (il existe des procédures de test de significativité de différence (Karoly and Schroder 2014) mais elles n'ont pas été implémentées ici). La valeur de 0.7 est interprétable comme une inégalité plutôt sévère. Lorsqu'elle atteint 1, l'inégalité est totale (un seul individu accumule 100% de la valeur).

Tableau 1: Evolution annuelle des indices de Gini, et courbe de Lorenz

G	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Libération	.68	.70	.68	.70	.72	.71	.68	.69	.69	.70
Le Figaro	.73	.72	.68	.69	.71	.72	.70	.72	.71	.72



Pour l'année 2016, nous donnons quelques indicateurs statistiques sur cette répartition (Tableau 2), qui montre une nouvelle fois une asymétrie forte avec une moyenne égale au 3ème quartile.

Tableau 2: Résumés statistiques des citations 2016

2016	*N*	*Min*	*25%*	*50%*	*Moy*	*75%*	*max*	*CV*
Libération	199	0	13	36	112	116	2200	2.06
Le Figaro	199	0	62	189	710	732	9210	1.82

Modélisation (Le Figaro, 2015)

Objectif

On souhaiterait étudier les inégalités de distribution du nombre de citations observées, tant en nombre que géographiquement, au moyen des variables explicatives collectées ainsi que de modèles de régression (linéaire ou généralisée). Il ne s'agit pas de prédiction, car les variables explicatives représentent des choses assez différentes, et il n'y a pas nécessairement de relation causale avec le Y; il s'agit donc plutôt d'expliquer et de commenter les liaisons entre variables. On cherchera enfin, après avoir retranché l'effet des variables les plus pertinentes (population, traffic, distance) à analyser les écarts résiduels qui pourraient indiquer des biais ou des explications sous-jacentes.

Mise en œuvre

Un premier modèle est testé sur les données 2015 du Figaro, entre le nombre de citations et la population des pays, dont on peut penser qu'elle est un bon prédicteur en ce qu'un pays peuplé a a priori plus de relations et donc d'information à partager avec les autres. Or ce modèle ne respecte absolument pas les hypothèses (normalité, hétéroscédasticité) du fait de la distribution des deux variables ; la répartition de la population montre une distribution encore plus « inégalitaire » que les citations avec un $G = 0{,}81$. Dans un premier temps une transformation des deux variables de type Box-Cox est envisagée, avec un paramètre optimisé $lambda = 0{,}1$; ce paramètre étant proche de zéro, on se trouve quasiment dans le cas de données à distribution *log-normale* et il est plus direct de prendre le

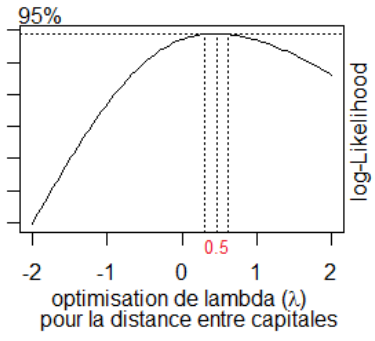


Figure 3

logarithme afin d'établir un modèle dit *log-log*. On fait le même constat sur la variable « passagers » mais pas sur la variable distance sur laquelle une transformation racine carrée ($lambda = 0{,}5$, voir Figure 3) plus pertinente.

Analyse exploratoire graphique

La représentation une à une des variables est présentée en différenciant selon le continent d'appartenance de chaque pays (avec leurs ellipses de confiance à 50%).

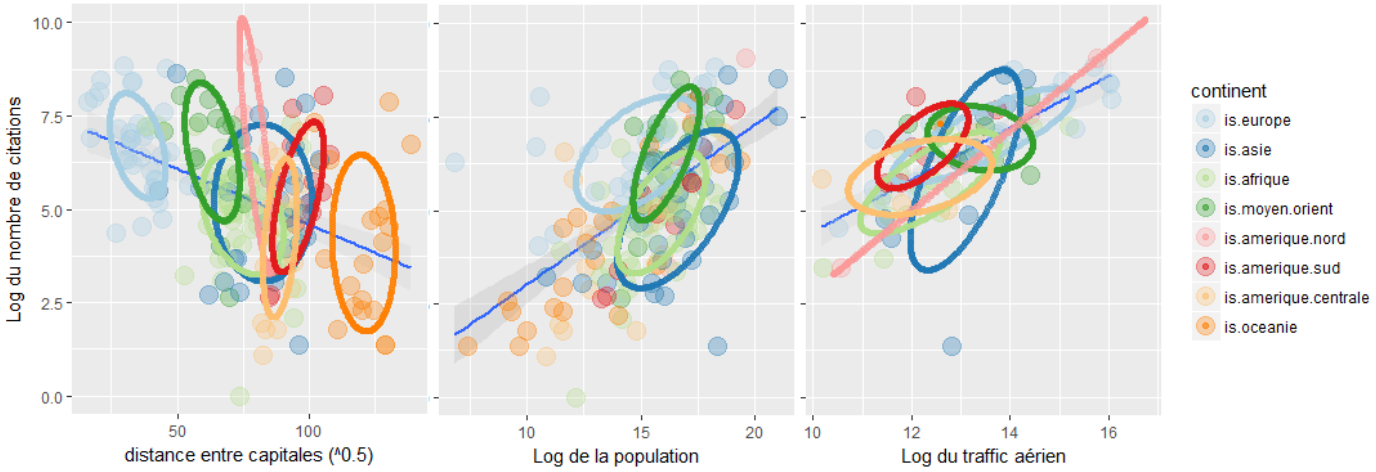


Figure 4: citations contre distance, population ou trafic aérien, couleur par continent et linéaire (bleu) (Figaro, 2015)

Le graphique sur la distance montre bien les « clusters » par continent, avec l'Europe en plus proche et l'Océanie lointaine. La relation semble correctement linéaire et à part pour le Moyen-Orient il n'y a pas de biais selon le continent (les barycentres des ellipses sont proches de la droite). On rencontre quelques outliers comme le Viet Nam (très peu cité malgré une forte population), Monaco (beaucoup cité, peut-être pour les domaines sportifs/touristique) et le Vatican (beaucoup cité du fait de sa nature religieuse). Ce n'est pas tout à fait le cas pour la population. Enfin, la même représentation pour le trafic aérien montre un meilleur ajustement que pour les deux autres variables, et moins de points isolés.

Corrélation	*Population (log)*	*Distance (^0.5)*	*Trafic aérien (Log)*
Coefficient Pearson*	0.34	-0.35	0.69

* calculé sur 90 pays seulement

Analyse du modèle complet

Un modèle complet est calibré sur les deux variables explicatives Population et Distance, nous mettons de côté le traffic aérien puisqu'il conduirait à exclure la moitié des observations. On obtient une forte significativité pour chacune des deux variables (Tableau 3). On valide ensuite les hypothèses de distribution des résidus (Figure 5); il faudra prendre quelques précautions à l'interprétation des coefficients qui est plus délicate pour un modèle *log-log.*

Tableau 3: analyse du modèle citations~population+distance (Figaro, 2015)

	Coefficient	*Std.Error*	*P-value*	R^2
Intercept	0.26	0.78	0.72	
Distance (^0.5)	-0.023	0.0036	10^{-9}	0.49
Population (Log)	0.44	0.043	0	

Le modèle donne un intercept non-significatif, ce qui est logique (un pays de population nulle n'est pas cité). Ainsi, un pays 2 fois plus peuplé (+100%) verra son nombre de citations augmenter de presque 44% pour un même éloignement. La magnitude des coefficients est donnée en recalibrant le modèle sur les variables standardisées, l'effet de la population est 60% plus fort que celui de la distance.

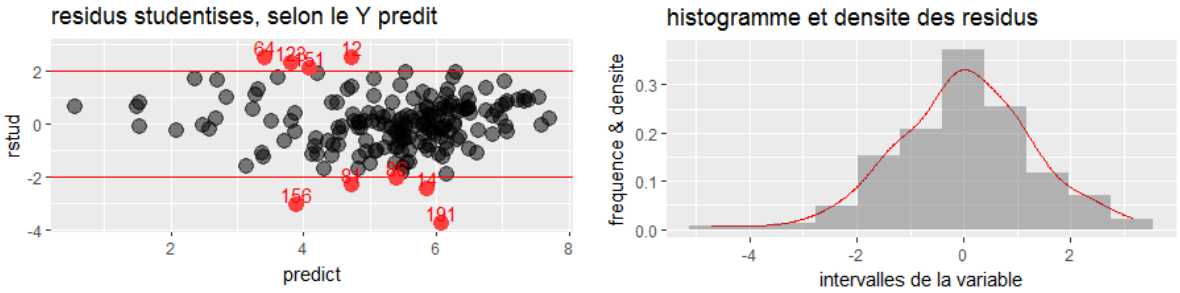


Figure 5: analyse des résidus du modèle citations~population+distance (Figaro, 2015)

Autres variables testées dans le modèle

A partir du modèle initial (citations~population+distance) on teste tour à tour l'ajout d'autres variables : traffic aérien, appartenance à l'UE, francophonie.

- Le traffic aérien qui avait été exclu est testé dans un second modèle avec un effet significatif fort également, et affaiblit légèrement les coefficients des deux premières variables, en valeur absolue ; c'est logique puisque le traffic aérien s'explique en partie par la distance et la population des pays.
- Les pays de l'Union Européenne sont tous biaisés positivement sur le modèle initial (les pays membres sont plus cités que ce que prédit le modèle). L'ajout de l'appartenance à l'UE au modèle, sous la forme d'une variable catégorielle (facteur) booléenne, apparait clairement significative. Toutefois, elle perd sa significativité lorsqu'elle est introduite conjointement au nombre de passagers, qui apporte donc une information redondante. On explique cela par le flux important et facilité de personnes à l'intérieur de l'Union Européenne (en particulier l'espace Schengen).
- L'appartenance d'un pays à la francophonie (en langue officielle ou non), testée dans le modèle. Le coefficient pour ce facteur est toujours négatif et non significatif, ce qui signifie que la francophonie n'a soit aucun impact sur la couverture médiatique, soit un impact négatif, ce qui est en somme assez contre-intuitif.

Biais résiduels par pays ou zone

A partir du modèle principal additionné de la variable UE (citations~population+distance+is.ue), on étudie les écarts de prédiction (résidus) graphiquement, par pays ou par zone. Certains pays, bien entendu, reflètent par leur abondance de citations une actualité forte, c'est le cas de la Syrie depuis quelques années, de la Grèce en 2015 par exemple, le Panama en 2016. Quelques pays sont systématiquement sur-représentés au fil des ans comme la Russie, l'Australie, la Suisse, l'Argentine, le Qatar, l'Islande pour les plus marquants.

Certaines zones sont systématiquement biaisées positivement du fait d'un fort impact touristique (les îles d'Océanie en particulier, avec aussi le Rugby !). Il aurait été intéressant de pouvoir cibler uniquement les thématiques politico-économiques par exemple, afin de s'affranchir de ce biais.

Certaines zones sont systématiquement biaisées négativement (quelle que soit l'année) comme l'Eurasie (pays d'Asie Centrale comme l'Azerbaidjan, le Kirghizistan, etc) et apparaissent complètement « oubliées » de la presse. C'est le cas également de nombreux pays d'Afrique Centrale. En Asie, le Viet-Nam semble extremement peu cité malgré une histoire francophone importante.

Répétition des analyses sur les données Libération

La reprise des mêmes analyses sur le journal Libération avec des résultats de modélisation identiques. Des différences apparaissent sur les pays sur- ou sous-représentés au cas par cas. Ainsi, citons une sur-représentation récurrente pour : USA, Australie, Israël, Islande, Cuba, Japon et Nouvelle-Zélande. A l'inverse, cinq petits pays ne sont même jamais cités.

On donne un exemple de représentation des résidus du modèle à trois variables sur le graphique suivant, où les noms de pays apparaissent proportionnellement à leur sur- ou sous-représentation Figure 6:

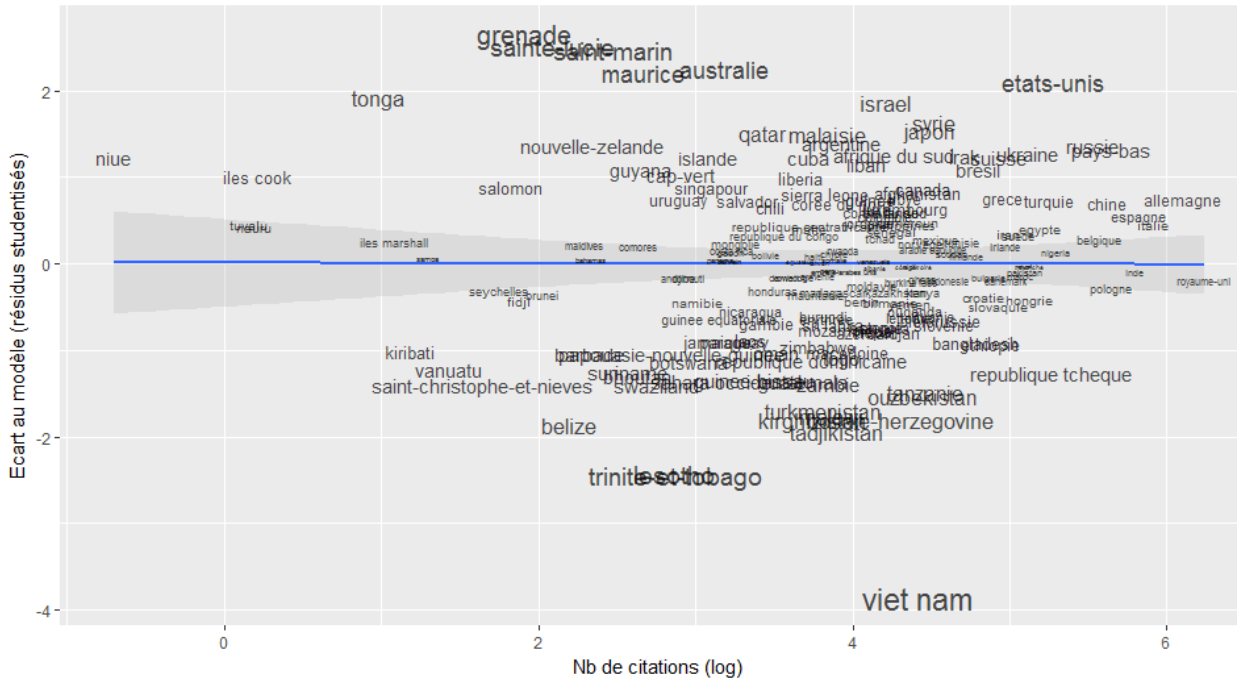


Figure 6: représentation des écarts au modèle, positifs (surreprésentation) ou négatifs (sous-représentation). Données : Libération, 2014

Conclusion

Nous avons tenté d'analyser puis d'expliquer l'évolution de la couverture médiatique des pays du monde chez deux grands quotidiens français, au moyen de variables indépendantes assez facilement accessibles. Pour certaines d'entre elles, on ne peut pas vraiment parler de relation de cause à effet : les mouvements de personnes (trafic aérien) et la médiatisation de telle ou telle destination peuvent s'influencer réciproquement. On aimerait avoir accès à d'autres variables intéressantes comme des données issues des douanes (volumes d'échanges économiques) ou de Facebook (nombre de relations sociales de pays à pays), mais difficiles à récupérer.

Nous avons constaté une très forte **inégalité** de représentation des pays, avec **16 pays accumulant la moitié des citations** totales dans des articles. Cette inégalité s'explique en partie (R^2=50%) par l'éloignement, la population ou les échanges de personnes.

Certains facteurs comme la langue ou une histoire liée à la France (les pays **francophones** étant presque tous des anciennes colonies) ne semblent pas jouer voire au contraire défavorisent la médiatisation.

Globalement, d'un journal à l'autre la structure de la **couverture médiatique est extrêmement proche**, avec quelques différences au cas par cas. L'omniprésence médiatique des **Etats-Unis** est très claire ; d'autres *outliers* (pays surreprésentés) demanderaient une analyse plus poussée afin de trouver un point commun entre eux ou d'autres facteurs explicatifs (échanges économiques, relations humaines). On peut aussi tout simplement imaginer que ces sujets trouvent un écho plus fort auprès du lectorat et sont apréciés des rédactions.

Annexes

Le script R utilisé pour l'analyse est disponible sur la page Gitlab.

Bibliographie

ACPM. «Classement presse quotidienne nationale 2016.» *Alliance pour les chiffres de la presse et des médias.* 31 12 2016. http://www.acpm.fr/Chiffres/Diffusion/La-Presse-Payante/Presse-Quotidienne-Nationale (accès le 05 05, 2017).

Europa.eu. «A propos de l'UE - Pays.» *Europa.eu.* s.d. https://europa.eu/european-union/about-eu/countries_fr (accès le 05 05, 2017).

EUROSTAT. «Database - Eurostat.» *Eurostat.* 05 01 2017. http://ec.europa.eu/eurostat/cache/metadata/en/avia_pa_esms.htm (accès le 05 05, 2017).

Karoly, Lynn, and Carsten Schroder. "Fast Methods for Jackknifing Inequality Indices." *RAND Corporation*, 2014.

Mayer, Thierry. «CEPII - outil GeoDist.» *CEPII.* 2011. http://www.cepii.fr/distance/geo_cepii.xls (accès le 05 05, 2017).

OTAN. «NATO - pays membres.» *NATO.* 18 12 2013. http://www.nato.int/cps/fr/natohq/nato_countries.htm (accès le 05 05, 2017).

Wikipédia. «ISO 3166-1.» *Wikipédia France.* 2013. https://fr.wikipedia.org/wiki/ISO_3166-1 (accès le 05 05, 2017).

—. «Liste des pays ayant le français pour langue officielle.» *Wikipédia France.* 08 04 2017. https://fr.wikipedia.org/wiki/Liste_des_pays_ayant_le_fran%C3%A7ais_pour_langue_officielle (accès le 05 05, 2017).

—. «Liste des pays par population.» *Wikipédia France.* 02 05 2017. https://fr.wikipedia.org/wiki/Liste_des_pays_par_population (accès le 05 05, 2017).

—. «Web scraping.» *Wikipédia France.* 17 01 2017. https://fr.wikipedia.org/wiki/Web_scraping (accès le 05 05, 2017).