Analyses de crimes par cantons

Jason Ola

April 2021

1 Introduction

Nous vivons dans une époque où fake-news, discriminations et stigmatisations sont répandus et relayés par les médias. Cela mélangé au fait que la Suisse accueille nombre d'étrangers ou demandeurs d'asiles alimente des frictions vis-à-vis de cette population et les politiques s'en mêlent, notamment l'UDC.

On voit souvent dans les médias des affaires de crimes où l'origine du malfaiteur est mentionnée, en particulier s'il est étranger. Aussi, il y a une théorie populaire qui avance la cause de criminalité comme étant la pauvreté [2]. Je voudrais donc voir ce qu'il se passe dans les analyses avec des variables comme le nombre de bénéficiaires d'aide sociale ou le nombre de boursiers, étudier les éventuels liens et corrélations, en gardant en tête que la corrélation ne signifie pas forcément la causalité.

2 Données

Nos données sont constituées de n = 26 cantons ainsi que p = 9 variables collectées sur le site de l'OFS [1]

Parmi ces variables nous avons 2 variables catégorielles et 7 variables numériques :

– population (variable numérique de taille) : C'est le nombre d'habitants par canton avec pour somme 8.5 mio d'hab pour la Suisse.

Statistiques descriptives

	N	Somme
population	26	8544527
N valide (liste)	26	

FIG. 1 – *Population et nombre de cantons*

C'est également la variable que j'ai décidé d'utiliser pour pondérer en créant la variable : poidseff = 26*population/8544527. Je pondère donc avec poidseff.

1. https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/donnees.html

– tendance_parti (variable catégorielle) : C'est la tendance gauche / centre / droite de la politique du canton obtenus en regardant le parti avec le plus de candidats élus depuis 2011. S'il y a égalité entre 2 cantons opposés j'ai mis centre.

  – C : Centre
  – D : Droite
  – G : Gauche

tendance_parti

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	C	2	9,6	9,6	9,6
	D	17	65,6	65,6	75,1
	G	6	24,9	24,9	100,0
	Total	26	100,0	100,0	

FIG. 2 – *Tableau de fréquences politique*

– frontiere_int (variable catégorielle) : Si le canton possède une frontière internationale, j'ai simplement regardé sur une carte de la Suisse avec les cantons.

  – O : Oui
  – N : Non

  J'ai mis non pour Berne, bien qu'il y ait un tout petit bout de frontière avec la France. Pour les variables numériques, j'ai récolté les variables de

frontiere_int

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	N	12	44,5	44,5	44,5
	O	14	55,5	55,5	100,0
	Total	26	100,0	100,0	

FIG. 3 – *Tableau de fréquences frontières*

  taille de chaque indicateur et ensuite divisé par la taille de la population afin d'avoir des variables de densité.
– enseignants (variable numérique de densité) : la proportion d'enseignants dans le canton.
– benef_aide_soc (variable numérique de densité) : la proportion de bénéficiaire d'aide sociale dans le canton.
– boursiers (variable numérique de densité) : la proportion de boursiers dans le canton.

– etrangers (variable numérique de densité) : la proportion d'étrangers dans le canton.
– requerants (variable numérique de densité) : la proportion de requérants dans le canton.
– crimes (variable numérique de densité) : proportion de crimes dans le canton. J'ai choisi quelques crimes qui paraissaient pertinents dans le contexte de l'exercice, ainsi je n'ai pas choisi certains crimes comme la fraude fiscale mais plutôt des crimes violents ou « urbains » : (meurtre, lésions corporelles graves, participation agression, vol, brigandage, escroquerie, extorsion/chantage, menaces, viol, discrimination raciale)

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
enseignants	26	.007764446	.011580243	.009319556	.001012152
benef_aide_soc	26	.006616848	.052807464	.023934034	.010582377
boursiers	26	.002601764	.010088529	.005534654	.002498931
etrangers	26	.112976154	.400296308	.251421173	.064541185
requerants	26	9.05240E-5	1.61201E-3	3.91947E-4	2.45734E-4
crimes	26	.001424590	.008281733	.003930937	.001492711
N valide (liste)	26				

FIG. 4 – *Statistiques descriptives des variables numériques*

3 Résultats et discussions

3.1 Test du χ^2

Posons l'hypothèse nulle $H0$: les variables frontiere_int et tendance_parti sont indépendantes avec un seuil de signification $\alpha = 0.05$, que nous gardons tout au long de cette analyse.

Voici les résultats du test χ^2 entre la tendance politique et s'il y a une frontière internationale. Nous voyons la valeur du $\chi^2 = 2.4$ ainsi que la va-

Tableau croisé frontiere_int * tendance_parti

			tendance_parti			
			C	D	G	Total
frontiere_int	N	Effectif	1	9	1	11
		Compte attendu	,9	7,5	2,6	11,0
	O	Effectif	1	8	5	14
		Compte attendu	1,1	9,5	3,4	14,0
Total		Effectif	2	17	6	25
		Compte attendu	2,0	17,0	6,0	25,0

FIG. 5 – *Tableau croisé*

Tests du khi-carré

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2,400[a]	2	,301
Rapport de vraisemblance	2,609	2	,271
N d'observations valides	25		

a. 4 cellules (66,7%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de ,88.

FIG. 6 – *Test du* χ^2

leur $p = 0.301$ (Fig. 6) ainsi, $p > \alpha$, nous ne pouvons pas rejeter l'hypothese d'indépendance des 2 variables. Cela est donc une bonne chose comme il est préférable d'avoir les variables les plus indépendantes possibles. Cependant on peut se permettre de douter de la pertinence du test au vu du nombre de cellules aux valeurs inferieures à 5 : 66.7% (Figure 6).

3.2 Régression linéaire

Pour commencer, j'ai recodé les variables nominales en variables numériques car le modèle ne prend pas les caractères. J'ai recodé de la manière suivante :

- frontiere_int : O = 1, N = 0
- tendance_parti : G = 0, C = 1, D = 2

J'ai également mis comme type de variable numérique limité pour les traiter comme des chiffres entiers non regroupables qui considère aussi le 0 comme non significatif.

Pour cette régression multiple j'ai choisi d'utiliser la variable dépendante : crimes. J'ai choisi comme variable indépendante :

- tendance_parti : éventuel lien politique lié aux crimes, peut-être il peut complémenter les chiffres d'aide sociale
- frontiere_int : élément qui pourrait complémenter la proportion d'étrangers afin de ne pas trop biaiser un éventuel lien crime/étrangers
- enseignants : la proportion d'enseignants peut être une indication de la qualité d'éducation publique, l'éducation à peut-être un rôle dans la criminalité
- benef_aide_soc : comme on veut regarder une composante socio-économique dans le lien avec le crime par rapport à la théorie populaire énoncée dans l'introduction
- boursiers : aussi une composante socio-économique qui est également en lien avec l'éducation pourrait s'avérer utile
- etrangers : je suis curieux de voir le poids de cette variable dans la régression
- requerants : pour complémenter la variable étrangers comme frontiere_int.

Posons comme hypothèse H0 : il n'y a pas de relation linéaire entre les variables indépendantes X et la variable dépendante Y.

Regardons maintenant les sorties SPSS : Commençons par évaluer la qualité

Récapitulatif des modèles[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,950[a]	,903	,865	.000547916

a. Prédicteurs : (Constante), requerants, etrangers, frontiere_int, tendance_parti, benef_aide_soc, enseignants, boursiers

b. Variable dépendante : crimes

FIG. 7 – *Evaluation du modèle*

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	,000	7	,000	23,936	,000[b]
	de Student	,000	18	,000		
	Total	,000	25			

a. Variable dépendante : crimes

b. Prédicteurs : (Constante), requerants, etrangers, frontiere_int, tendance_parti, benef_aide_soc, enseignants, boursiers

FIG. 8 – *ANOVA*

de la régression, nous avons dans le tableau récapitulatif (Fig. 7) la valeur du R^2 qui est la proportion de variance expliquée. Ici cette valeur vaut 90,3%, notre modèle est particulièrement robuste. Nous voyons sur le tableau d'ANOVA (Fig. 8) la valeur F qui est le rapport de la somme des carrés moyens inter et intra-groupe. Nous voyons aussi que la valeur p de l'ANOVA est très faible et inférieure à α, cette valeur est la probabilité de retrouver la valeur F lorsque l'hypothèse nulle est vraie, nous pouvons alors rejeter l'hypothèse $H0$. Dans le tableau des coefficients (Fig. 9), nous voyons aussi la valeur p associée aux coefficients de régression des différentes variables. Nous pouvons alors analyser l'amplitude de ces coefficients et leur signification. Ici, la régression donne plus d'importance à la variable requerants : 1.221 ainsi qu'aux variables enseignants et boursiers avec respectivement -0.259 et 0.292. La valeur p de requerants et de boursiers sont inférieures au seuil de significativité $\alpha = 0.05$ et sont donc pertinentes, par contre, celle d'enseignants ne l'est pas. Nous voyons également dans le tableau que les variables tendance parti et frontière internationales ne sont pas pris en compte dans la régression avec des coefficients nuls. De ce tableau on pourrait déjà imaginer ne garder que boursiers et requérants, nous verrons plus tard ce qu'il en est dans l'analyse en composantes principales. Le signe des coefficients nous indiquent le sens de la corrélation, ici nous pouvons interpréter le signe négatif d'enseignants : moins il y a d'enseignants, plus il y a de crimes. Pour le signe positif de requérants on peut dire : plus il y a de requérants,

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	-,001	,002		-,277	,785
	tendance_parti	,000	,000	,156	1,284	,215
	frontiere_int	-,001	,000	-,332	-3,115	,006
	enseignants	-,259	,183	-,176	-1,413	,175
	benef_aide_soc	,033	,016	,233	2,096	,050
	boursiers	,292	,075	,489	3,873	,001
	etrangers	,016	,002	,711	6,700	,000
	requerants	1,221	,506	,201	2,411	,027

a. Variable dépendante : crimes

FIG. 9 – *Coefficients de régression linéaire*

plus il y a de crimes. Attention encore tout de fois à ne pas faire de lien de causalité, ces coefficients sont calculés sur la base des variables à disposition et il pourrait en avoir d'autres qui contre-balanceraient ces coefficients différemment.

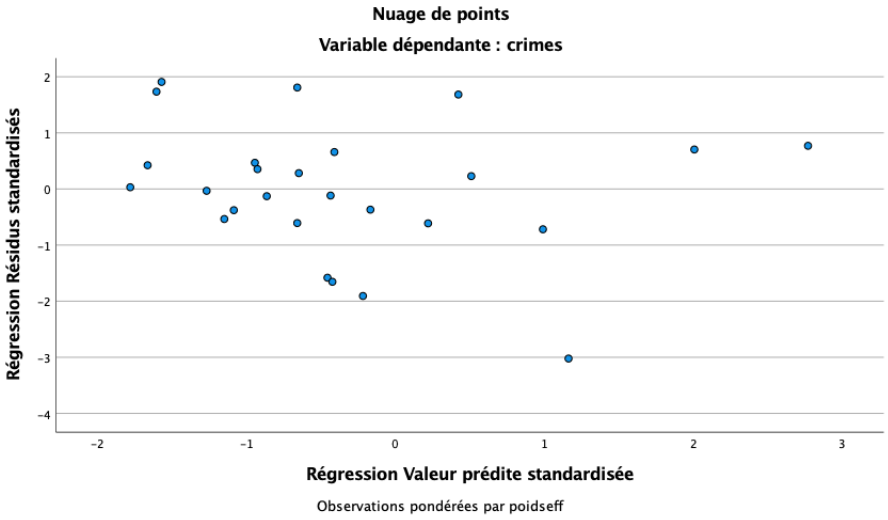

FIG. 10 – *Diagramme des résidus*

Le diagramme de résidus (Fig. 10) peut également servir afin de vérifier les hypothèses de travail : normalité, homoscédasticité et linéarité ([1]). C'est un nuage de points avec pour abscisse les valeurs standardisées des $y*$ prédits et en ordonnées les résidus e. Ici nous voyons que le nuage se centre autour de 0 sur les 2 axes et se répartissent pas trop autour, juste 2 valeurs proche de 3

fois l'écart-type. Le diagramme nous montre que les hypothèses de travail sont vérifiées.

3.3 Analyse en composantes principales

Procédons maintenant à une analyse en composantes principales, essayons d'extraire les composantes les plus importantes de nos variables. Comme l'analyse en composantes principales se base sur les corrélations entre variables, je choisis les 5 variables numériques indépendantes que nous avions utilisées : enseignants, benef_aide_soc, boursiers, etrangers et requerants. Il convient de com-

Indice KMO et test de Bartlett

Indice de Kaiser-Meyer-Olkin pour la mesure de la qualité d'échantillonnage.		,730
Test de sphéricité de Bartlett	Khi-carré approx.	38,384
	ddl	10
	Signification	,000

FIG. 11 – *Test de Bartlett*

mencer par regarder la pertinence d'une analyse en composantes principales, posons comme hypothèse nulle H0 : les variables sont globalement indépendantes. Le test de sphéricité de Bartlett (Fig. 11) donne une valeur $p = 0$, on peut donc rejeter l'hypothèse nulle.

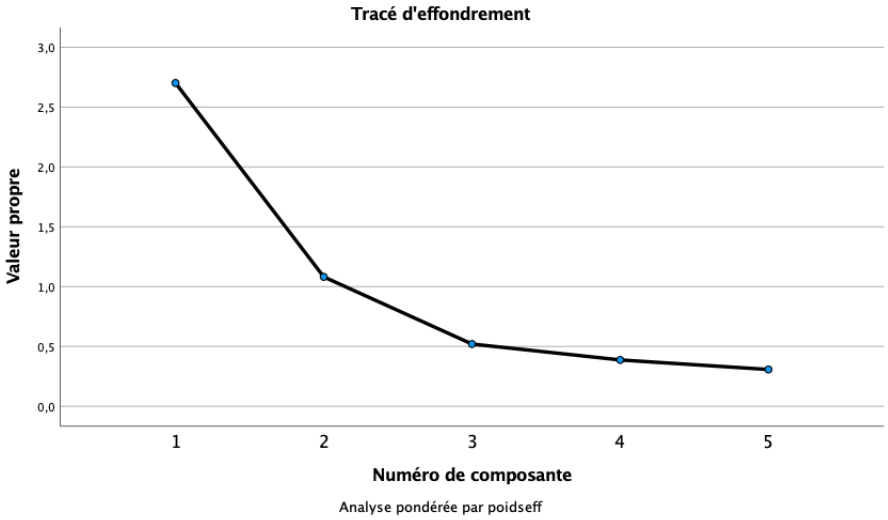

FIG. 12 – *Screegraph*

Si dessus nous avons le screegraph (Fig. 12), on peut y voir les valeurs propres associées aux composantes et le pourcentage cumulé. Il y a également les détails dans le tableau de variance totale expliquée (Fig. 13). Si l'on se réfère au critère

Variance totale expliquée

Composante	Valeurs propres initiales Total	% de la variance	% cumulé
1	2,702	54,041	54,041
2	1,082	21,638	75,679
3	,521	10,412	86,091
4	,388	7,753	93,844
5	,308	6,156	100,000

Méthode d'extraction : Analyse en composantes principales.

FIG. 13 – *Proportion de variance expliqueé en détails*

de Joliffe ici, nous devrions retenir 90% de variance expliquée cumulée, donc 4 composantes. Si l'on se réfère au critère du coude de Cattell, on devrait en garder 1, l'angle me parait plus important au numéro de composante 2, il faut donc l'exclure.

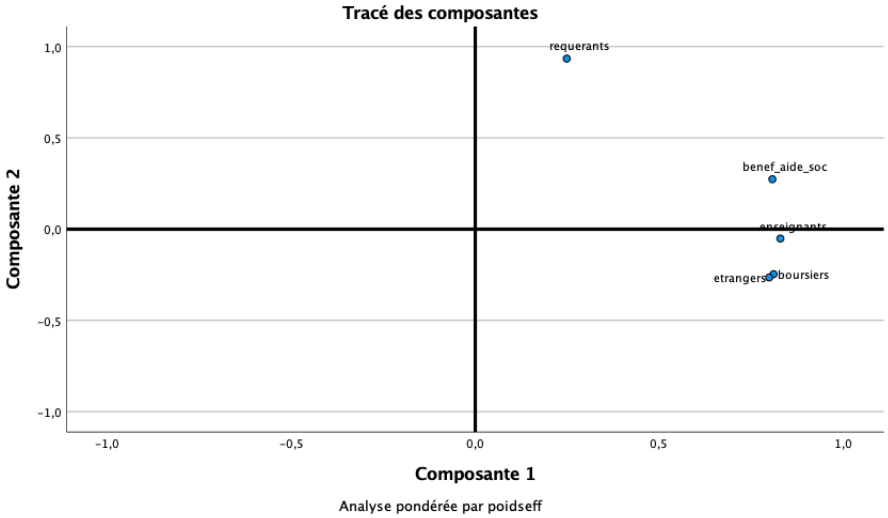

FIG. 14 – *Cercle des corrélations*

Voici ci-dessus le cercle des corrélations (Fig. 14), basé sur les corrélations entre variables et composantes. On peut voir 2 groupes : un proche de l'axe de composante 2 composé juste de requerants et l'autre proche de la composante 1 composé de benef_aide_soc, enseignant, etrangers et boursiers. Il est difficile de nommer les composantes ici mais on peut essayer, le groupe de la composante 1 s'apparente plutôt à un groupe socio-économique. Le groupe de la composante 2 est fortement représenté par requerants et un peu par benef_aide_soc, mais avec une légère contradiction avec etrangers et boursiers, appelons cette composante : type de population.

J'ai aussi enregistré les scores factoriels afin de construire un diagramme

de scores factoriels (Fig. 15) avec les cantons en libellé que vous trouverez ci dessous :

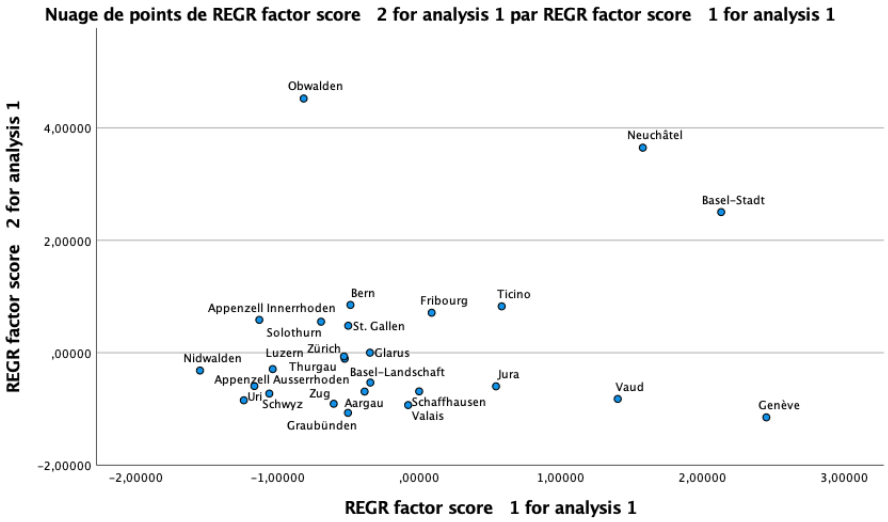

FIG. 15 – *Diagramme de scores factoriels*

Pour rappel, nous avons nommé le score 1 : socio-économique, et score 2 : type de population. On peut voir un centre autour de 0 sur les axes et quelques cantons qui se démarquent, notamment Obwald, Neuchâtel, Bâle, Vaud et Genève. J'ai remarqué sur ce graphe que l'axe score 1 était particulièrement lié à la tendance politique avec sur la droite des cantons à tendance politique gauche et sur la gauche des cantons à tendance de droite, probablement lié aux tendances sociales de la gauche pour les cantons de gauche par exemple. Pour l'axe du score 2, c'est assez lié à la proportion de requérants dans le canton, où l'on voit Obwald, Neuchâtel et Bâle ville comme fer de lance.

3.4 Classification hiérarchique ascendante

La classification hiérarchique ascendante permet de regrouper des individus afin de les classifier. J'ai pris comme variables : enseignants, benef_aide_soc, boursiers, etrangers et requerants, frontiere_int et tendance_parti. J'ai choisi de faire la classification avec la méthode de Ward en utilisant comme mesure d'intervalle le carré de la distance euclidienne et en enregistrant 3 valeurs de clusters dans le fichier de données. Voyons ce que donne le dendrogramme (Fig. 16).

Nous avons ici 3 groupes principaux, que l'on sépare aux distances 25 et 11, on revoit Vaud, Genève et Bâle Ville ensemble dans le 3ème groupe, encore un effet de la composante sociale politique. On voit aussi que le 2ème groupe contient que des cantons de Suisse centrale. J'ai par contre du mal à voir ce qui regroupe les cantons du premier groupe mis à part un sous-groupe montagneux

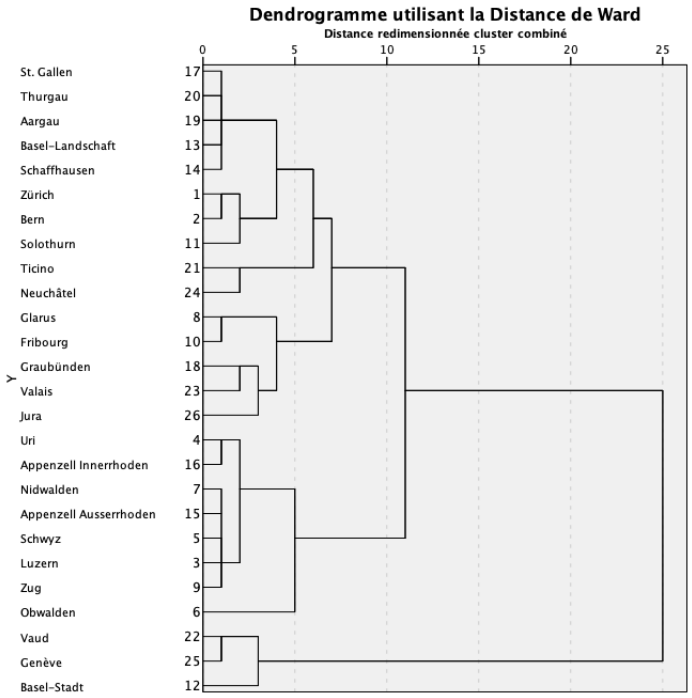

FIG. 16 – *Dendrogramme*

avec les Grisons, le Jura et le Valais.

4 Conclusion

On a pu voir lors de ce travail qu'avec les données que nous avons, quelques analyses ont pu être faites ainsi que quelques reflexions. Les résultats montrent à la fois dans l'ACP et dans la régression linéaire, que la variable requérants prend beaucoup de poids. Malgré le fait qu'il y ait peu de données, les tests sont significatifs pour ces 2 analyses. Cependant, on se rend compte que la grande influence de la variable requérants peut aussi être suspecte de la faible influence des autres et je remets en doute le choix de mes variables. Je pense également qu'avoir beaucoup plus de variables aurait rendu l'analyse plus interessante, en voyant le cercle des corrélations (Fig. 14) je vois beaucoup de vide sur la gauche et le bas du cercle et je pense que j'aurai pu mieux nommer les axes et ainsi imaginer plus de relations avec les crimes.

Références

[1] François Bavaud. *Méthodes quantitatives III : analyse des données multivariées*. 2018.

[2] Laurent Lemasson. *La pauvreté est-elle la cause de la délinquance?* 2017.