Introduction

« Au cours des 10 dernières années de 2008 à 2018, les vols ont considérablement augmenté. En 2008, c'est 386 000 vols et tentatives de vols contre la résidence principale qui ont été enregistré en France. Dix ans plus tard, en 2018, le nombre de vols a augmenté de 171 000. C'est une augmentation de plus de 44%, des cambriolages en France ! » (www.portes-etserrures.fr). Autre chiffre, d'après l'enquête CVS (Cadre de Vie et Sécurité), « En 2018, près de 221 000 personnes de quinze ans ou plus ont été enregistrées comme victimes de coups et blessures volontaires par la police et la gendarmerie en France métropolitaine, ce qui représente 4 victimes pour 1 000 habitants. ». Nombre de cambriolage et nombre de coups et blessures volontaires (tant intrafamiliale qu'extrafamiliale) sont en hausse sur le territoire. Ces deux facteurs font partie intégrante des délits liés à l'insécurité.

L'insécurité peut se définir comme un environnement physique ou social favorisant les atteintes aux personnes et aux biens (Larousse). Dans ce contexte, il est intéressant de comprendre les déterminants des actes d'insécurité en France car ces derniers peuvent impacter la qualité de vie urbaine, l'environnement physique ou bien social liés aux personnes et aux biens.

Précédemment nous nous sommes interrogés sur ces principaux facteurs contribuant aux faits d'insécurité sur le territoire métropolitain, plus précisément les actes de cambriolages et actes de coups et blessures volontaires pour l'année civile 2018.

Pour cela, nous avons collecté un ensemble de variables diverses et variées pour chacun des départements français métropolitain afin de déterminer le modèle le plus adéquat pour mettre en exergue la relation entre les variables sélectionnées et l'insécurité. (Tableau de données, Voir Annexe)

Matrice de corrélation et algorithme de sélection de modèle nous ont permis de choisir l'ensemble des variables formant le meilleur modèle. Dans un souci d'amélioration du modèle en vue d'obtenir une estimation future meilleur, nous avons appliqué plusieurs transformations logarithmiques sur nos variables dans l'équation du modèle. Le modèle final choisi était un modèle log-log (log(Délit)~ log(Population)+log(Immigration)+log(Scolarisation) disposant du coefficient de détermination ajusté (R2 ajusté) et du critère d'information d'Akaike le plus avantageux.

En conclusion, nous étions sensibles à l'idée que certaines variables non retenues pour le modèle pour cause de suspicion multicolinéarité pouvaient également apporter un éclaircissement à la problématique posée. De plus, de forte corrélation n'impliquer pas forcément une potentiel causalité.

C'est pourquoi l'objet du présent rapport vise à introduire un regard et plus approfondi sur la notion de multicolinéarité et d'endogénéité dans le but d'avoir une meilleure réponse quant à notre problématique : Quels sont les déterminants responsables des actes d'insécurité au travers des départements français ?

Les statistiques descriptives de nos variables et les points clés du modèle final constitueront le premier volet. Par la suite, nous appliquerons des méthodes de traitement de la multicolinéarité. Enfin nous évoquerons les possibilités d'endogénéité vis-à-vis de nos variables et notre modèle.

1. Rappel modèle final

a. Statistique descriptive

Visualisation de l'endogène

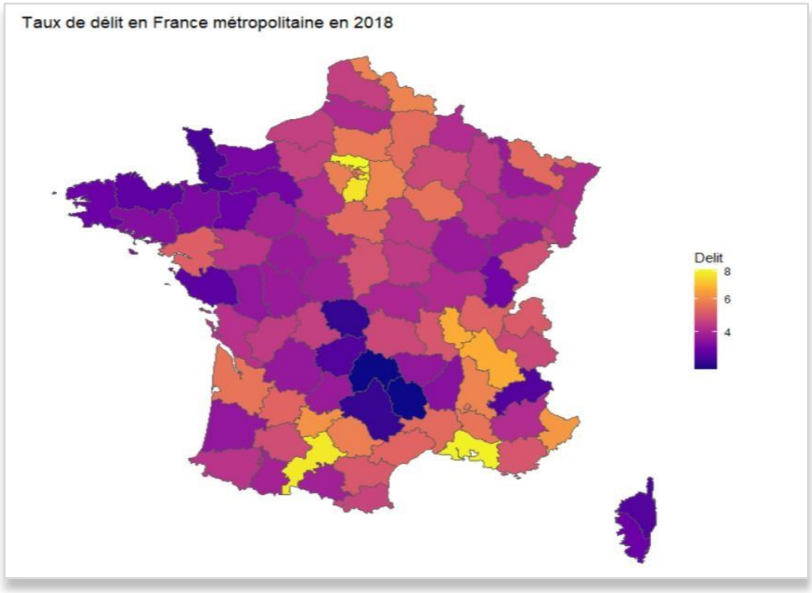


Figure n°1 – Carte de France en fonction du taux de délit

Les variables exogènes

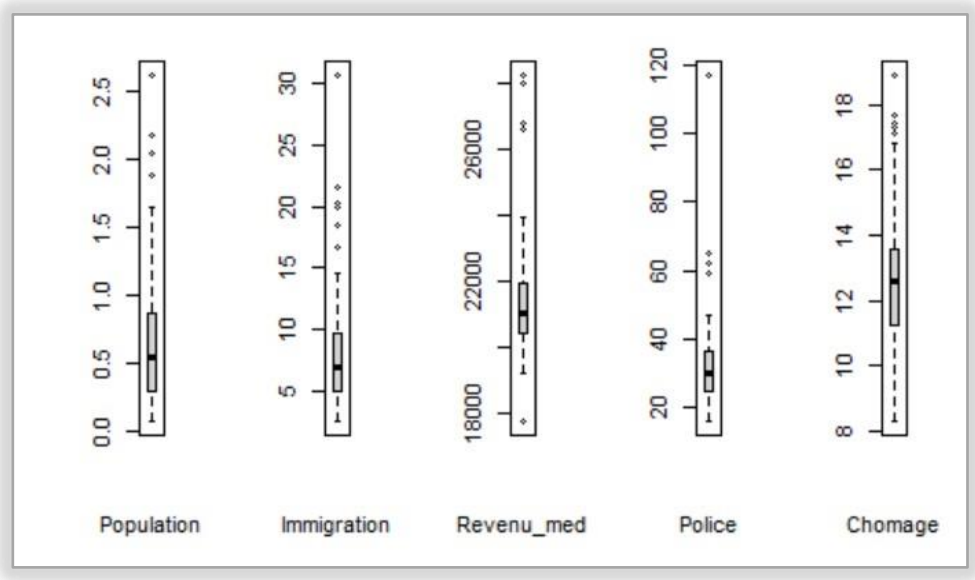


Figure n°2 – Distributions de variables exogènes (1)

Ces boîtes à moustaches mettent en évidence la pluralité d'individus « atypiques » aussi bien au-dessus du maximum déterminé par **min(max,Q3+1.5*(Q3-Q1))** que du minimum déterminé par **max(min,Q1-1.5*(Q3-Q1))**. Par conséquent, on observe une importante variance entre les individus faisant écho au contraste d'un département à un autre.

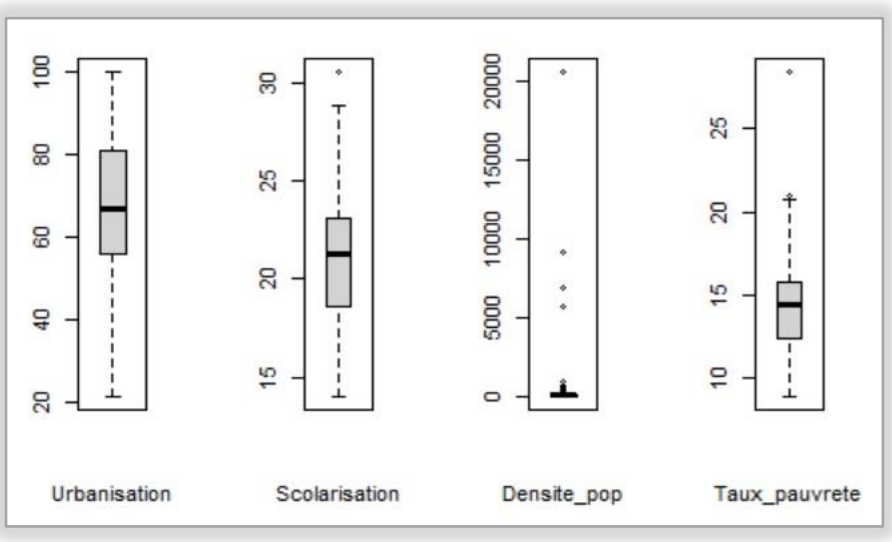
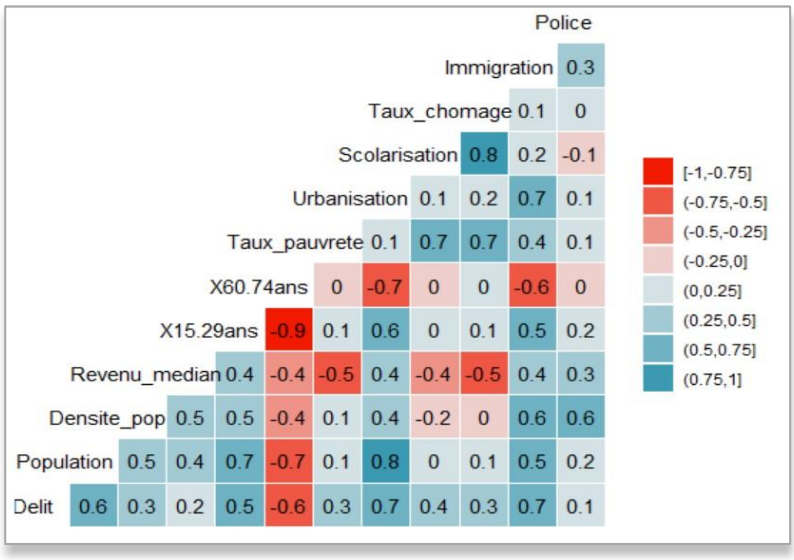


Figure n°3 – Distributions de variables exogènes (2)

La boîte à moustache de la densité de la population met en évidence la densité dominante du département de Paris et des départements de la première couronne parisienne. Environ 20 000 habitants/km² à Paris! Les trois autres boxplots ne comptent que peu, voire pas de valeurs atypiques.

b. Modèle final (log-log)



Modèle log-log (Estimation)

	Dependent variable:
	log(Delit)
log(Population)	0.169***
	(0.023)
log(Immigration)	0.257***
	(0.036)
log(Scolarisation)	0.545***
	(0.107)
Constant	-2.897***
	(0.439)
Observations	96
R2	0.714
Adjusted R2	0.705
Residual Std. Error	0.157 (df = 92)
F Statistic	76.500*** (df = 3; 92)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure n°4 – Corrélogramme

• La matrice de corrélation indique une seul corrélation négative avec l'endogène, c'est la variable X60.74ans, cela peut laisser penser qu'un département dont la part des 60-74ans est importante a une probabilité plus faible d'observer des cas de délit en 2018. On note que les variables Population, Immigration et Urbanisation ont une bonne corrélation positive avec Delit. Enfin, on remarque également d'importante corrélation entre exogène (ex : Urbanisation et Population).

• Le modèle log-log choisi initialement disposé d'un R2 ajusté de R2 avec des coefficients ayant un seuil de significativité très élevé, avec une probabilité d'erreur inférieure à 0.001.

2. Multicolinéarité

La multicolinéarité peu engendré une non fiabilité des coefficients de régression. Précédemment pour remédier à la multicolinéarité nous avions supprimé du modèle les variables fortement corrélés entre elles malgré parfois leurs bonnes corrélation avec l'endogène. Cette fois-ci nous transformerons les variables au travers de plusieurs méthodes.

Nous verrons tout d'abord les techniques de Ridge et Lasso basées sur la pénalisation des moindres carrés par pénalités de type L1 pour Lasso et L2 pour Ridge. Nous verrons également les techniques de PCR et PLS basées sur la réduction de dimensions des prédicteurs. Nous évaluerons ses modèles à l'aide du R2 (qualité d'ajustement du modèle) et du RMSEP (évalue les erreurs de prédictions)

Avant d'appliquer les méthodes nous porterons un regard sur le VIF des variables. Ce dernier est une mesure évaluant le degré de corrélation entre chaque variable explicative et les variables explicatives du modèle. Plus le VIF est élevé, plus la corrélation est forte. Un VIF égale à 1 indique une absence de corrélation, supérieur à 1 implique une corrélation positive, à partir de 5 on peut considérer la corrélation comme problématique. (Formule ci-dessous)

$$VIF(\hat{\beta}_k) = \frac{1}{1 - R_k^2}$$

VIF :

```
   Population   Densite_pop Revenu_median     X15.29ans     X60.74ans Taux_pauvrete  Urbanisation Scolarisation  Taux_chomage
Immigration        Police
     3.725985      3.621609      4.075740      7.258824      9.392143      6.129637      4.530457      4.054445      4.594260
5.150951      2.004747
    Politique
     1.261515
```

Un bon nombre de variables sont proches d'un vif de 5 voir supérieur à 5 notamment les variables Immigration, Taux_pauvrte, X15.29ans et X60.74ans. Les méthodes que nous verrons par la suite auront un réel intérêt au vue des VIF observés.

a. Méthodes de réductions de dimensions

Tout d'abord déterminons le nombre de composantes à conserver qui maximiseront le R2 et minimiseront le RMSEP.

PCR (Régression sur composantes principales) :

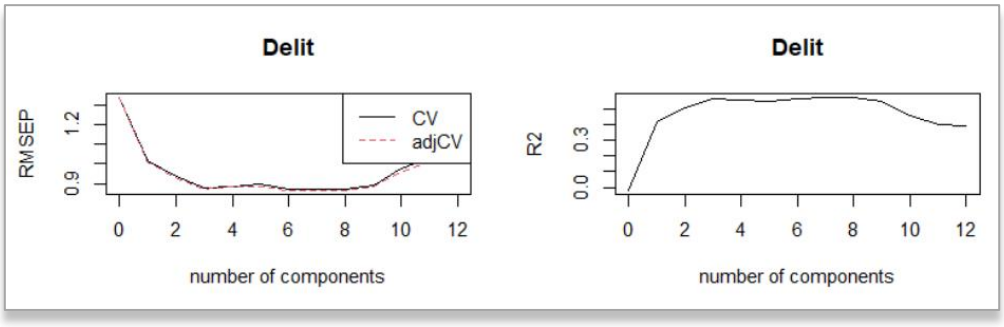


Figure n°5 – PCR Validation Plot RMSEP & R2

Au vue des deux graphiques et en s'appuyant sur le résumé statistiques via la commande « summary » nous choisirons un nombre de composantes égale à 7. Ci-après l'évaluation des modèles :

R2 = 65.69
Estimation du RMSEP = 0.8638

PLS (Régression des Moindres Carrés Partiels) :

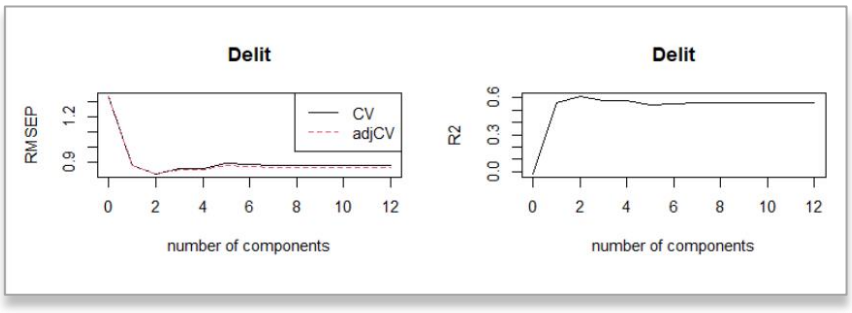


Figure n°6 – PLS Validation Plot RMSEP & R2

b. Méthodes de régularisation

Afin d'obtenir un résultat de régression le plus précis, choisirons la valeur de lambda la plus optimale qui n'est d'autre que le paramètre de pénalité. C'est au moyen de la « Validation croisée » que nous sélectionnerons le meilleur lambda. Les graphiques ci-dessous montreront comment les coefficients évoluent en fonction du paramètre de régularisation lambda.

Régularisation Ridge

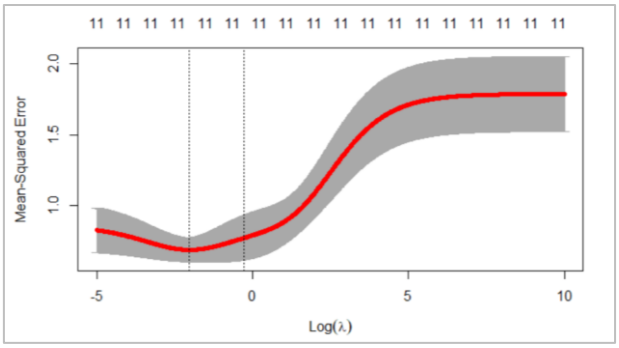


Figure n°7 – Ridge Cross Validation Plot

Après application de la méthode Ridge nous obtenons les résultats ci-contre :

R2 = 0.6997791
RMSEP = 0.8088936
MSE = 0.6543088

Régularisation Lasso

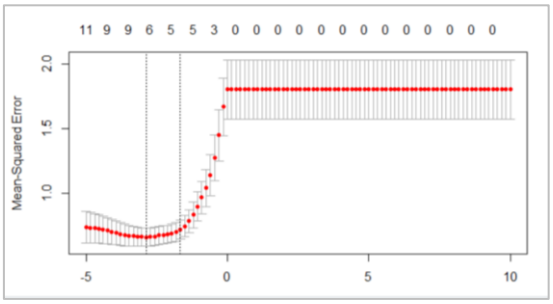


Figure n°8 – Lasso Cross Validation Plot

Après application de la méthode Lasso nous obtenons les résultats ci-contre :

R2 = 0.6928804
RMSEP = 0.8137986
MSE = 0.6622681

Elastic net:

Combinant à la fois Ridge et Lasso nous avons appliqué la méthode cv.glmnet avec un alpha équivalant à 0.5 afin de bénéficier des avantages de Ridge et Lasso. Controlant la régularisation du modèle, un alpha strictement différent de 0 ou de 1 impliquera que le modèle n'utilise pas uniquement la régularisation Ridge ou Lasso.

Nous obtenons un R2 de 0.6939 et un RMSEP de 0.8088.

Table des résultats par méthodes

METHODES	R2	RMSEP
PCR	65.09	0.8638
PLS	65.46	0.82
RIDGE	0.6997	0.8088
LASSO	0.6928	0.8137
ELASTIC NET	0.6939	0.8088

Nous avons essayé plusieurs méthodes pour déterminer celle qui est la plus appropriée pour notre jeu de donnée. D'après les résultats obtenus, les modèles de régularisation (Ridge, Lasso et Elastic Net) semblent être plus performants (sur notre jeu de donnée) que les méthodes de réductions de dimension en termes de R2 et de RMSEP. Lasso, Elastic net et Ridge affiche des résultats assez proche, pour notre jeu de donnée l'une de ces 3 méthodes semblent les plus adéquat avec un avantage pour la Ridge. Raison pour laquelle nous choisirons la méthode Ridge parmi les méthodes employées car elle présente le plus haut R2 parmi les modèles testés, avec un R2 de 0.6997 et un RMSEP de 0.8088.

L'approche différente entre méthode de régularisation et méthode de réduction de dimensions peut expliquer les résultats obtenus. Les méthodes de régularisation réduisent la multicolinéarité en pénalisant les coefficients de régression tandis que les méthodes de réduction de dimensions réduisent directement le nombre de variables grâce aux composantes principales qui capturent l'essentiel de l'information contenue dans les variables d'origine. Le fait de pouvoir conserver toutes les variables dans le modèle dans le processus de réduction de la multicolinéarité peut potentiellement être une explication quant au meilleur résultat.

Enfin, en comparaison avec le R2 de notre modèle initiale (log-log), on remarque qu'ils sont assez proche et que nous avons l'avantage avec la méthode Ridge de bénéficier de l'apport de chacune de nos variables explicatives, par conséquent un modèle avec une méthode Ridge pourrait être privilégié afin de mieux comprendre l'influence les facteurs influençant notre taux de délit.

3. Endogénéité

En économétrie il est coutume d'analyser les facteurs de manière « Ceteris Paribus ». Dans certains cas il peut y avoir de l'endogénéité dans le modèle, c'est-à-dire qu'une variable explicative est corrélée avec le terme d'erreur. L'endogénéité peut entrainer un biais dans les coefficients estimés et rendre les résultats de l'analyse peu fiable. C'est pourquoi nous analyserons ci-dessous les potentielles sources d'endogénéité.

• Erreur de mesure

La matrice de corrélation donne une idée des relations entre les variables, il ne semble pas y avoir de corrélation inattendues laissant indiquer une possible erreur de mesure. Les boxplots et le « summary » du jeu de donnée confirme l'absence d'erreur de mesure, il y a des valeurs atypiques dans le jeu de donnée liés à la diversité des départements français.

Résumé statistique de notre jeu de données, des visualisations des distributions statistiques et de la matrice de corrélation, ne montre pas de possible variables omises ou d'erreur de mesure

• Variables omises

Il est toujours possible que des variables potentiellement pertinentes ne soient pas incluses dans un modèle économétrique, car les relations économiques peuvent être difficiles à modéliser de manière exhaustive. Après réflexion, d'autres pistes de variables supplémentaires peuvent être intéressante pour le modèle tel que le taux de croissance du département ou encore les dépenses dans des politiques de préventions ou de sécurité. Cependant elles ne seraient pas considérées comme des variables importantes absentes du modèle.

• Simultanéité

Une approche de variables instrumentales, peut être utilisée pour vérifier la simultanéité entre les variables de l'ensemble de donnée. Pour implémenter cette approche sur Rstudio, la fonction nommé ivreg() peut être pratique.

Que peut-on dire du biais d'endogénéité ?

Afin de détecter l'endogénéité nous avons employé des tests de corrélations entre les résidus de la régression et chacune des variables explicatives. C'est à l'aide de la fonction « cor.test » du logiciel Rstudio. Exemple ci-dessous avec la variable Immigration :

```
        Pearson's product-moment correlation

data:  residus and delit$Immigration
t = 4.9673e-16, df = 94, p-value = 1
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2004859  0.2004859
sample estimates:
         cor
5.123428e-17
```

Le test de corrélation de Pearson entre les résidus de la régression et la variable Immigration est de -5.12e-17, ce qui indique une corrélation très faible. Ainsi le test ne montre pas de preuve de corrélation significative entre les résidus et la variable Immigration, car la p-value est supérieur au seuil de 0,05. Ainsi, il est peu probable que la variable Population soit endogène dans ce modèle.

Après application du test sur l'ensemble des variables, il s'avère qu'aucune n'est significative au niveau de 5%, nous avons des corrélations très faibles avec les résidus. Ainsi il semble peu probable que notre modèle comporte de biais d'endogénéité.

A noter qu'il est toujours intéressant de confirmer ses résultats à l'aide d'autres méthodes car un test ne capture pas forcément tout type de biais ou autre problématique. Le modèle à variables instrumentales peut par exemple être utilisé.

Causalité ?

Pour vérifier la causalité entre la variable Delit et l'une de nos variables explicatives, un test de causalité à l'aide de variable instrumentale pourrait être réalisé. Pour cela, l'idée est de trouver une variable instrumentale, c'est-à-dire une variable qui est corrélé avec la variable d'intérêt (variable explicative) mais qui n'est pas directement corrélée avec la variable à expliquer (Delit), sauf par l'intermédiaire de la variable d'intérêt. Le test de Wald via la commande waldtest() sur Rstudio est réalisable, ce dernier a pour hypothèse nulle que l'instrument n'a aucun effet sur la variable à expliquer (dans notre cas Delit). Une p-value inférieure au seuil de significativié (généralement 5%) alors on rejette l'hypothèse nulle et on conclut que l'instrument est valide et que la variable d'intérêt a un effet causal sur la variable à expliquer. A noter que pour réaliser ce type de test la variable instrumentale choisie doit être approprié.

Conclusion

Pour conclure, nous avons au travers de ce rapport résolu et traité les différents problèmes rencontrés auparavant liés à la multicolinéarité. Plusieurs méthodes ont été mise en œuvre dans le but d'optimiser et fiabiliser l'analyse. Les résultats en termes de métrique sont assez équivalents par rapport à la première étude. De bon modèle fiable pour notre étude.

Au vu des résultats obtenus, nous pouvons conforter l'idée du premier rapport que des facteurs d'accroissement du nombre d'habitants et des facteurs d'inégalités en termes de niveau d'éducation et de vitalité économique sont des éléments majeurs pouvant être considéré comme déterminants responsables des actes d'insécurité au travers des départements français.

Supprimer les variables faiblement corrélés avec l'endogène et ajouter de nouvelles variables fortement liés pourrait être une prochaine étape pour améliorer le modèle et trouver également les variables causales à notre taux de délit.

Identifier les variables causales permettrait de justifier des politiques publiques ou encore aider les autorités à mieux appréhender ces facteurs de risque pour mieux prévenir.

Annexe

Ci-dessous, le descriptif et l'origine des variables :

Code variable	Définition de la variable	Unité de la variable	Source de l'information
Delit (variable endogène)	Moyenne entre le taux de cambriolages pour 1000 logements et le taux de coups et blessures volontaires pour 1000 personnes (par département)	En (%)	www.insee.fr
Population	Nombre d'habitants recensés dans un département en 2018	Nombre d'habitant en milliers	https://public.opendatasoft.com/
Densite_pop	Nombre d'habitants au km² dans le département	Habitants/km²	www.observatoire-desterritoires.gouv.fr
Revenu_median	Revenu médian de la population active du département	En euros	www.insee.fr
X15.29ans	Part des habitants ayant entre 15 et 29 ans dans le département	En (%)	www.observatoire-desterritoires.gouv.fr
X60.74ans	Part des habitants ayant entre 60 et 74 ans dans le département	En (%)	www.observatoire-desterritoires.gouv.fr
Taux_pauvrete	Part des ménages dans le département vivant en dessous du seuil de pauvreté. (*Ce seuil est fixé à 60% du niveau de vie médian en France.)*	En (%)	www.insee.fr
Urbanisation	Taux d'urbanisation, la part de la population vivant en ville par rapport à la population totale.	En (%)	www.insee.fr
Scolarisation	Taux de scolarisation, la part des peu ou pas diplômé parmi les 16-24 ans qui ne sont pas en études en 2017-2018.	En (%)	www.insee.fr
Taux_chomage	Pourcentage des personnes de la population active qui sont au chômage.	En (%)	www.observatoire-desterritoires.gouv.fr
Immigration	Correspond à un taux d'immigration, la part de la population immigrée selon les pays de naissance autre que la France (UE et Reste du monde).	En (%)	www.insee.fr
Police	Taux de policiers et gendarmes pour 10 000 habitants dans les départements.	Nombre de policiers + gendarmes /10 000 habitants	www.insee.fr
Politique	Orientation politique départementale	/	www.interieur.gouv.fr

Pourcentage contribution au projet : Yann (40%), Nabil (30%) et Ibrahim (30%)