Rapport de Projet : Data Mining

Arthur DESBIAUX et Abdurrahman SEN

Abstract Dans ce rapport, nous décrivons les procédures et méthodologies utilisées pour l'exploration et l'extraction d'informations d'un jeu de données sur le nombre de victimes d'homicides par année dans le monde.

Introduction

L'objectif de ce projet est d'explorer un dataset choisi grâce à l'application de méthodes et outils présentés en cours. À ces fins, nous présenterons le dataset sélectionné (source et contenu) ainsi que les différentes méthodes appliquées ainsi que l'analyse que nous avons pu en tirer.

Questions principales

À travers le dataset sélectionné, nous souhaitions répondre aux questions suivantes :

- «*Quels sont les pays les plus meurtriers dans le monde ?*»
- «*Retrouve-t-on des régions géographiques très meurtrières ?*»
- «*Correspondent-elles aux continents mondiaux ?*»
- «*Retrouve-t-on des pays présentant des tendances meurtrières similaires ?*»

Dataset sélectionné

Le dataset est intitulé «*Intentional homicide victims by sex, counts and rates per 100,000 population*» et provient du site des Nations Unies (UN data) au format XLS. Il répertorie le nombre de victimes de meurtres et le taux de meurtres pour chaque pays selon le sexe des victimes.

Les données sont structurées telles que dans l'exemple donné dans le tableau 1. En l'occurence, on trouve des cellules vides et des en-têtes très

verbeux : «*Number of victims of intentional homicides by sex*» par exemple.

Prétraitement des données

La première étape avant tout consiste à nettoyer nos données. Nous avons des données manquantes selon les années, des en-têtes verbeux qui rendent pénible la manipulation du jeu de données et certaines colonnes qui ne nous serviront pas à étudier les éventuelles tendances

Colonne superflue

La colonne «*Source*» contient un sigle représentant la source d'un instance du dataset. Les sources, quant à elles, sont listées à la suite du tableau de valeurs. Puisque cette colonne et ces sources ne sont pas pertinentes à la manière dont nous souhaitons exploiter le dataset, nous avons décidé de les supprimer.

Renommage d'en-têtes

Les en-têtes de premier niveau sont très verbeux et long, nous les avons donc renommé en «*Count*» et «*Rate*» ainsi que la première colonne contenant les informations liées au pays (région, sous région, nom du pays, sexe des victimes) en «*Info*».

Données manquantes

Une question importante vis-à-vis des données manquantes dans notre dataset revient à leur signification : il peut s'agir de données nulles ou bien de données qui n'ont pas été reportées. Ainsi, le remplissage des valeurs manquantes par des zéros sous entend que le meurtre était inexistant dans le pays cette année là.

					Count			Rate		
Region	Subregion	Country	Sex	Source	2000	...	2021	2000	...	2021
Africa	N. Africa	Algeria	Female	CTS			160			0.74
Africa	N. Africa	Algeria	Male	CTS			535			2.38

TABLE 1. *Exemple des deux premières lignes du dataset avec leurs en-têtes*

Nous préférons nous servir de la moyenne du nombre de victimes ainsi que celle du taux de meurtre pour remplir nos données manquantes.
Il était également envisageable d'utiliser la plus petite valeur renseignée pour comparer les résultats, mais ce ne fut pas une piste que nous avons décidé d'explorer.

Après ces prétraitements, nous avons sauvegardé le dataset nettoyé dans un fichier CSV sous le nom «*clean_data.csv*».

Exploration des données

Pays les plus meurtriers

En rajoutant une colonne «*Total*» pour le nombre et le taux de meurtres à notre dataset, on peut déjà retrouver les pays les plus meurtriers. Ainsi, la figure 1 ci-dessous permet de visualiser l'évolution du nombre de victimes dans les 10 pays avec le plus grand nombre de victimes de meurtre sur la période 2000 - 2021.

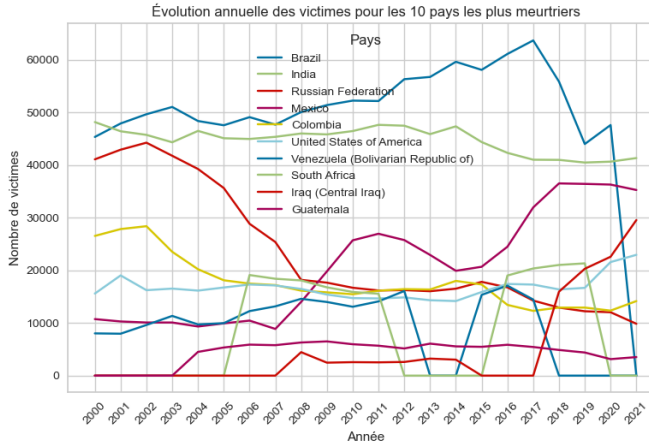

FIGURE 1. *Graphe des pays les plus meurtriers (selon le Count)*

On retrouve de grands pays avec une très grande population (e.g. Brésil, Méxique, Russie, Inde, États Unis...), mais également l'Afrique du Sud, qui possédait en 2022 un population de 59 millions d'habitants contre les 67 millions de la France.

Pour permettre une comparaison plus rapide et significative, on a également

cherché les 10 pays avec les plus grand taux de meurtres ramenés à 100 000 de population. La figure 2 présente l'évolution du taux de meurtres annuel parmi ces 10 pays là sur la période 2000 - 2021.

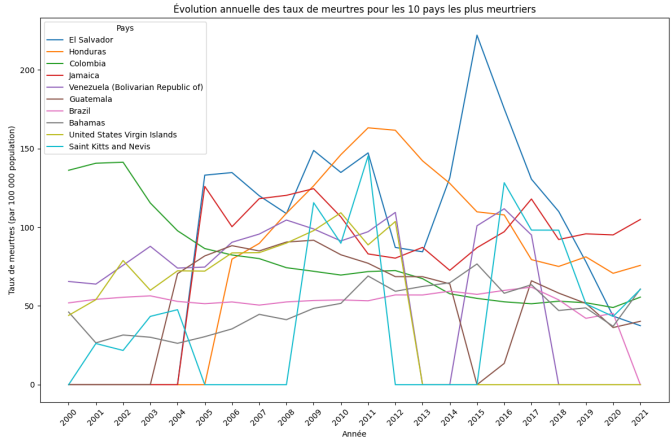

FIGURE 2. *Graphe des pays les plus meurtriers (selon le Rate)*

On retrouve la Colombie, le Guatemala et le Venezuela. On retrouve également le Brésil, mais cette fois-ci, à la 6ème place contrairement à la tête du classement précédent. La plupart du classement est occupé par de plus petits pays (e.g. Jamaïque, El Salvador, Honduras, Bahamas ...).

Sexe des victimes

Il nous semble pertinent de comparer ces nombres et taux de victimes de meurtres avec des classements par sexe des victimes (masculin/féminin). À ces fins, nous avons donc obtenu les classements présentés dans la figure 2 ci-dessous.

On remarque systématiquement que le nombre de victimes masculines est bien plus élevé (de 2 à 10 fois plus) dans chaque pays. Dans le cas des pays d'Amérique Latine, cela s'explique par le fort taux de criminalité en bande organisée.

On peut maintenant se poser la même question avec le taux de meurtres national cumulé sur la période 2000 - 2021, pour lequel on obtient les classements présentés dans le tableau 3.

On retrouve le Salvador en tête de classement à nouveau et le taux de meurtre des victimes masculines est 2 à 9 fois plus élevé que l'autre.

Rang	Féminin		Masculin	
	Pays	Nb Victimes	Pays	Nb Victimes
1	Inde	368897	Brésil	1006754
2	Russie	127340	Inde	615455
3	Brésil	88784	Mexique	405856
4	États Unis d'Amérique	80760	Russie	385161
5	Mexique	50573	Colombie	357596
6	Colombie	31415	États Unis d'Amérique	286672
7	Afrique du Sud	26725	Venezuela	189342
8	Guatemala	13649	Afrique du Sud	158797
9	Venezuela	11392	Irak	102075
10	Ukraine	10862	Guatemala	81741

TABLE 2. *Classement des 10 pays les plus meurtriers (Count) selon le sexe des victimes sur la période 2000-2021*

Rang	Féminin		Masculin	
	Pays	Taux de meurtre	Pays	Taux de meurtre
1	El Salvador	204.05	El Salvador	1822.98
2	Jamaïque	168.74	Honduras	1601.10
3	Russie	164.32	Colombie	1596.72
4	Honduras	144.23	Jamaïque	1540.39
5	Guatemala	137.81	Venezuela	1355.28
6	Colombie	137.29	Brésil	1043.05
7	St Vincent & Grenadines	112.11	Guatemala	996.86
8	Guyane	108.23	US Virgin Islands	966.36
9	Bahamas	104.66	Bahamas	955.40
10	Belize	103.54	Puerto Rico	902.61

TABLE 3. *Classement des 10 pays les plus meurtriers (Rate) selon le sexe des victimes sur la période 2000-2021*

En conclusion, on retrouve donc souvent El Salvador, la Jamaïque, le Brésil, le Venezuela, la Colombie, le Guatemala et le Honduras. Le nombre et taux de victimes masculines est systématiquement plus élevé (de 2 à 9 fois plus) que le nombre de victimes féminines. On peut expliquer cette différence par la présence de crime organisé dans le pays.

Dans la suite, on se demande s'il existe des groupes de pays avec des statistiques de meurtres similaires.

Clustering

En premier point de départ, nous souhaitions essayer d'appliquer un clustering KMeans pour chercher à retrouver les continents du dataset. En ayant enlevé la colonne «*Region*» et en utilisant un encodage OneHot pour les nom de pays, les sous régions et le sexe, on obtient le clustering présenté dans la figure 3.

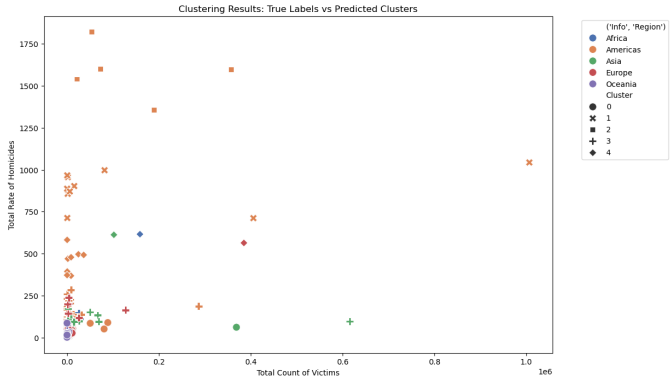

FIGURE 3. *Clustering Region obtenu. La couleur indique la véritable région, tandis que la forme représente le cluster associé.*

Sans prendre en compte les contraintes de régionalisation, on ne s'attendait pas à obtenir un résultat correct. On observe que des pays d'Amérique sont présent dans chacun des clusters et qu'aucun cluster ne contient qu'un seul continent.

Ce résultat s'explique par la quantité d'informations que l'on a passée au KMeans (nombre de victimes ET taux de meurtre, nom du pays encodé) mais également par la non utilisation de contraintes régionales.

Nous avons pris la décision de ne pas continuer à faire la tâche de clustering, car une tâche de régionalisation serait plus adaptée.

Regionalisation

Étant donné que nos données correspondent à des pays, il semble plus que pertinent de se servir de la Regionalisation pour améliorer le clustering. Or, nous ne possédons pas les latitudes et longitudes des pays. Nous devons donc compléter nos données.

Nous avons donc utilisé la bibliothèque «*CountryInfo*», seuls quelques pays

n'étaient pas trouvé/reconnu dans cette bibliothèque, nous avons donc complété les dernières informations à la main. Après ces modifications, nous avons sauvegardé le dataset contenant le nom du pays, sa position en latitude et sa position en longitude dans un fichier CSV sous le nom «*clean_data_pos.csv*».

Nous avons pris la décision de rassembler les données tout sexe confondu, mais également de conserver seulement les informations sur le taux de meurtre.

Pour réaliser notre régionalisation, nous utiliserons «*AgglomerativeClustering*» nous comparerons les résultats avec ou sans matrice de connectivité, mais également en changeant le nombre de cluster.

Dans un premier temps, on essayera de faire 7 clusters (un cluster par continent avec l'Amérique non unifiée)

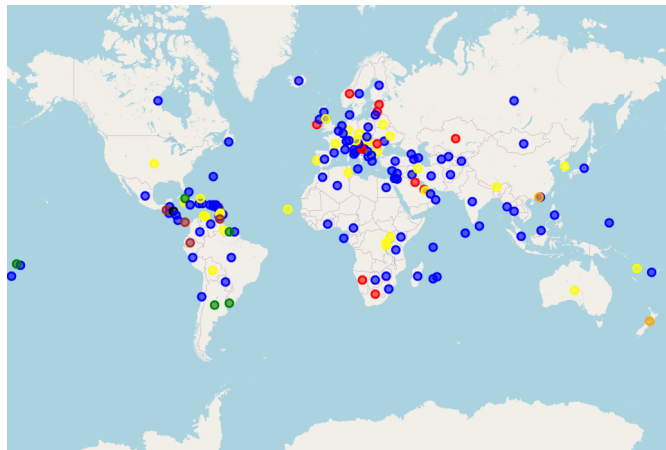

FIGURE 4. *Regionalisation 7 cluster avec matrice de connectivité*

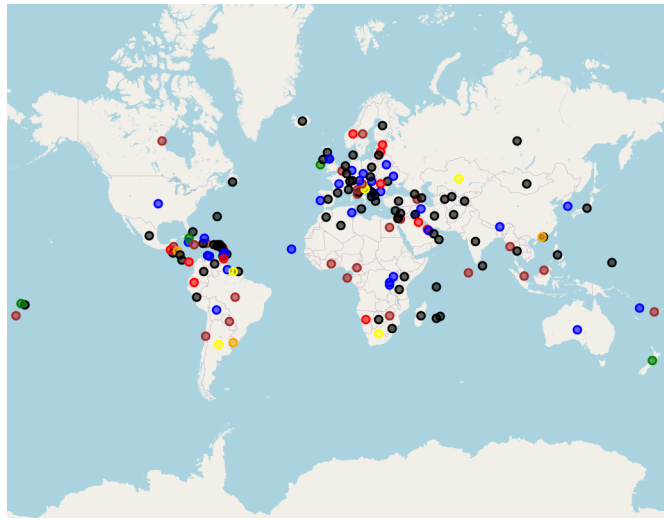

FIGURE 5. *Regionalisation 7 cluster sans matrice de connectivité*

Comme on le remarque dans la figure 4 ou 5, il est impossible avec nos données de pouvoir observer les 7 continents. On peut cependant voir quelques régions similaires comme l'Europe et le Moyen-Orient. On remarque peu de différence avec ou sans matrice de connectivité, ce qui s'explique du fait que l'algorithme peut plus facilement respecter les contraintes avec un nombre de clusters peu élevé.

Nous voudrons maintenant voir si la régionalisation nous permet d'observer les différentes sous-régions, donc un total de 17 clusters (ce nombre est calculé à partir du nombre de sous-régions différentes dans notre dataset).

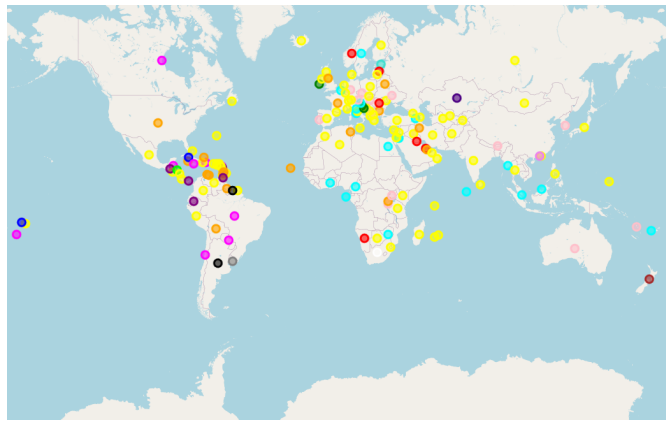

FIGURE 6. *Regionalisation 17 cluster avec matrice de connectivité*

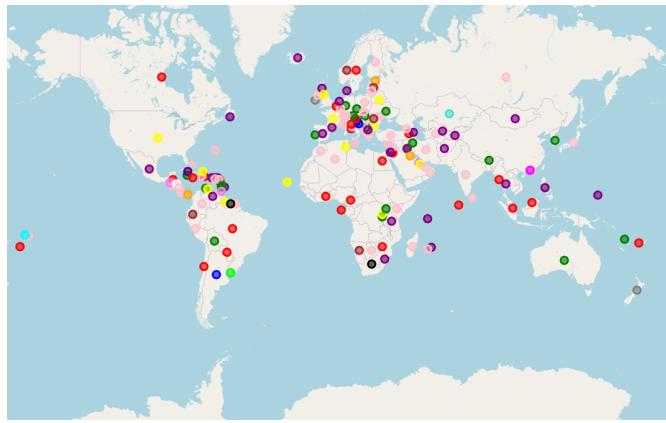

FIGURE 7. *Regionalisation 17 cluster sans matrice de connectivité*

Cette fois-ci, dans les figures 6 et 7 on remarque de grandes différences. Tout d'abord, sans matrice de connectivité, on semble avoir des résultats beaucoup plus aléatoires, alors qu'avec on peut observer quelques sous régions très semblables telle que l'Europe de l'Est ou encore les îles autour de Porto Rico. Cette différence peut s'expliquer dû au grand nombre de clusters : la matrice de connectivité limite le partitionnement et influence davantage la structure des clusters.

En conclusion, pour notre regionalisation, les résultats ne sont pas parfaits mais on peut reconnaître quelques régions. On peut attribuer cette difficulté au

peu de données que l'on possède mais également au taux de meurtre qui ne permet pas à lui seul de nous permettre de retrouver précisément des régions ou sous-régions.

Temporalisation

Notre dataset inclue deux séries de données temporelles : le nombre de victime par an et le taux de meurtre par an. Réaliser une étude sur les caractéristiques temporelles de nos données paraît plus que pertinent. L'objectif étant de répondre à la question «*Retrouve-t-on des pays présentant des tendances meurtrières similaires ?*».

Dans cette section, nous avons décidé de séparer notre DataFrame uni en deux, un pour le nombre de victimes et un pour le taux de meurtres. Par ailleurs, nous nous concentrons uniquement sur le taux de meurtres sans distinction du sexe des victimes puisque celui-ci permet une comparaison des pays sur un pied d'égalité.

Time Series KMeans

Nous avons réalisé un KMeans en cherchant 7 clusters (les continents) dont un exemple de résultat est montré par la figure 8. On observe qu'aucun cluster n'est vraiment satisfaisant (notamment par rapport à la moyenne en rouge), que 2 clusters regroupent l'enorme majorité des pays à l'exception d'une petite poignée qui se retrouve dans les 7 clusters restants.

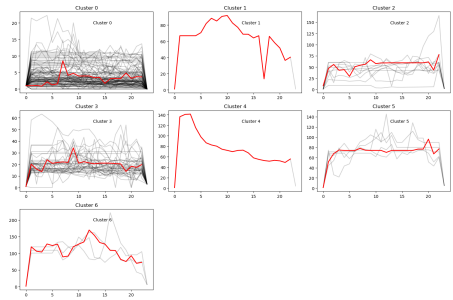

FIGURE 8. *Clustering temporel sur le taux de meurtre*

Périodicité (ACF)

Une question qui peut se poser à l'échelle d'un pays est : retrouve-t-on des tendances meurtrières au cours des deux décennies ? des périodes à haut taux de meurtres ?

Pour répondre à cette question, nous décidons de calculer la fonction d'Autocorrelation pour chacun des pays. La série temporelle considérée est le taux de meurtre annuel totu sexe confondus.

Le résultat complet est disponible en taille complète sur le dépôt GitHub à cette adresse.

De manière générale, on ne retrouve pas de périodicité dans les ACF de chaque pays. C'est un résultat plutôt attendu, il n'existe pas de saison annuelle de chasse à l'Homme. Cependant, on retrouve souvent des pays avec un ACF qui diminue progressivement de proche en proche. C'est le cas par exemple de l'Estonie et de la République Dominicaine (fig 9).

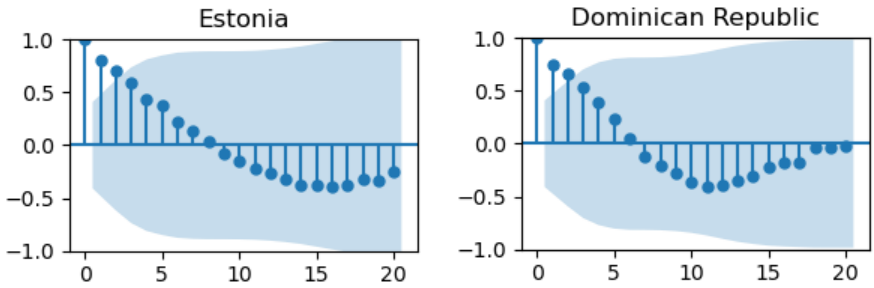

FIGURE 9. *ACF de l'Estonie et de la République Dominicaine*

Cette diminution progressive montre que les taux de meurtres évoluent petit à petit chaque année, contrairement à, par exemple, la Nouvelle Zélande et le Nicaragua (fig 10).

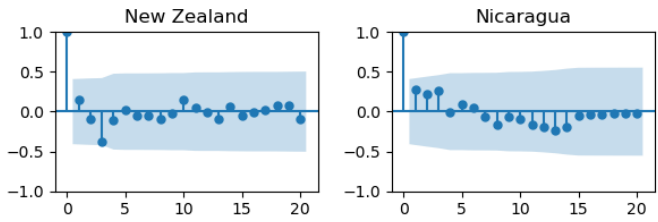

FIGURE 10. *ACF de la Nouvelle Zélande et du Nicagarua*

Pays voisins et Dynamic Time Warping

Qu'en est-il d'une sélection de pays voisins ? retrouve-t-on des tendances meurtrières proches ?

Pour répondre à cette question, nous avons décidé de sélectionner un sous-ensemble de pays spécifiques (France, Italie, Allemagne et Suisse) et de calculer la matrice de distances des scores DTW (*Dynamic Time Warping*).

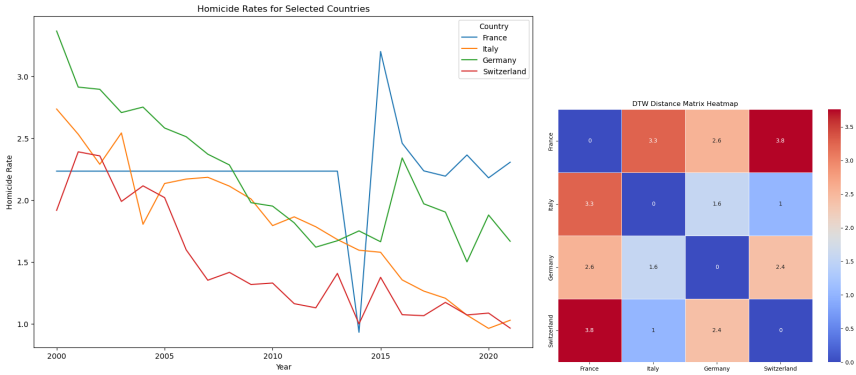

FIGURE 11. *Taux de meurtres de pays voisins de la France et Heatmap des distances entre scores DTW*

En s'appuyant sur cette matrice et la courbe d'évolution du taux de meurtres de ces 4 pays (fig 11, on observe que l'Italie et la Suisse sont les plus similaires. La France, quant à elle, se démarque de ses voisins avec la plus grande distance. On peut l'expliquer avec le haut taux de meurtres à partir de 2015, liés aux attentats du 13 novembre 2015.

Graph

Dans cette partie, la question que nous nous sommes posée est la suivante «*Peut on transformer notre dataset en graphe ? Et peut on en tirer des informations (closeness, betweenness. . .) ?*»

Nous avons pris la décision de rassembler les données tout sexe confondu, mais également de conserver seulement les informations sur le taux de meurtres et de remplir les informations manquantes par la moyenne.

Pour réaliser le graphe, nous avons fait une matrice de corrélation seulement sur les données des différentes années, puis on garde seulement les données les plus corrélées (nous avons choisi un seuil de 0.70). Une fois le graphe

construit, nous le sauvegardons en graphml, ainsi nous pouvons le manipuler avec Gephi.

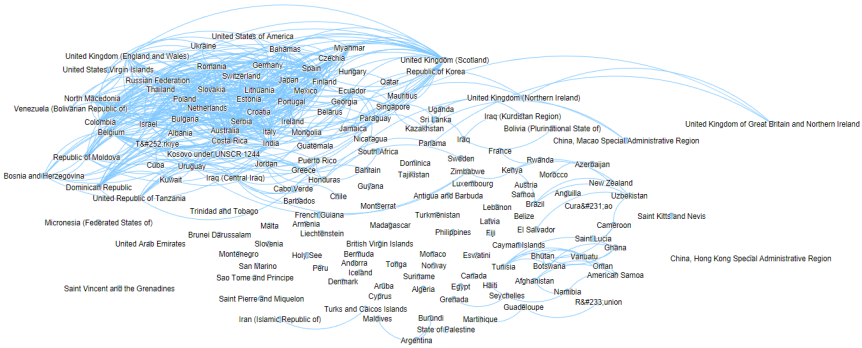

FIGURE 12. *Dataset transformé en graph*

Ce qu'on cherche maintenant à étudier est les différentes métriques de description de graphe pour voir si on peut tirer des informations de celui-ci, telles que la «*Closeness*» qui permet de savoir si l'on est "proche" du global, la «*Betweenness*» qui permettrait de savoir les pays qui servent d'intermédiaire entre deux pays avec un taux de criminalité fort et enfin la «*Degree centrality*» pour voir les pays les plus fortement corrélés avec d'autres pays.

Closeness

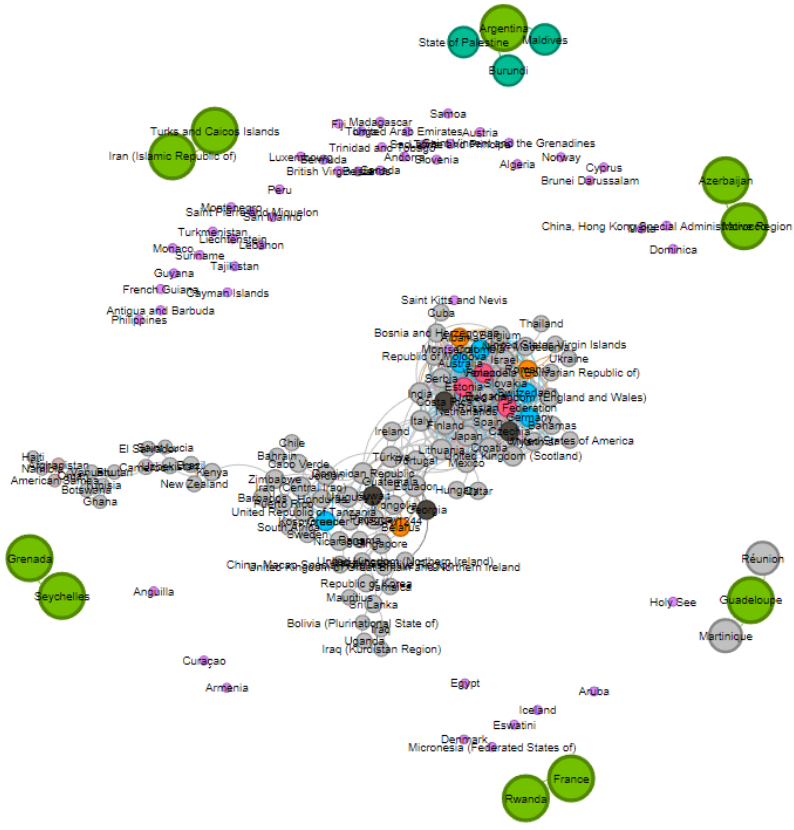

FIGURE 13. *Closeness sur le graph*

Dans la figure 13, les nœuds de couleur verte sont les plus hauts en closeness (Maroc, Rwanda, Seychelles, Argentine, Grenade, Guadeloupe, Îles Turques et Caïques, Iran, Azerbaïdjan, France) ils devraient donc être les plus proches au global en terme de criminalité, mais comme on peut le remarquer, ils ne sont pas connectés avec tous les nœuds. Il ressemble donc seulement au global des nœuds sur lesquels ils sont reliés. Pour avoir une analyse plus intéressante, nous allons retirer ces nœuds et garder seulement le plus grand graphe connexe.

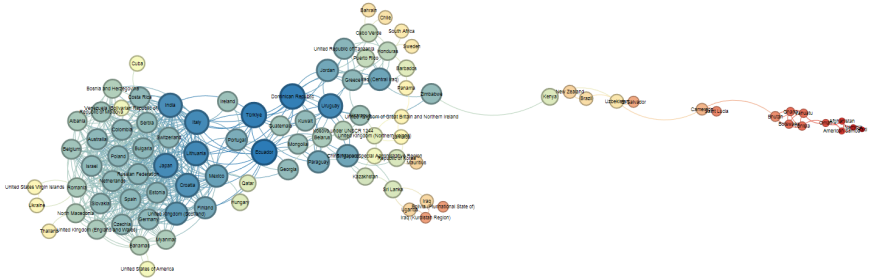

FIGURE 14. *Closeness sur le plus grand graphe connexe*

On se retrouve avec le graphe 14, on retrouve ici des nœuds de taille plus ou moins similaire. Dans le top 5 des plus gros pays en terme de closeness 4 on retrouve l'Équateur, la République Dominicaine, la Turquie, l'Italie et l'Uruguay, et donc que leur taux de criminalité est proche du global. Tout d'abord, on voit que la closeness la plus grande est de 0.26, ce qui est plutôt bas. On ne peut pas non plus tirer de grosses conclusions sur ce top. Cependant, il est intéressant de voir qu'on retrouve des pays venant de différents continents et non pas d'un seul (Amérique du Nord, Amérique du Sud, Europe et Asie).

Pays	**Closeness**
Équateur	0.3392
République Dominicaine	0.3345
Turquie	0.3345
Uruguay	0.3244
Italie	0.3244

TABLE 4. *Closeness les plus grandes*

Degree centrality

Passons maintenant à la métrique de «*Degree centrality*», ce que nous cherchons à voir est quels sont les pays les plus corrélés entre eux. Cela nous permettrait de voir notamment si les pays ayant un grand nombre de liens avec d'autres pays pourraient être des pays avec des comportements de criminalité qui influencent ou sont influencés par plusieurs autres pays. Pour cette métrique nous réutilisons notre graphe originel, car nous souhaitons conserver tous les liens.

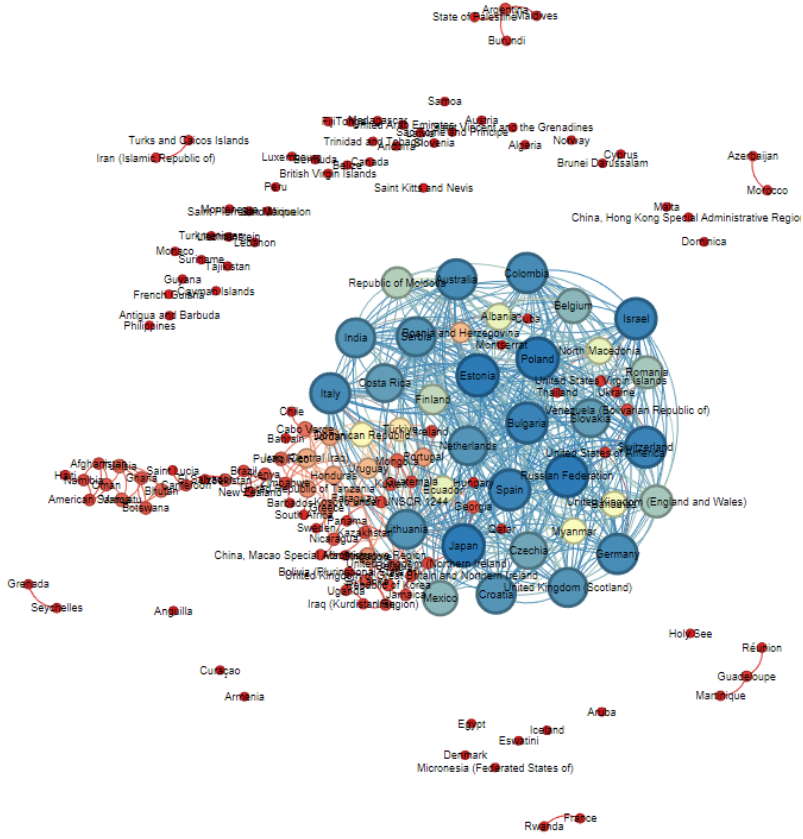

FIGURE 15. *Degree centrality sur le graph*

On se retrouve avec le graphe 15, on retrouve ici des nœuds de taille plus ou moins similaire. Dans le top 5 des plus gros pays en terme de degree centrality 5 on retrouve le Japon, la Pologne, la Russie, l'Estonie et Israël. Tout d'abord, on voit qu'ils ont tous une degree centrality plutôt proche (30). Ce qui est intéressant ici, c'est que ces pays viennent de deux régions : l'Asie et l'Europe de l'Est. Cela suggère que les pays ayant une degree centrality élevée ont des comportements de criminalité qui ne sont pas confinés à une seule région géographique, mais traversent plusieurs zones du monde. On observe aussi des pays extrêmement proches tels que l'Estonie, la Pologne et la Russie, on pourrait donc imaginer qu'ils s'influencent entre eux.

Pays	Degree
Russie	31
Pologne	31
Estonie	31
Japon	31
Israël	30

TABLE 5. *Degree les plus grands*

Betweenness

Finissons avec la métrique de «*Betweenness*», ce que nous cherchons à voir est quels sont les pays qui servent d'intermédiaire entre deux pays avec un taux de criminalité fort.

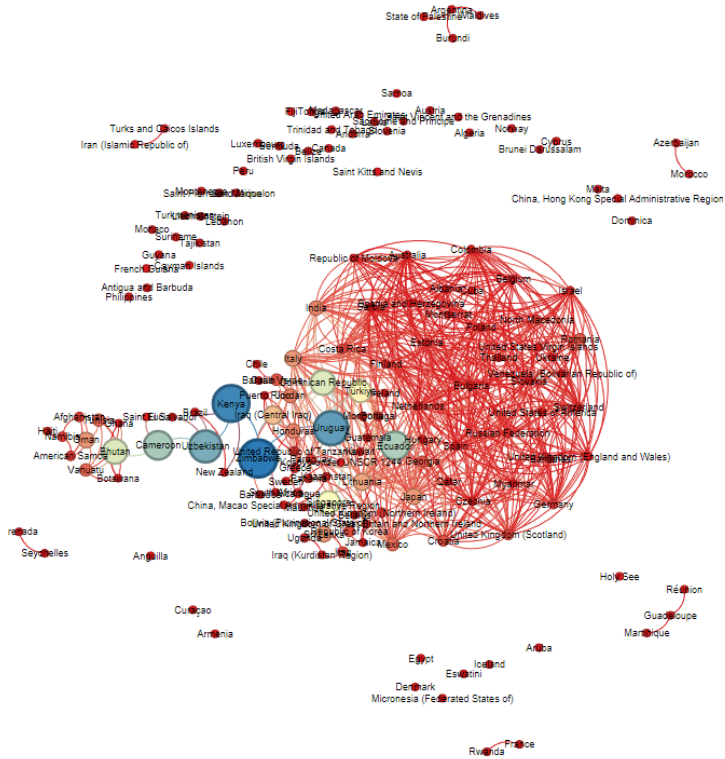

FIGURE 16. *Betweenness sur le graph*

On se retrouve avec le graphe 16, on retrouve ici des très petit nœuds

impliquant une faible betweenness dans l'ensemble du graphe. Dans le top 5 des plus gros pays en terme de betweeness 6, on remarque que les valeurs sont très petites, on ne peut donc pas déterminer clairement des pays qui relieraient d'autres pays avec un taux de criminalité similaire Il n'y a donc pas de pays qui joue un rôle de connecteur dans un réseau de criminalité.

Pays	**Betweenness**
Zimbabwe	0.1030
Kenya	0.0987
Uruguay	0.0901
Ouzbékistan	0.0836
Cameroun	0.0724

TABLE 6. *Betweenness les plus grandes*

Conclusion

D'après nos différents résultat, nous pouvons conclure que notre dataset ne nous permet pas de faire d'analyse très approfondie sur les tendances meurtrières des pays. La temporalisation, la régionalisation et la transformation en graph semblent tout de même être les meilleures techniques de Data Mining à appliquer pour extraire des informations de ce dataset.