

Projet crime : Projet Traitement de l'information Semestre 5

Maël PAUL

Nathan THIVIN

Louis-Victor LADAGNOUS

Léo-Paul MAZIERE

Décembre 2021

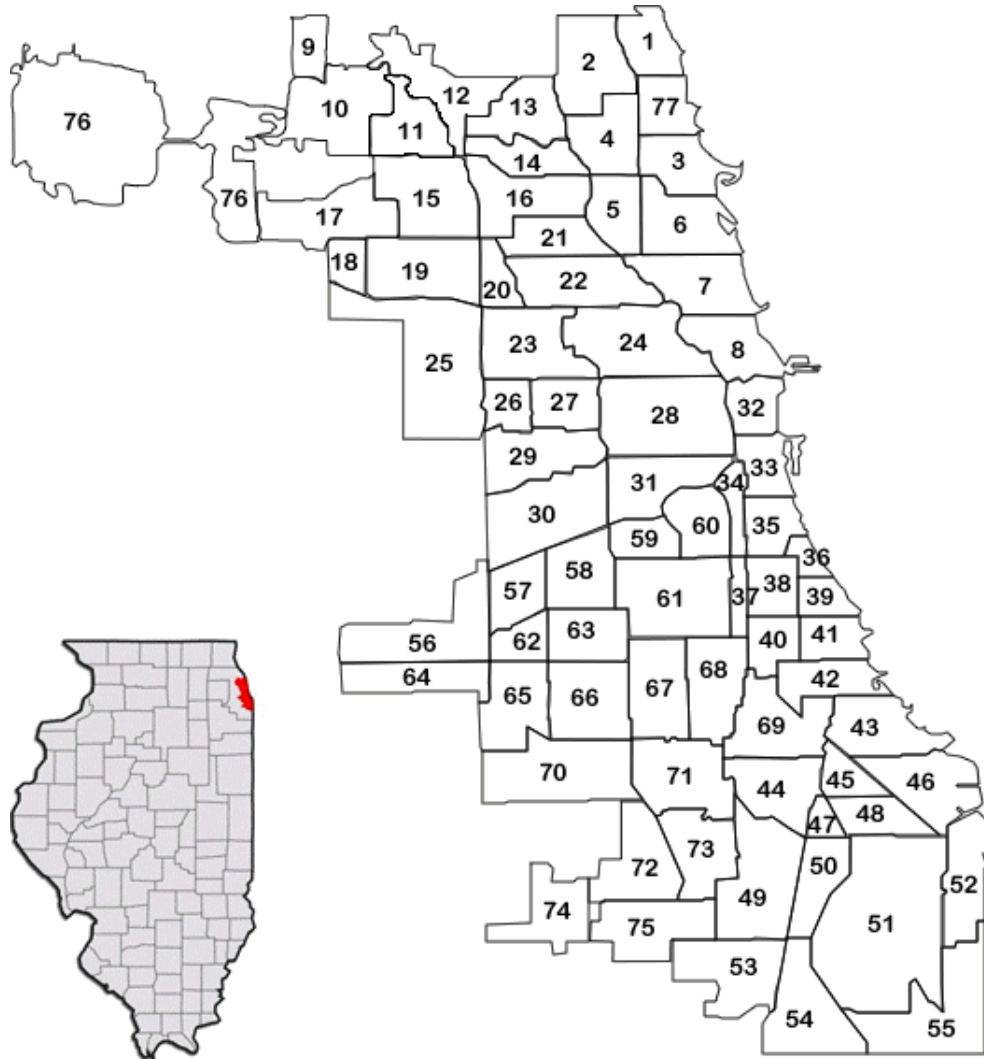


Table des matières

1	Introduction	3
2	Présentation des données	4
2.1	La Base de données	4
2.2	Les Variables utiles	4
3	Présentation de la méthode	6
3.1	Qu'est-ce que l'AFC ?	6
3.2	Tableau des effectifs ou tableau de contingence	6
3.3	Indépendance	6
3.4	Application à notre cas d'étude	6
4	Analyse des résultats	8
4.1	AFC sur les secteurs de Chicago	8
4.2	AFC sur les arrestations par crime	8
5	Sources	10

1 Introduction

Ce projet se base sur les crimes ayant eu lieu à Chicago entre 2012 et 2017. Le but est d'étudier, en premier lieu, le lien entre certains types de crimes et les secteurs de Chicago, mais aussi, dans un second temps, de visualiser les crimes où l'arrestation du criminel est plus ou moins probable. Pour ce faire, nous allons appliquer deux Analyses Factorielles des Correspondances ou AFC.


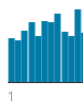


2 Présentation des données

2.1 La Base de données

La base de données utilisée dans cette analyse est une base de données répertoriant les crimes et délits s'étant produits à Chicago (voir lien dans les sources) entre 2001 et 2017. Cette base de données étant trop grande pour réaliser notre analyse simplement, le site nous propose de la restreindre aux crimes ayant eu lieu entre 2012 et 2017. C'est donc cette base de données que nous utiliserons.

La base de données est composée de 23 variables décrivant notamment le type de crime commis, les coordonnées GPS du lieu du crime et le booléen indiquant si son auteur a été arrêté ou non.

Crimes in Chicago from 2001 - 2017					
Primary Type	Description	Location Descript...	Arrest	# District	
THEFT 23%	SIMPLE 10%	STREET 23%			
BATTERY 18%	\$500 AND UNDER 9%	RESIDENCE 16%			
Other (863554) 59%	Other (1170078) 80%	Other (892713) 61%			
BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	10.0	
BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	3.0	
PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET	False	15.0	
BATTERY	SIMPLE	SIDEWALK	False	15.0	
THEFT	\$500 AND UNDER	RESIDENCE	False	15.0	
BATTERY	AGGRAVATED: HANDGUN	STREET	False	6.0	
BATTERY	SIMPLE	CHA HALLWAY/STAIRWELL /ELEVATOR	False	1.0	
BATTERY	SIMPLE	RESIDENCE PORCH/HALLWAY	False	2.0	

2.2 Les Variables utiles

Dans ce projet nous allons réaliser deux Analyses Factorielles des Correspondances (AFC) afin d'étudier les corrélations de différentes variables de la base de données.

Les variables de la base de données utilisées dans ces analyses sont :

- Le type de crime ou délit commis : il s'agit de données qualitatives nominales

```
Crime_type = ['HUMAN TRAFFICKING', 'OTHER NARCOTIC VIOLATION', 'PUBLIC INDECENCY', 'CONCEALED CARRY LICENSE VIOLATION', 'NON-CRIMINAL', 'OBSCENITY', 'INTIMIDATION', 'STALKING', 'KIDNAPPING', 'LIQUOR LAW VIOLATION', 'GAMBLING', 'ARS ON', 'HOMICIDE', 'SEX OFFENSE', 'INTERFERENCE WITH PUBLIC OFFICER', 'CRIM SEXUAL ASSAULT', 'PROSTITUTION', 'OFFENSE INVOLVING CHILDREN', 'PUBLIC PEACE VIOLATION', 'WEAPONS VIOLATION', 'CRIMINAL TRESPASS', 'ROBBERY', 'MOTOR VEHICLE THEFT', 'DECEPTIVE PRACTICE', 'BURGLARY', 'OTHER OFFENSE', 'ASSAULT', 'NARCOTICS', 'CRIMINAL DAMAGE', 'BATTERY', 'THEFT']
```

- Le district de la ville dans lequel s'est produit le crime : il s'agit de données qualitatives ordinales

```
district = ['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38', '39', '40', '41', '42', '43', '44', '45', '46', '47', '48', '49', '50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60', '61', '62', '63', '64', '65', '66', '67', '68', '69', '70', '71', '72', '73', '74', '75', '76', '77']
```

- Le booléen indiquant si l’auteur du crime a été arrêté ou non : il s’agit de données qualitatives nominales

```
arrested = ['False', 'True']
```

La première AFC nous permettra d’étudier la corrélation entre le type de crime commis et le district dans lequel s’est produit le crime, nous chercherons à savoir si certains districts sont plus concernés que d’autres par la criminalité ou alors plus favorables à un type de crime en particulier.

La seconde AFC nous permettra d’étudier la corrélation entre le type de crime commis et l’arrestation éventuelle de son auteur, nous chercherons à savoir si les chances d’être arrêté sont les mêmes quelque soit le crime commis.

3 Présentation de la méthode

3.1 Qu'est-ce que l'AFC ?

L'AFC ou Analyse Factorielle des Correspondances est une méthode de traitement de données qui s'applique sur des données de type qualitatives (nominales ou ordinales) : c'est-à-dire des catégories (ordonnées ou non) bien définies comptant chacune un certain nombre n d'éléments ou individus.

Cette méthode fait intervenir deux variables ou questions que l'on nommera A et B pouvant prendre plusieurs valeurs ou réponses (A_i , B_j , etc). L'objectif est d'établir ou non la corrélation entre A et B grâce à l'utilisation d'un tableau croisé de leurs valeurs (réponses) sur le même ensemble N (qui contient les n éléments (individus)).

3.2 Tableau des effectifs ou tableau de contingence

La manière de remplir le tableau est simple, il suffit de parcourir les n éléments et ajouter 1 à la case correspondante:

Si l'individu a répondu A_i à la question A et B_j à la question B on ajoute 1 à la case (A_i, B_j) .

L'étape suivante consiste à transformer ce tableau des effectifs en tableau de probabilités, pour cela on divise tous ses éléments par n .

Grâce aux valeurs obtenues, on peut calculer les probabilités marginales de chaque ligne et chaque colonne, pour ce faire on somme les valeurs de la ligne ou de la colonne correspondante.

Ces valeurs nous seront utiles car elles permettent d'établir le tableau des effectifs théoriques dont nous avons besoin plus tard pour étaler ou non l'indépendance des variables A et B .

Pour remplir ce second tableau, dans une case donnée on met la valeur du produit entre: l'effectif total n , la probabilité marginale de la ligne et la probabilité marginale de la colonne obtenues précédemment.

Dans la case (A_i, B_j) du nouveau tableau, on met le produit: $n \times$ masse ligne $i \times$ masse colonne j .

On est maintenant en mesure d'évaluer l'indépendance.

3.3 Indépendance

On applique le test du χ^2 pour évaluer la distance entre les effectifs réels et théoriques.

La valeur du χ^2 est la somme des distances entre les deux tableaux: la somme des distances entre chaque case et son homologue dans l'autre tableau.

Pour calculer la distance, on prend le carré de la différence entre les valeurs divisée par l'effectif théorique.

La distance entre (A_i, B_j) réel et (A_i, B_j) théorique est:

$$\left((A_i, B_j)_{\text{réel}} - (A_i, B_j)_{\text{théorique}} \right)^2 \div (A_i, B_j)_{\text{théorique}}$$

Ne reste plus qu'à calculer le nombre de degrés de liberté:

(nombre de lignes - 1) \times (nombre de colonnes - 1) ou $(i - 1) \times (j - 1)$.

On peut ensuite, avec l'aide d'une table du χ^2 , en déduire la p-value qui nous permet en fonction de sa valeur de juger de la corrélation.

3.4 Application à notre cas d'étude

Correspondance avec notre étude:

Premier cas: Corrélation entre le type de crime et le district dans lequel il s'est produit

– Variable A: le type de crime

```
Crime_type = ['HUMAN TRAFFICKING', 'OTHER NARCOTIC VIOLATION', 'PUBLIC INDECENCY', 'CONCEALED CARRY LICENSE VIOLATION', 'NON-CRIMINAL', 'OBSCENITY', 'INTIMIDATION', 'STALKING', 'KIDNAPPING', 'LIQUOR LAW VIOLATION', 'GAMBLING', 'ARS ON', 'HOMICIDE', 'SEX OFFENSE', 'INTERFERENCE WITH PUBLIC OFFICER', 'CRIM SEXUAL ASSAULT', 'PROSTITUTION', 'OFFENSE INVOLVING CHILDREN', 'PUBLIC PEACE VIOLATION', 'WEAPONS VIOLATION', 'CRIMINAL TRESPASS', 'ROBBERY', 'MOTOR VEHICLE THEFT', 'DECEPTIVE PRACTICE', 'BURGLARY', 'OTHER OFFENSE', 'ASSAULT', 'NARCOTICS', 'CRIMINAL DAMAGE', 'BATTERY', 'THEFT']
```

– Variable B: le district

```
district = ['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38', '39', '40', '41', '42', '43', '44', '45', '46', '47', '48', '49', '50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60', '61', '62', '63', '64', '65', '66', '67', '68', '69', '70', '71', '72', '73', '74', '75', '76', '77']
```

Deuxième cas: Corrélation entre le type de crime et l'arrestation

– Variable A: le type de crime

```
Crime_type = ['HUMAN TRAFFICKING', 'OTHER NARCOTIC VIOLATION', 'PUBLIC INDECENCY', 'CONCEALED CARRY LICENSE VIOLATION', 'NON-CRIMINAL', 'OBSCENITY', 'INTIMIDATION', 'STALKING', 'KIDNAPPING', 'LIQUOR LAW VIOLATION', 'GAMBLING', 'ARS ON', 'HOMICIDE', 'SEX OFFENSE', 'INTERFERENCE WITH PUBLIC OFFICER', 'CRIM SEXUAL ASSAULT', 'PROSTITUTION', 'OFFENSE INVOLVING CHILDREN', 'PUBLIC PEACE VIOLATION', 'WEAPONS VIOLATION', 'CRIMINAL TRESPASS', 'ROBBERY', 'MOTOR VEHICLE THEFT', 'DECEPTIVE PRACTICE', 'BURGLARY', 'OTHER OFFENSE', 'ASSAULT', 'NARCOTICS', 'CRIMINAL DAMAGE', 'BATTERY', 'THEFT']
```

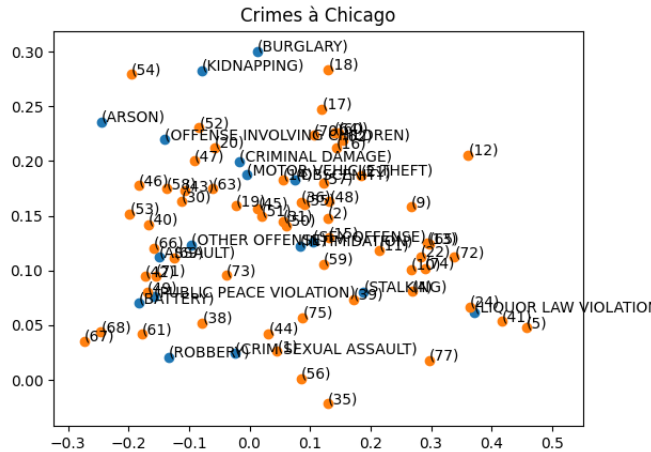
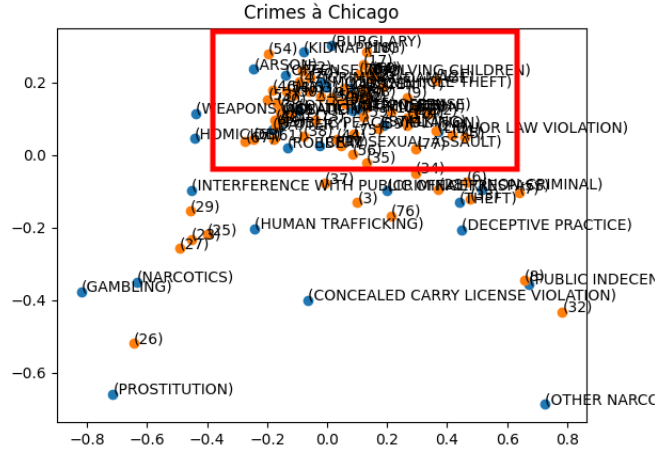
– Variable B: l'arrestation

```
arrested = ['False', 'True']
```

4 Analyse des résultats

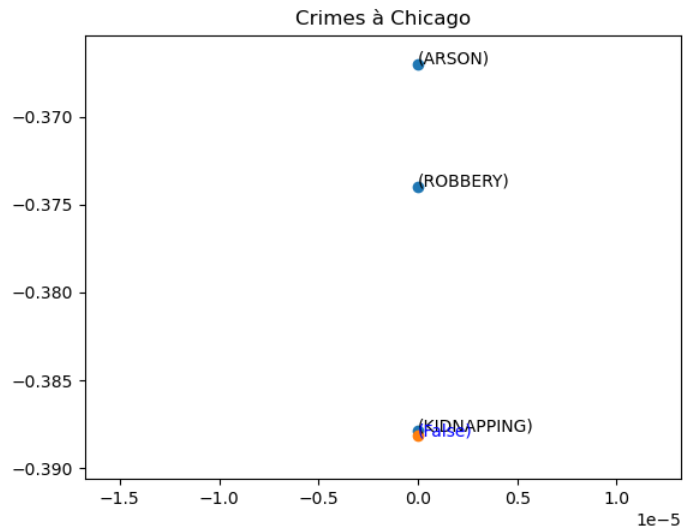
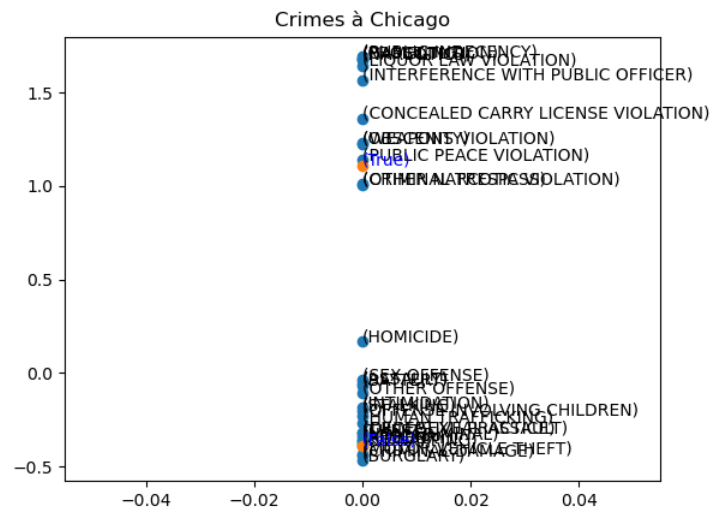
4.1 AFC sur les secteurs de Chicago

Cette première AFC consiste à analyser les corrélations entre certains types de crimes et les secteurs de Chicago. Le χ^2 confirme bien qu'il y a une corrélation entre les secteurs et les types de crimes. On peut observer, par exemple, que la prostitution, le trafic de drogues et les crimes liés aux jeux d'argent sont proches du secteur 26, ce qui peut suggérer la présence de mafias et de gangs dans ce secteur. En zoomant sur la zone en rouge, on peut aussi observer les zones les plus denses en crime.



4.2 AFC sur les arrestations par crime

La deuxième AFC consiste à analyser la relation entre le type de crime et l'arrestation ou non du criminel. Les données étant projetées sur une seule ligne, il est difficile de les visualiser correctement. Néanmoins, en zoomant sur les points True et False pour les arrestations, on retrouve des crimes comme le kidnapping (où le criminel est très rarement arrêté) et d'autres crimes tels que l'homicide qui est beaucoup plus équilibré dans ses arrestations. (La deuxième image ci-dessous est un zoom sur certains points du côté False).



5 Sources

Source de la base de données :

<https://www.kaggle.com/currie32/crimes-in-chicago>

Site internet utilisé pour la rédaction du rapport :

<https://www.overleaf.com>