



Master 2 Ingénierie Statistique

Elie LOLE

---

Fabrication et mise à disposition de fonds de carte et  
de modèles d'indicateurs spatiaux à partir de  
données pénitentiaires

---

**Référent universitaire :**

Aymeric Stamm

**Encadrante Stage :**

Camille Dorion



28 Août 2024

# Remerciements

Je souhaite exprimer ma profonde gratitude à Madame Camille Dorion pour la confiance qu'elle m'a accordée et pour m'avoir donné l'opportunité de travailler sur le projet cartographique. Son accompagnement et son encouragement ont été essentiels au bon déroulement de ce stage. Merci de m'avoir transmis votre savoir-faire.

Je tiens également à remercier Monsieur Aymeric Stamm et Madame Lise Bellanger, ainsi que l'ensemble du corps enseignant du Master Ingénierie Statistique de Nantes Université pour la qualité de la formation que j'ai reçue.

Enfin, je souhaite remercier chaleureusement les membres du Service de la Statistique, des Études et de la Recherche pour leur accueil convivial et leur contribution à un cadre de travail agréable et enrichissant.

# Glossaire

Ce glossaire répertorie les définitions et abréviations des termes spécifiques au monde pénal et pénitentiaire, afin de faciliter la compréhension du lecteur.

**PPSMJ - Personnes placées sous main de justice** : Mettre une personne sous main de Justice signifie placer cette personne sous l'autorité de la Justice. Cette personne fait donc l'objet d'une mesure restrictive ou privative de liberté par décision de Justice ou d'une peine en milieu fermé ou en milieu ouvert.

**Milieu Ouvert** : Une personne placée sous main de justice en milieu ouvert est suivie pour des mesures sans écrou telles que les alternatives aux poursuites, les mesures pré-sentencielles, les ajournements ainsi que les mesures post-sentencielles et les mesures de sûreté.

**Milieu Fermé** : Une personne placée sous main de justice en milieu fermé est dite écrouée.

**Écroué** : Est écrouée toute personne ayant fait l'objet d'une mise sous écrou. L'écrou est l'acte par lequel est établie la prise en charge par l'administration pénitentiaire des personnes placées en détention provisoire ou condamnées à une peine privative de liberté. La personne écrouée peut être hébergée au sein d'un établissement pénitentiaire ou non.

**Détenu** : Est détenue toute personne faisant l'objet d'une mesure privative de liberté à l'intérieur d'un établissement pénitentiaire.

**Non détenu** : Est non détenue toute personne ayant fait l'objet d'une mise sous écrou mais ne faisant pas l'objet d'une mesure privative de liberté à l'intérieur d'un établissement pénitentiaire. Il s'agit des personnes placées en détention à domicile sous surveillance électronique ou en placement à l'extérieur sans hébergement.

**CO - Condamné** : Est condamnée toute personne déclarée coupable d'un délit ou d'un crime. Une personne écrouée condamnée peut être détenue ou non détenue, selon les modalités d'aménagement ou d'exécution de sa peine.

**PR - Prévenu** : Est prévenue toute personne poursuivie se trouvant dans l'attente d'un jugement ou n'ayant pas encore été définitivement condamnée.

**CP - Condamné-prévenu** : Est à la fois condamnée et prévenue toute personne ayant été condamnée définitivement dans une ou plusieurs affaires et prévenue dans une ou

plusieurs autres affaires. Une personne peut en effet être concernée par plusieurs affaires différentes.

**SL - Semi-liberté** : La semi-liberté permet à une personne condamnée de bénéficier d'un régime particulier de détention l'autorisant à quitter durant la journée son établissement pénitentiaire afin d'exercer une activité professionnelle, de participer de manière essentielle à sa vie de famille, de suivre un traitement médical ou de s'investir dans tout autre projet d'insertion ou de réinsertion de nature à prévenir les risques de récidive. Elle peut être exécutée soit dans le cadre d'un aménagement de peine soit dans le cadre d'une libération sous contrainte.

**Quantum de peine** : Durée de la peine prononcée après le jugement.

**Reliquat** : Peine restant à purger .

**MA - Maison d'arrêt** : Les maisons d'arrêt reçoivent généralement les personnes prévenues en détention provisoire ainsi que les personnes condamnées dont la peine ou le reliquat de peine n'excède pas deux ans.

**CD- Centre de détention** : Les centres de détention sont des établissements pour peine qui accueillent les détenus condamnés à une peine supérieure à 2 ans.

**DISP** : Direction Interrégionale des Services Pénitentiaires

**SPIP** : Service Pénitentiaire d'Insertion et de Probation , est une structure chargée de l'accompagnement et de la réinsertion des personnes placées sous main de justice

**ASPIP** : Antenne SPIP

**TJ** : Tribunal Judiciaire

**CA - Cour d'appel** : Une juridiction qui examine les recours contre les jugements rendus en première instance par les tribunaux .

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Contexte du projet cartographique . . . . .	2
1.2	Structure d'accueil . . . . .	2
<b>2</b>	<b>Fondamentaux cartographiques</b>	<b>4</b>
2.1	Introduction aux fonds de cartes . . . . .	4
2.2	Types de fonds de cartes . . . . .	5
2.2.1	Cartes matricielles ou raster . . . . .	5
2.2.2	Cartes vectorielles . . . . .	6
2.3	Les projecteurs . . . . .	7
2.4	Catégories de cartes . . . . .	8
2.4.1	Cartes topographiques . . . . .	8
2.4.2	Cartes thématiques . . . . .	8
2.4.3	Types de visualisations . . . . .	10
<b>3</b>	<b>Création de fonds de cartes judiciaires et pénitentiaires</b>	<b>12</b>
3.1	Présentation de GESTSR . . . . .	12
3.2	Fonds de carte de base des communes . . . . .	13
3.3	Fonds de cartes des TJ et CA . . . . .	16
3.4	Cartes des ressorts pénitentiaires . . . . .	18
3.4.1	Liens entre les ressorts et les communes . . . . .	18
3.4.2	Fonds de carte obtenus . . . . .	20
3.5	Géolocalisation des TJ et centres pénitentiaires . . . . .	22
3.5.1	Cas des TJ . . . . .	22
3.5.2	Géolocalisation des établissements pénitentiaires . . . . .	24
<b>4</b>	<b>Création d'indicateurs spatiaux</b>	<b>26</b>
4.1	Présentation de GENESIS . . . . .	27
4.2	Présentation des données . . . . .	28
4.3	Pré-traitement des données . . . . .	30
4.3.1	Les données manquantes . . . . .	30
4.3.2	Restructuration de la base de données . . . . .	31
4.3.3	Ajout de variables supplémentaires . . . . .	33
4.4	Analyses exploratoires des données . . . . .	34
4.4.1	Analyse en composantes principales . . . . .	34
4.4.2	Classification des maisons d'arrêt . . . . .	37
4.5	Création d'un modèle de prédiction du taux de surpopulation des maisons d'arrêt . . . . .	39
4.5.1	Méthodes d'apprentissage statistique . . . . .	39

4.5.2	Comparaison des résultats . . . . .	40
4.6	Représentation spatiale des taux de surpopulation des Maisons d'Arrêt .	43
<b>5</b>	<b>Conclusion et appréciation de l'expérience du stage</b>	<b>45</b>

# Table des figures

2.1	Exemple de supersitions de couches [1]. . . . .	4
2.2	Différences entre les fonds de cartes raster et vectoriels [1]. . . . .	6
2.3	Déformations observées avec le WGS84 . . . . .	7
2.4	Nombre d’infractions confondues pour 100k habitants jugés dans la CA de Rennes . . . . .	9
2.5	Proportion d’infractions confondues jugées dans la CA de rennes . . . . .	10
2.6	Comparaison des taux d’excès de vitesse par commune à Rennes . . . . .	10
2.7	Aperçu d’un fond de carte généré avec leaflet . . . . .	11
3.1	Exemple de lien père-fils . . . . .	13
3.2	Fond de carte initial des communes de la France entière de l’Insee . . . . .	14
3.3	Fond de carte initial après rapprochement . . . . .	15
3.4	Fond de carte des TJ en France entière . . . . .	17
3.5	Fond de carte des CA en France entière . . . . .	17
3.6	Résumé des jointures clés effectuées . . . . .	20
3.7	Fond de carte des DISP en France hexagonale . . . . .	21
3.8	Fond de carte des SPIP en France hexagonale . . . . .	21
3.9	Fond de carte des ASPIP en France hexagonale . . . . .	22
3.10	Aperçu du fond de carte interactive des TJ en France entière et les COM . . . . .	23
3.11	Zoom sur les TJ de la France Hexagonale . . . . .	24
3.12	Zoom sur le TJ de Nantes . . . . .	24
3.13	Etablissement pour Mineurs d’Orvault . . . . .	25
4.1	Evolution du nombre de détenus depuis 2022 . . . . .	26
4.2	Les tables issues de GENESIS . . . . .	28
4.3	Pourcentage des données manquantes des variables . . . . .	30
4.4	Pourcentage des données manquantes après imputation . . . . .	31
4.5	Pourcentage de variance expliquée par les composantes principales . . . . .	35
4.6	Cercle de corrélations des variables . . . . .	36
4.7	Coordonnées des individus dans le plan 1-2 . . . . .	37
4.8	Plan factoriel des individus avec clusters 3-means . . . . .	38
4.9	Valeurs d’entraînement ajustées par chaque modèle . . . . .	41
4.10	Valeurs prédites par chaque modèle . . . . .	42
4.11	Prédiction du taux de surpopulation moyenne des MA par DISP au mois d’Août 2024 . . . . .	44

# Chapitre 1

## Introduction

Le ministère de la Justice, en tant qu'institution clé de l'État, joue un rôle crucial dans l'administration de la justice et la gestion des affaires pénales et pénitentiaires, et c'est au sein d'un de ses services statistiques que j'ai effectué mon stage. La production et l'analyse de données statistiques sont essentielles pour évaluer l'efficacité des politiques judiciaires, identifier les tendances et prendre des décisions éclairées. Cependant, l'analyse des données statistiques peut être limitée si elle ne prend pas en compte la dimension géographique des phénomènes observés. La visualisation d'indicateurs statistiques sur des fonds de cartes spécifiques permet de surmonter cette limitation en offrant une analyse par territoire judiciaire, qu'il s'agisse de tribunaux judiciaires ou de ressorts pénitentiaires. La question de la mise à disposition de ces fonds de cartes n'a pas encore été résolue de manière satisfaisante. Cette lacune pose plusieurs problématiques : comment créer et mettre à disposition des fonds de cartes pour garantir une couverture géographique complète et précise des territoires judiciaires ? Comment faciliter l'accès et l'utilisation de ces fonds de cartes pour les différentes équipes du ministère ? Comment valoriser les données métier sous forme d'indicateurs spatiaux pour améliorer la prise de décision ?

L'objectif principal de ce stage est de répondre à ces problématiques en développant une solution innovante qui permet de fabriquer et de mettre à disposition des fonds de carte spécifiques au ministère de la Justice. Dans un second temps, il s'agira de caractériser la surpopulation des maisons d'arrêt en fonction des caractéristiques des établissements et des personnes écrouées, de prédire le taux de surpopulation dans ces maisons d'arrêt, puis enfin de mettre en valeur ces prédictions sur des fonds de carte créés.

Ce rapport présente l'ensemble des travaux que j'ai effectués lors de mon stage de deuxième année de master en Ingénierie Statistique à Nantes Université. Je commencerai par exposer le contexte du projet cartographique, puis je présenterai la structure d'accueil de mon stage. Ensuite, j'aborderai les fondamentaux de la cartographie, notamment les différents types de cartes et les projecteurs. De plus, je présenterai GESTSR et les fonds de carte créés à partir des données de ce centre, qui m'ont permis de créer des fonds de cartes pénitentiaires allant des plus simples, comme celles des TJ, aux plus complexes, comme celles des ressorts pénitentiaires. Je présenterai également GENESIS, le centre de données qui m'a permis de créer des indicateurs sur la surpopulation des maisons d'arrêt, ainsi que toutes les étapes du prétraitement mises en œuvre, l'analyse exploratoire de ces données, et les méthodes d'apprentissage statistiques. Enfin, je mettrai en valeur ces indicateurs sur des fonds de carte.



## 1.1 Contexte du projet cartographique

Le besoin de visualiser des indicateurs sur des fonds de cartes spécifiques au ministère de la Justice a été exprimé par de nombreuses équipes du SSER. La visualisation d'indicateurs statistiques sur une carte offre une perspective géographique immédiate, essentielle pour l'analyse par territoire judiciaire, que ce soit pour les tribunaux judiciaires ou les services pénitentiaires. Cette dimension géographique est cruciale pour comprendre et analyser les phénomènes observés, mais les solutions actuelles présentent des insuffisances.

Bien que le site internet de l'observatoire des territoires fournisse certains fonds de cartes, il néglige certaines juridictions, notamment en outre-mer, et la question de la mise à disposition des données reste en suspens. En termes de logiciels de cartographie, le plus courant est QGIS. Cependant, il présente certaines difficultés d'appropriation au sein de notre bureau. En effet, QGIS nécessite une formation spécifique et une certaine expertise pour être utilisé efficacement. Par exemple, au sein du BSPP, seule une personne dispose de ce logiciel et maîtrise son utilisation. Cette situation limite la collaboration et la flexibilité dans la réalisation des projets cartographiques.

En revanche, R est plus accessible et peut être appris relativement rapidement par les membres de l'équipe, grâce à sa syntaxe intuitive et à la disponibilité de nombreuses ressources en ligne. De plus, l'utilisation de R permet la génération de rapports automatisés, ce qui correspond parfaitement aux besoins du bureau. La solution envisagée dans le cadre de ce stage est de créer et de mettre à disposition des fonds de carte en utilisant donc le logiciel R, en s'appuyant sur les bases de données du référentiel du ministère qui contient les organismes judiciaires rattachés. Ce projet vise à pallier les insuffisances actuelles en produisant des cartes précises et adaptées aux besoins spécifiques du ministère, tout en valorisant les données métier sous forme d'indicateurs spatiaux. Cette initiative permettra ainsi une meilleure visualisation et compréhension des données, facilitant les analyses et prises de décision au sein du ministère de la Justice.

## 1.2 Structure d'accueil

Mon stage se déroule au Service Statistique Ministériel (SSM) de la Justice depuis le 15 avril 2024, renommé le Service de la Statistique, des Études et de la Recherche (SSER). Ce service, qui était auparavant une sous-direction appelée la Sous-Direction des Statistiques et des Etudes (SDSE), élabore des statistiques publiques sur les thématiques de la justice et les diffuse à des fins d'information générale, est dirigé par Pascal Chevalier et réparti sur deux sites : Nantes et Paris. Le SSER dépend du Secrétariat Général (SG) du ministère de la Justice, dirigé par Carine Chevrier, tandis que le ministère est sous la direction du Garde des Sceaux Éric Dupond-Moretti. Le SSER est organisé autour de quatre bureaux, trois missions et un projet d'entrepôt de données, et compte environ 70 agents.

Le service comprend trois missions :

- La Mission Recherche, destinée à accueillir des chercheurs, bien qu'elle ne compte actuellement aucun agent ni chercheur.
- La Mission Programmation et Moyens (MPM), qui gère les tâches des ressources humaines, de secrétariat, et le maquettage de certaines publications du SSER. Les 5 agents de cette mission sont basés à Nantes et Paris.

- La Mission Pilotage et Accompagnement (MPA), qui entretient des relations transversales avec les différents bureaux et coordonne les projets du service. Elle est composée de 4 agents et basée à Paris.

Le projet Entrepôt de Données Brutes Historisées (DOBHIS) a pour but de consolider, dans un même espace de stockage, les données issues des différentes sources du ministère de la Justice. Il est composé d'une seule personne et situé à Nantes.

Le service est aussi structuré en quatre bureaux :

- Le Bureau de l'Informatique Statistique (BIS), composé de 15 agents et basé à Nantes, se charge de mettre en œuvre et maintenir des chaînes informatiques pour la collecte et le traitement statistique des opérations du SSM (via la Section Développement Informatique), d'exploiter et maintenir l'infrastructure informatique statistique (via la Section Exploitation Infrastructure) et d'assister à l'outillage data (via le chef de projet outillage data).
- Le Bureau des Sources Civiles, Accès aux Droits et Enquêtes (BUSCADE), composé de 13 agents et situé à Nantes, a des missions pour la justice civile, et se divise en deux sections : la Section Enquête et la Section Sources Civiles.
- Le Bureau des Études, de l'Évaluation et de la Communication (BEEC), divisé en deux sections, la Section Études réalise les études du SSER, et la Section Diffusion, Réponse à la demande gère la diffusion externe et les réponses aux demandes. Ce bureau, composé de 15 agents, est situé à Paris.
- Le Bureau des Sources Pénales et Pénitentiaires (BSPP), où j'effectue mon stage, composé de 14 agents et basé à Nantes, produit des données et des indicateurs sur la justice pénale et le milieu carcéral.

Le BSPP élabore et met en œuvre le dispositif statistique permettant d'éclairer le débat public dans le domaine de la justice pénale. Au sein du SSER, le bureau des sources pénales et pénitentiaires conçoit et assure la production de données statistiques des domaines pénal et pénitentiaire, dirigé par Wali Rostam (chef de bureau) et son adjointe Maryline Laille. Le bureau se compose de 11 responsables de production statistique et d'une assistante statisticienne. Mon encadrante de stage, Camille Dorion, est responsable de production statistique et référente du sujet des exécutions et aménagements des peines, ainsi que l'une des membres du groupe de travail de cartographie qui est le sujet sur lequel est axé mon stage.

# Chapitre 2

## Fondamentaux cartographiques

### 2.1 Introduction aux fonds de cartes

Avant mon stage, je n'avais aucune connaissance de la cartographie. En intégrant l'entreprise, j'ai été rapidement introduit à la fabrication des cartes, une compétence essentielle pour maîtriser le processus de production. Cette initiation m'a permis de développer des compétences précieuses, validées par la mise en place d'ateliers de cartographie en R et par ma capacité à répondre aux demandes au sein des équipes.

Un fond de carte est une représentation graphique composé d'un ensemble d'éléments géographiques. Ces éléments géographiques doivent être sélectionnés avec soin en fonction du niveau de détail souhaité, du support de représentation (papier ou numérique), de l'objectif de la carte et du public cible. Il existe trois types d'éléments géographiques de base : le point, la ligne et le polygone. Ces éléments peuvent être créés selon deux modes : vectoriel ou matriciel (raster). Ils sont liés entre eux par des relations spatiales (proximité, éloignement, contiguïté), ainsi que par leur taille et leur forme. Ces éléments peuvent être géoréférencés et stockés de manière cohérente dans un système d'information géographique (SIG). Ils forment des couches qui peuvent être superposées, croisées et utilisées pour construire un fond de carte.

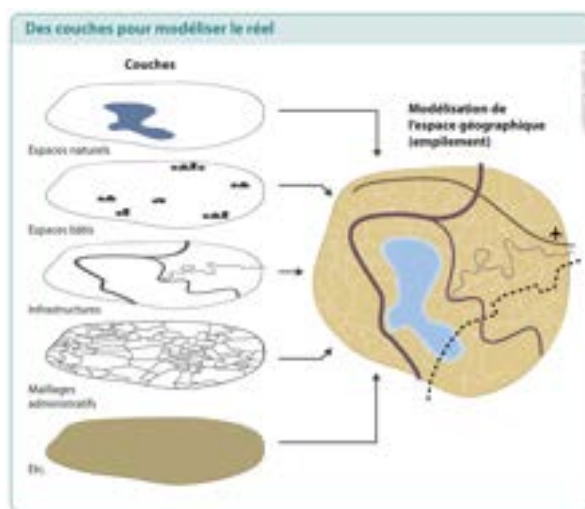


FIGURE 2.1 – Exemple de supersitions de couches [1].

Le fond de carte sert de référence pour la superposition d'autres informations géographiques ou thématiques. Ces éléments forment une base sur laquelle des données supplémentaires peuvent être ajoutées pour créer des cartes plus détaillées et spécifiques [2]. Ils sont essentiels pour contextualiser les données superposées, offrant un cadre spatial qui aide à situer et à comprendre les informations représentées. Ils sont utilisés dans divers domaines comme la géographie, l'urbanisme, l'environnement et les systèmes d'information géographique.

Cependant, ces fonds offrent une base commune qui garantit la cohérence et la comparabilité des données cartographiques, ce qui est crucial pour les analyses temporelles et spatiales. Ils rendent aussi les cartes plus accessibles et utilisables pour un large éventail d'utilisateurs, des professionnels aux amateurs et simplifient la lecture des cartes en fournissant des repères visuels familiers. En guise d'exemple, dans les applications de navigation GPS, les fonds de carte permettent aux utilisateurs de se situer rapidement et de comprendre leur environnement immédiat.

En outre, les fonds de carte de qualité, souvent créés par des institutions reconnues, garantissent une représentation précise et fiable des éléments géographiques de base, ce qui renforce la confiance des utilisateurs dans les cartes produites et l'Institut Géographique National (IGN) en France produit des fonds de carte largement utilisés pour leur précision et leur fiabilité.

Afin de mieux comprendre les différentes utilisations et applications des fonds de carte, il est important de maîtriser trois éléments indispensables, à savoir leur différents types, les projecteurs et les maillages.

## 2.2 Types de fonds de cartes

### 2.2.1 Cartes matricielles ou raster

Les fonds de cartes raster sont composés de pixels ou de cellules, chaque pixel représentant une valeur ou une couleur spécifique des éléments géographiques. Les données en mode raster sont issues d'un géocodage de plans scannés, d'images satellitaires. Cette structure permet de capturer des détails très fins et des textures complexes, ce qui en fait un choix idéal pour les images aériennes et les photographies satellites[3]. La qualité de l'image dépend de la résolution, c'est-à-dire du nombre de pixels par unité de surface. Plus la résolution est élevée, plus l'image est détaillée, mais cela entraîne également des fichiers plus volumineux. Les modifications apportées aux cartes raster peuvent entraîner une perte de qualité, surtout lors de l'agrandissement, car les pixels sont simplement redimensionnés sans ajout de nouvelles informations. Les fonds de cartes raster sont particulièrement adaptés aux représentations détaillées et les images. De plus, ils sont efficaces pour l'analyse de la couverture terrestre et des changements environnementaux, grâce à leur capacité à capturer des détails fins et des variations subtiles et sont utilisés, par exemple, pour les prévisions météorologiques.

Les cartes topographiques, qui généralement ont un fond de carte raster, sont une catégorie de cartes et sont des outils indispensables pour représenter les caractéristiques physiques et naturelles de la surface terrestre. Elles offrent une vue détaillée des reliefs, des élévations, des cours d'eau, des forêts, et de nombreux autres éléments naturels. Ces

cartes sont particulièrement utiles pour la navigation, la planification urbaine, l'ingénierie, et les activités de plein air comme la randonnée.

Cependant, en raison des analyses spatiales spécifiques nécessaires à notre travail, nous n'avons pas eu recours aux cartes topographiques. En effet, notre travail nécessite des outils qui permettent de visualiser et d'analyser des données spécifiques et thématiques. C'est pourquoi nous nous intéressons davantage à l'autre catégorie de carte qui sont les cartes thématiques qu'on détaillera plus bas.

## 2.2.2 Cartes vectorielles

Pour les fonds de cartes vectoriels, les éléments géographiques quant à eux, c'est à dire les points, lignes et polygones, sont définis par des coordonnées géographiques. Cette structure permet de représenter des éléments géographiques de manière précise et flexible. La qualité de l'image reste constante, quel que soit le niveau de zoom, car les éléments sont redessinés à chaque affichage. Cela signifie que les cartes vectorielles peuvent être agrandies ou réduites sans perte de qualité, ce qui est un avantage majeur par rapport aux cartes raster. De plus, les fichiers vectoriels sont généralement plus légers que les fichiers raster, ce qui facilite leur stockage et leur manipulation. Les fonds de cartes vectoriels sont faciles à modifier et à mettre à jour sans perte de qualité, ce qui en fait un choix idéal pour les représentations précises et les analyses spatiales. Ils sont parfaits pour la cartographie urbaine, où les réseaux routiers, les bâtiments et les infrastructures doivent être représentés avec précision.

Ils sont également couramment utilisés pour les cartes thématiques, telles que les cartes de population, de répartition des ressources, etc. Les fonds de cartes vectoriels sont également efficaces pour les analyses de réseaux, les calculs de distances et les études de densité, grâce à leur capacité à représenter des éléments géographiques de manière précise et flexible.

La figure (2.2) résume de manière visuelle les définitions récentes et les différences clés entre les deux types de fonds de cartes. Il vous aidera à visualiser les avantages et les inconvénients de chaque méthode.

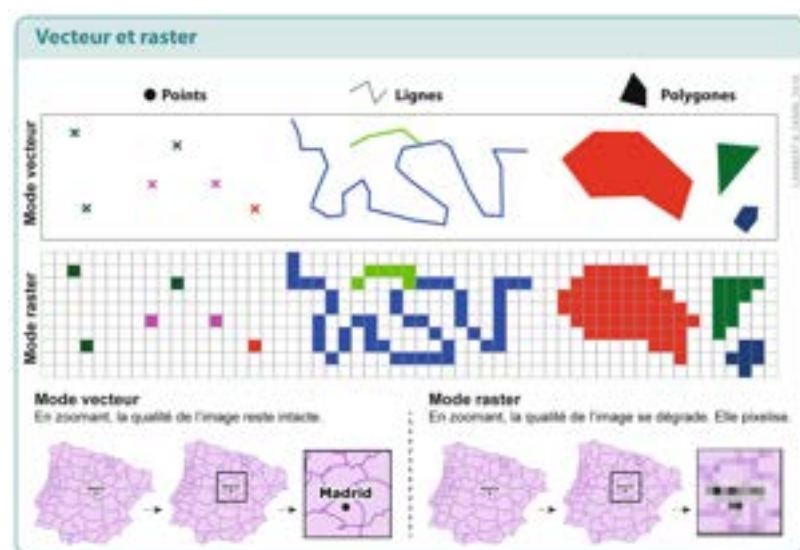


FIGURE 2.2 – Différences entre les fonds de cartes raster et vectoriels [1].

## 2.3 Les projecteurs

En cartographie, les systèmes de projection sont essentiels car ils permettent de représenter une surface initialement courbe, comme celle de la terre, sur une surface plane, celle de la carte. Ces projections sont essentielles pour minimiser les distorsions et garantir une représentation précise des distances, des angles et des superficies. Parmi les systèmes de projection les plus couramment utilisés, deux se distinguent particulièrement : le Lambert 93 et le World Geodetic System, créé en 1984 (WGS84).

Le Lambert 93 est une projection conique conforme, largement utilisée en France hexagonale . Cette projection est particulièrement adaptée aux cartes à grande échelle, car elle préserve les angles et les formes des objets géographiques. Elle est souvent utilisée pour les cartes topographiques, offrant une précision accrue aux applications locales et régionales. Le Lambert 93 est également apprécié pour sa capacité à représenter fidèlement les distances et les superficies sur des zones géographiques spécifiques. De plus, les cartes fournies par l'IGN en France, sont issues de ce projecteur dont l'identifiant European Petroleum Survey Group (EPSG) est 2154 .

Le WGS84 ou projecteur de Mercator avec pour EPSG 4326, quant à lui, est un système de coordonnées géographiques mondial qui sert de référence pour les systèmes de positionnement par satellite (GPS). Il est utilisé pour les cartes à petite échelle et pour les applications nécessitant une précision géodésique. Le WGS84 comme son nom l'indique, est particulièrement important pour les applications globales, telles que la navigation, la cartographie mondiale et les systèmes d'information géographique. Ce système de projection permet une représentation cohérente et précise des coordonnées géographiques à l'échelle mondiale. C'est ce projecteur que nous utiliserons quand il s'agira de manipuler les fonds de cartes de la France entière.

Cependant, il est important de noter qu'aucune projection cartographique n'est parfaite et que toutes les cartes représentées comportent des déformations par rapport à la réalité.



FIGURE 2.3 – Déformations observées avec le WGS84

Il est important d'apporter une attention particulière aux projecteurs, car ils jouent un rôle essentiel dans la précision et la fiabilité des cartes. Par exemple, superposer des cartes qui n'ont pas le même système de projection peut entraîner des erreurs significatives, un problème auquel j'ai été confronté lors de la création de fonds de carte.

## 2.4 Catégories de cartes

### 2.4.1 Cartes topographiques

Les symboles utilisés sur les cartes topographiques sont standardisés et permettent de représenter une variété d'éléments naturels et artificiels. Par exemple, les rivières, les lacs, les forêts, et les montagnes sont souvent représentés par des symboles spécifiques. Ces symboles sont comme un langage universel qui vous aide à comprendre rapidement ce que vous allez rencontrer sur votre chemin.

Bien que les cartes topographiques soient des outils précieux pour représenter les caractéristiques physiques et naturelles de la surface terrestre, elles ne répondent pas entièrement à nos besoins d'analyse statistique.

### 2.4.2 Cartes thématiques

Les cartes thématiques, quant à elles, se concentrent sur des sujets spécifiques ou des thèmes particuliers, comme par exemple la démographie. Elles sont conçues pour visualiser et analyser des données spécifiques, souvent pour des études, des rapports, ou des présentations. Dans la catégorie des cartes thématiques, il existe plusieurs types de cartes adaptées en fonction de ce que l'on veut mettre en valeur. Bien qu'il y en ait un certain nombre, nous nous contenterons d'aborder celles que j'ai eu l'occasion de côtoyer pendant mon stage. Il s'agit des cartes choroplèthes, des cartes à ronds proportionnels et des cartes carroyées.

Lors de mon stage, j'ai eu l'opportunité d'animer un atelier sur la cartographie en utilisant le langage de programmation R, et des données issues des infocentres MINOS et CASSIOPEE. L'atelier s'est avéré très enrichissant et a permis aux participants de comprendre l'importance de la cartographie pour la prise de décision et l'identification des tendances et des disparités géographiques. Cet atelier a été conçu pour initier les participants aux techniques de visualisation de données géographiques et avait pour objectif de leur montrer comment importer, manipuler et traiter des fonds de cartes avec la librairie **sf**, réaliser des cartes basiques statiques choroplèthe et cercles proportionnels avec **mapsf** qui est une librairie R française développée par l'Insee, et enfin un carroyage. Parmi les travaux réalisés, j'ai été amené à créer chacune de ces cartes que je vous présenterai en guise d'illustration. Dans l'idée, nous sommes partis des exemples plus simples comme les choroplèthes, puis au fur et à mesure nous avons appréhender les plus complexes, c'est à dire les cartes carroyées.

#### 2.4.2.1 Cartes choroplèthes

Les cartes choroplèthes sont des outils de cartographie qui permettent de visualiser des données statistiques spatiales de manière claire et compréhensible. Elles sont particulièrement efficaces pour montrer la distribution géographique de variables quantitatives. Pour

ce faire, elles utilisent des couleurs ou des teintes pour représenter les différentes valeurs de la variable étudiée dans diverses zones géographiques. Une légende ou une échelle de couleurs est généralement incluse pour aider à interpréter les informations. Les données sont souvent réparties en plusieurs catégories pour simplifier la lecture. Ces cartes sont très utiles pour l'analyse spatiale, la communication de données complexes et la prise de décision, car elles permettent de repérer rapidement les tendances et les disparités géographiques. Les cartes choroplèthes servent surtout à illustrer les variables quantitatives relatives, telles que les taux et les proportions, et ces types de données sont fréquemment utilisés en cartographie statistique.

La figure (2.4) ci dessous, met en évidence le nombre total des excès de vitesse toutes catégories confondues jugés par les tribunaux judiciaires (TJ) dépendant de la cour d'appel (CA) de Rennes entre 2019 et 2023. En effet, nous remarquons que cette carte utilise un nombre de classes (breaks) pour définir un seuil d'appartenance moyen . De plus, sans avoir à disposition les données, cette carte permet de visualiser de manière facile que les automobilistes commettent moins d'infractions sur le réseau routier nantais , tandis qu'à Saint-Malo ou Saint-Brieuc où ils commettent en moyenne plus d'excès de vitesse.

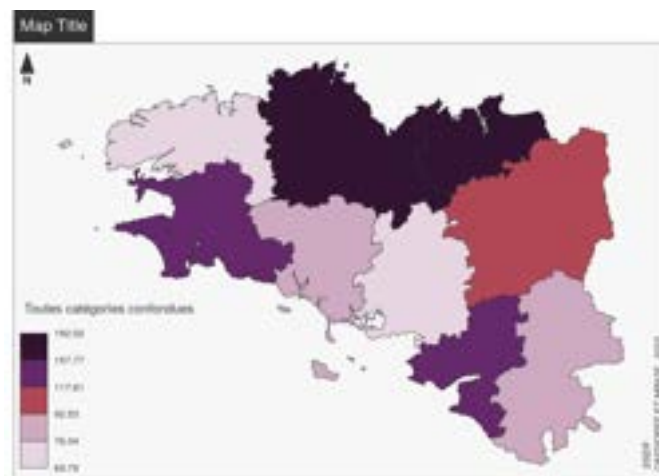


FIGURE 2.4 – Nombre d'infractions confondues pour 100k habitants jugés dans la CA de Rennes

#### 2.4.2.2 Cartes de proportions

Ce type de carte thématique est utilisé pour représenter des données quantitatives absolues en ajustant la taille des symboles ou des figures en fonction des valeurs des données. Contrairement aux cartes choroplèthes, qui utilisent des couleurs, les cartes proportionnelles se servent de la taille des symboles pour indiquer les variations quantitatives . Ces cartes utilisent des symboles, comme des cercles, des carrés ou des icônes, dont la taille varie en fonction des valeurs des données.



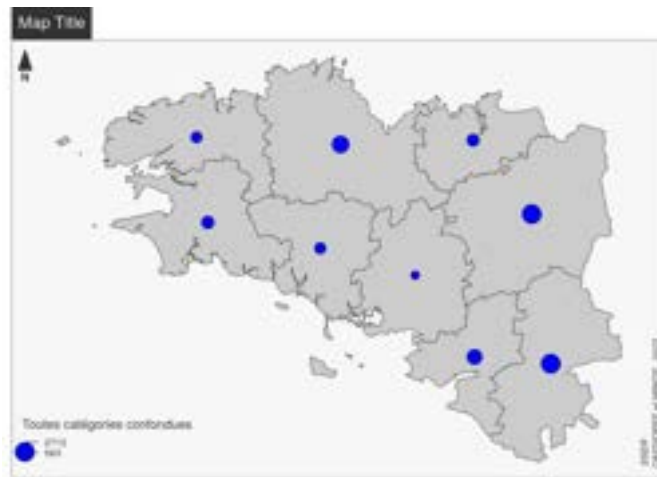
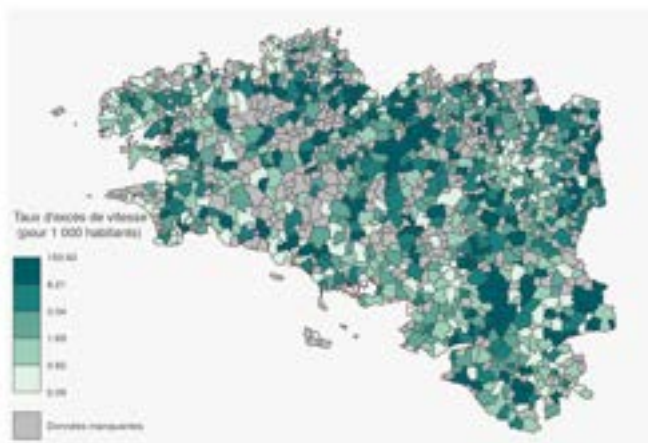


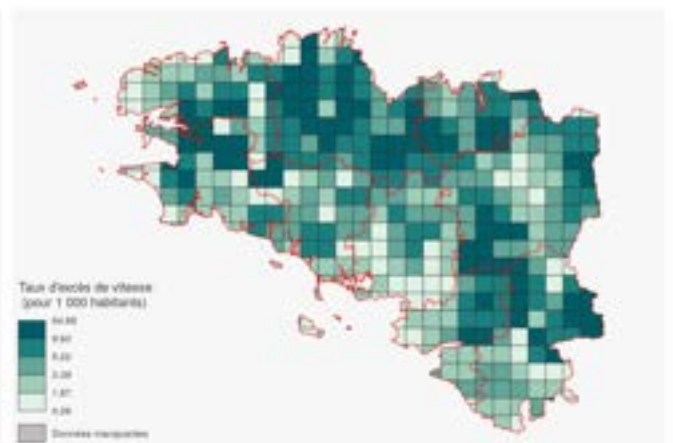
FIGURE 2.5 – Proportion d’infractions confondues jugées dans la CA de Rennes

### 2.4.2.3 Cartes carroyées

La méthode du carroyage consiste à segmenter l’espace géographique en une grille composée de carrés uniformes dans une projection donnée. Cette technique permet de se libérer de l’arbitraire et de l’irrégularité d’un découpage administratif. Elle révèle les grandes tendances de la distribution spatiale d’une donnée en divisant le territoire en carreaux égaux et localisés. La donnée est ensuite répartie sur ce quadrillage régulier, en proportion de la surface représentée. Chaque carreau est coloré ou rempli en fonction de la valeur de la variable quantitative dans cette zone. Cependant, ce type de carte peut s’avérer un peu plus compliqué à mettre en évidence, car il nécessite une préparation minutieuse des données et une attention particulière sur le maillage à mettre en valeur.



(a) Carte initiale de la CA de Rennes



(b) Carte de la CA de Rennes carroyée

FIGURE 2.6 – Comparaison des taux d’excès de vitesse par commune à Rennes

### 2.4.3 Types de visualisations

Toutes les types de cartes qui ont été présentées peuvent être visualisées selon ce que l’on veut mettre en évidence. Il existe deux types de visualisation : la visualisation statique

et la visualisation dynamique. La visualisation statique consiste à représenter les données de manière fixe, souvent sous forme de cartes imprimées ou d'images numériques. Elle est idéale pour des présentations claires et détaillées, mais ne permet pas d'interagir avec les données. En revanche, la visualisation dynamique permet aux utilisateurs d'interagir avec les données en temps réel, en zoomant, en déplaçant la carte ou en filtrant les informations. Cette approche est particulièrement utile pour l'exploration et l'analyse des données, offrant une flexibilité et une profondeur d'information supplémentaires. Chaque type de visualisation a ses avantages et peut être choisi en fonction des besoins spécifiques de l'utilisateur et des objectifs de la présentation.

Cependant, il est important de noter que l'utilisation de ces visualisations peut entraîner la création de fichiers volumineux, ce qui peut poser des problèmes de performance et de stockage. L'un des objectifs de ce stage est de maîtriser les bonnes pratiques pour optimiser ces visualisations sans sacrifier la qualité de l'analyse. Cela inclut des techniques telles que la réduction de la taille des données, l'utilisation de formats de fichiers plus efficaces, et l'optimisation du code pour minimiser la charge de traitement. En adoptant ces pratiques, nous pouvons créer des visualisations dynamiques qui sont non seulement informatives, mais aussi légères et performantes.

Pour faire du dynamique avec R, nous utilisons couramment les packages comme `leaflet`. Cette librairie nécessite exclusivement d'avoir comme projecteur le WGS84 car ce dernier est le projecteur référence mondiale et donc les fonctions de `leaflet` ne répondent qu'à ce projecteur, sinon aucune génération de fond de carte serait possible.



FIGURE 2.7 – Aperçu d'un fond de carte généré avec `leaflet`

# Chapitre 3

## Création de fonds de cartes judiciaires et pénitentiaires

Dans ce chapitre, je présenterai les fonds de carte que j'ai créés. Ces cartes couvrent une variété de sujets et de complexités, allant des représentations simples des tribunaux judiciaires et cours d'appel aux cartes plus complexes, les ressorts pénitentiaires. Chaque fond de carte a été conçu pour répondre à des besoins spécifiques et pour offrir une visualisation claire et précise des données géospatiales pertinentes.

### 3.1 Présentation de GESTSR

La GESTion du Système de Référence, dont l'acronyme est GESTSR, met à disposition les tables de données provenant du Système de Référence Judiciaire. Ce dernier regroupe l'ensemble des nomenclatures utilisées dans les différents applicatifs, qu'il s'agisse des catégories d'infractions, de la classification des mesures ou, dans notre cas, de la définition de l'ensemble des entités juridiques.

On y trouve les éléments de structure qui décrivent les acteurs concernés par l'exercice de la justice comme les établissements pénitentiaires, les tribunaux judiciaires(TJ), les cours d'appel (CA), les directions interrégionales des services pénitentiaires (DISP) et les services pénitentiaires d'insertion et de probation (SPIP) ainsi que les natures d'infractions, peines et mesures (les sanctions prises par l'autorité judiciaire ) et les événements (les jugements, enquêtes, etc...).

Chaque domaine est décomposé en plusieurs Types d'Eléments de Structure (TES), non domaines. De la même manière, chaque Type d'Eléments de Structure peut être décomposé en sous-types. Ce type de lien entre TES constitue une relation hiérarchique, par exemple, le « Type » TJ est le fils du « Type » CA . On pourra considérer qu'un élément de structure est une occurrence ou une instance d'un type d'éléments de structure. Un élément de structure peut être rattaché à un père et avoir des fils en cohérence avec les relations hiérarchiques établies entre leurs TES respectifs. La figure ci-dessous illustre le lien père-fils dans la DISP de Rennes .

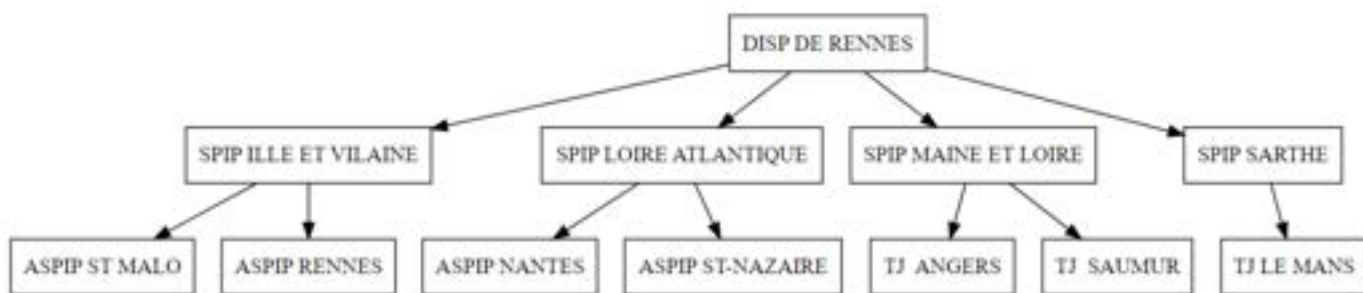


FIGURE 3.1 – Exemple de lien père-fils

Le ressort territorial d'un élément de structure est défini par son secteur géographique. Un secteur géographique peut être composé d'un ensemble de communes et/ou d'autres secteurs géographiques. Dans ce dernier cas, on parlera de regroupement de secteurs géographiques.

Dans GESTSR, il existe de nombreuses bases de données. Cependant, dans le cadre de notre étude, nous nous servirons de certaines d'entre elles, à savoir :

- Éléments de Structure (**ELE\_STR**) : Cette table regroupe tous les éléments de structure, c'est à dire toutes les structures organisées, interlocuteurs ou acteurs en relation avec l'activité du Ministère de la Justice.
- Secteurs Géographiques (**SEC\_GEO**) : Le ressort de compétence des juridictions judiciaires est fixé par décret. Le secteur géographique représente le ressort de compétence d'une juridiction judiciaire. Les secteurs géographiques sont composés de communes réelles INSEE ou (exclusif) d'autres secteurs géographiques. Une occurrence de secteurs géographiques est identifiée par un numéro (NU\_SGE). Le secteur géographique d'un TJ est composé de 1 à n communes. Cette information est conservée dans la table Regroupement secteurs géographiques .
- Regroupement Secteurs Géographiques (**REG\_SEC\_GEO**) : Un lien entre un Secteur Géographique Fils et un Secteur Géographique Père s'identifie par le trinôme Numéro de Secteur Géographique Fils (NU\_SGE\_FIL), un Numéro de Secteur géographique Père (NU\_SGE\_PER) et une Date d'Application du lien.
- Compétences Territoriales (**COM\_TER**) : Cette table regroupe les compétences territoriales, c'est-à-dire les liens entre les Éléments de Structure et les Secteurs Géographiques. Un élément de structure exerce ses compétences sur un territoire donné.
- Communes Secteurs géographiques (**COM\_SEC**) : Cette table contient les liens entre les communes et les secteurs géographiques. Cette table assure aussi le rôle d'historisation des liens.
- Commune : Cette table regroupe toutes les communes réelles et fictives, inscrites au fichier du CODE OFFICIEL GEOGRAPHIQUE de l'INSEE.

## 3.2 Fonds de carte de base des communes

La création de fonds de carte comprend plusieurs étapes. Dans ce rapport, nous appellerons France entière la France incluant les Départements et Régions d'Outre-Mer (DROM) c'est à dire la Guyane, Martinique, Guadeloupe, Mayotte et la Réunion , mais excluant les

Collectivités d’Outre-Mer (COM) c’est à dire la Nouvelle-Calédonie, Polynésie Française, Saint-Barthélemy, Saint-Pierre-et-Miquelon et par France hexagonale, nous entendrons la France sans les DROM et les COM mais comprenant la Corse.

Tout d’abord, il est nécessaire de disposer d’un fond de carte de base des communes de la France entière, qui fournit les géométries sous forme de vecteur multidimensionnel de toutes les communes du pays. Pour se le procurer, l’INSEE met à disposition en interne, sur le site CreaCartes[4], des fonds de cartes des deux France précitées. Ces fonds incluent des délimitations au niveau des communes, départements, régions, ainsi que des cartes des deux Frances sans aucune délimitation . Le choix du fond de base dépend de l’objectif à atteindre.

Les fichiers mis à disposition par l’INSEE sont des fichiers GeoPackage. Le GeoPackage est un format de fichier ouvert qui permet de stocker des données géospatiales vectorielles et raster, ainsi que des métadonnées associées. Pour obtenir les géométries des communes, il suffit de charger le fichier GeoPackage dans R avec la fonction **st\_read** du package **sf**, et d’extraire les données nécessaires.

D’une manière plus générale, en partant du fond de carte des communes de la France entière, il est possible de remonter aux autres délimitations et de passer à la France hexagonale . En effet, le fichier comprend, en plus des géométries des communes, d’autres variables très pertinentes comme les codes de communes , des départements et des régions qui permettent de passer du niveau communes aux niveaux supérieurs. Par exemple, pour passer du niveau communes au niveau département, il suffit de filtrer les communes, de les regrouper par numéros des départements, puis de faire une union des géométries.

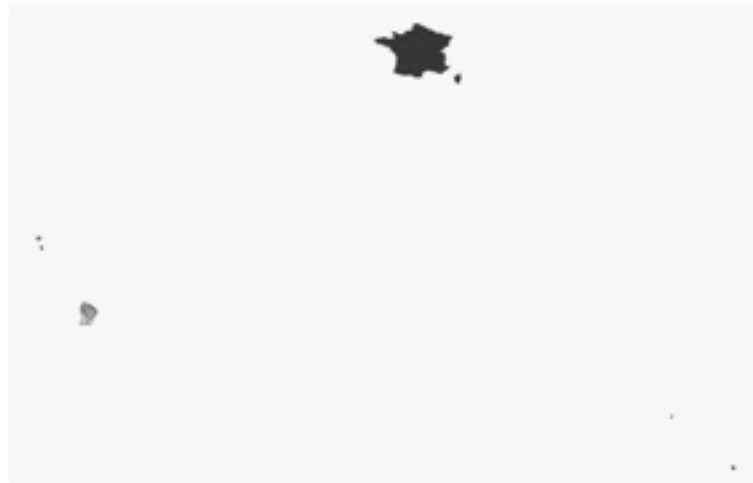


FIGURE 3.2 – Fond de carte initial des communes de la France entière de l’Insee

La figure (3.2) ci-dessus représente le fond de carte initial des communes de la France entière dont nous disposons. En examinant ce fond de carte, on s’aperçoit qu’il présente de nombreuses imperfections. En effet, il n’est pas rapproché, et les DROM ne sont pas bien intégrés à la France hexagonale, ce qui limite son utilité pour certaines analyses spatiales. Par contre, avec le fond de carte de la France hexagonale, l’unique problème sera le non rapprochement du fond de carte car il ne comprend pas les DROM.

Pour pallier ce problème, nous avons décidé d’effectuer des retouches spatiales afin de

rapprocher le fond de carte et de mieux intégrer les DROM. Pour ce faire, à l'aide d'une notice de l'Insee , nous avons procédé à un traitement particulier pour chaque région comme suit :

- Pour la Corse, on filtre le code Insee de la région 94, puis on commence par modifier son projecteur en Lambert 93 qui correspond au numéro EPSG 2154, puis aux géométries on applique une translation par le vecteur de coordonnées  $(-166400,0)$ . Les coordonnées sont ensuite réinitialisées et définies dans le système de coordonnées EPSG : 4326 c'est à dire WGS84. Tandis que les régions des DROM ont chacune un projecteur spécifique qu'on initialise avant de faire les mêmes opérations.
- Partant de ce principe, pour la Guadeloupe et la Martinique ayant le même projecteur spécifique, les coordonnées sont transformées en utilisant le système de coordonnées EPSG : 5490 , sont ajustées respectivement en faisant des homothéties de rapport 1.32 et 1.85 et en ajoutant un vecteur de translation de coordonnées  $(-699983, 4269050)$  et  $(-1134525, 3517169)$  puis réinitialisées et définies dans le système de coordonnées WGS84 .
- En ce qui concerne la Guyane et la Réunion, avant de rendre les coordonnées en WGS84, les projecteurs spécifiques sont initialisés qui sont respectivement les EPSG : 2972 et 2975. Cette fois les rapports des homothéties sont respectivement 0.25 et 1.75 et les vecteurs des translations ont pour coordonnées  $(118687, 6286270)$  et  $(-422169, -7132230)$ .
- Pour Mayotte, ils sont initialisés en utilisant le système de coordonnées EPSG : 4471, puis les paramètres de la similitude sont respectivement 2.45 pour l'homothétie et  $(-1082300, -14868225)$  comme coordonnées pour le vecteur de translation.
- Enfin, pour rapprocher le fonds de carte définitif, on filtre toute la France hexagonale et l'initialisant en Lambert 93 , puis on lui effectue comme opération la similitude de paramètres 2.78 et  $(-1634506, -12046235)$  avant de rendre le projecteur en WGS84 comme projecteur final .

Après avoir effectué toutes ces opérations, on obtient comme résultat le fond de carte suivant .

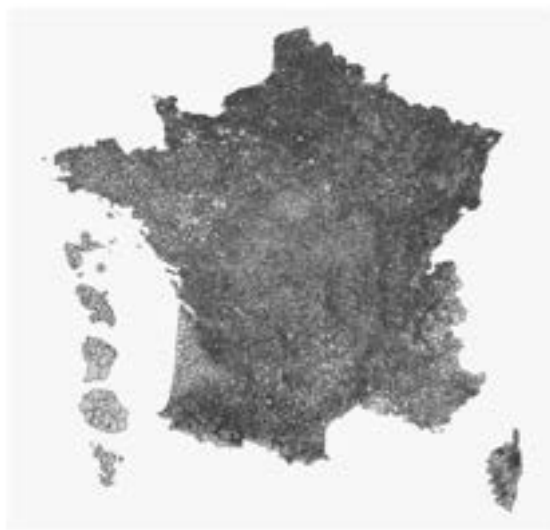


FIGURE 3.3 – Fond de carte initial après rapprochement

C'est ce fond de carte obtenu qui nous servira comme fond de base pour la suite de nos travaux.

### 3.3 Fonds de cartes des TJ et CA

Dans notre service, la mise à disposition de fonds de carte des Tribunaux Judiciaires et des Cours d'Appel est une étape essentielle pour plusieurs raisons. Ces cartes nous permettent de visualiser clairement la répartition géographique des instances judiciaires, ce qui est crucial pour comprendre et analyser les dynamiques juridiques et administratives. Les TJ, par exemple, jouent un rôle central dans les affaires pénitentiaires, en gérant les procédures judiciaires et en garantissant l'application équitable de la loi. De même, les CA sont indispensables pour traiter les appels et superviser les décisions des tribunaux de première instance, assurant ainsi la cohérence et la justice dans notre système judiciaire.

En disposant de ces fonds de carte, nous pouvons améliorer la planification et l'organisation des ressources judiciaires, et identifier les zones nécessitant une attention particulière en matière de sécurité publique et de justice. Ces cartes nous aident également à mieux comprendre les besoins en infrastructures judiciaires et à planifier les investissements nécessaires pour améliorer l'efficacité et l'accessibilité des services judiciaires. De plus, elles constituent un outil précieux pour les études et les recherches, permettant aux analystes et aux décideurs de visualiser les tendances et les patterns géographiques liés aux affaires judiciaires, c'est à partir de cette idée que le choix de générer ces fonds de carte a été fait.

Avant de commencer, il est important de comprendre comment les tribunaux sont répartis en France et comment ils sont organisés géographiquement. En France, chaque ville ou commune dépend d'un TJ bien défini. Ces tribunaux sont responsables de traiter toutes les affaires judiciaires liées à leur juridiction. Cette organisation permet de garantir que chaque citoyen a accès à un tribunal compétent pour traiter ses affaires, qu'elles soient civiles ou pénales. Il existe un ordre administratif comportant des tribunaux administratifs, des cours administratives d'appel et le Conseil d'Etat à côté de l'ordre judiciaire.

En France entière, nous disposons de 164 TJ. Ces tribunaux sont eux-mêmes regroupés en ressorts, qui forment les 34 CA de la France entière. Chaque CA supervise plusieurs TJ et est responsable de traiter les appels des décisions rendues par ces tribunaux. Cette structure hiérarchique assure une cohérence et une uniformité dans l'application de la justice à travers le pays.

Pour générer le fonds de carte des TJ, nous avons suivi un processus détaillé de manipulation et de traitement des données des ressorts géographiques. Tout d'abord, nous avons chargé les données des communes et de leur rattachement aux TJ à partir d'une base de données GESTSR qui fournit les ressorts historisés des juridictions au niveau commune, suivant le type de juridiction. Ensuite, nous avons effectué plusieurs corrections pour rectifier des erreurs spécifiques dans les codes communes INSEE et les dates de fermeture des communes. Par exemple, nous avons corrigé le code commune INSEE pour une commune spécifique (Les Trois Lacs) qui était erroné depuis 2021 [5] et supprimé les dates de fermeture pour certaines communes comme Haut Valromney qui était considéré comme fermée ainsi que Lyon et Marseille qui avaient des problèmes sur les dates des secteurs géographiques.

Après ces corrections, nous avons filtré les données pour ne conserver que les communes dont la date d'application était antérieure ou égale au 01-01-2024, et dont les dates de fermeture étaient soit nulles, soit postérieures à cette date.

Nous avons ensuite fusionné ces données géographiques avec le fond de carte des communes France entière figure (3.3) pour obtenir une vue complète et précise de la répartition géographique des TJ. Pour vérifier la qualité des données, nous avons effectué un contrôle d'identification des codes des communes dont la géométrie était manquante.

Enfin, nous avons regroupé les données par les codes d'établissement des TJ ainsi que leur code CA auquel ils sont rattachés et fusionné les géométries correspondantes à l'aide de la fonction `st_union` de *mapsf* pour obtenir une représentation claire et unifiée des zones géographiques associées aux TJ.



FIGURE 3.4 – Fond de carte des TJ en France entière

Après avoir obtenu le fond de carte de la figure (3.4) , pour passer des TJ aux CA, il nous a fallu regrouper les TJ en fonction des codes d'établissement des CA auxquels ils sont rattachés. Ensuite, il a été nécessaire de refaire une union des géométries des TJ en utilisant la même fonction mentionnée précédemment.



FIGURE 3.5 – Fond de carte des CA en France entière



Ces fonds de carte ont été relativement facilités car, dans GESTSR, les données sur les ressorts des TJ et des CA sont suffisamment bien prédisposées, ce qui nous a permis d'effectuer ces travaux avec une telle précision, ainsi que des redressements et quelques contrôles de qualité.

## 3.4 Cartes des ressorts pénitentiaires

Les ressorts pénitentiaires en France sont des zones géographiques où sont organisés et gérés les directions interrégionales des services pénitentiaires, les services pénitentiaires d'insertion et de probation, et leurs antennes.

Parler du monde pénitentiaire, c'est s'intéresser aux ressorts de ces structures pénitentiaires, chacune jouant un rôle crucial dans la gestion et la réinsertion des personnes placées sous main de justice.

Les DISP supervisent et coordonnent les actions des SPIP et des établissements pénitentiaires à l'échelle régionale. Elles assurent une gestion cohérente et efficace des politiques de réinsertion et de probation, veillant à ce que chaque individu reçoive le soutien nécessaire pour se réinsérer dans la société.

Les SPIP, quant à eux, sont en première ligne pour accompagner les détenus et les personnes placées sous main de justice. Ils mettent en œuvre des mesures de probation et préparent la sortie de prison, offrant un soutien personnalisé et adapté aux besoins de chacun.

Les antennes des SPIP, implantées localement, permettent une décentralisation des services, rendant l'accompagnement plus accessible et adapté aux besoins spécifiques des individus. Elles facilitent le suivi des mesures judiciaires et l'accompagnement des détenus, en étant présentes là où elles sont le plus nécessaires.

Disposer de fonds de carte précis et détaillés des ressorts pénitentiaires est un atout inestimable pour le ministère de la Justice, ces fonds de cartes permettront de suivre, à travers une représentation visuelle, la répartition des ressorts pénitentiaires sur le territoire français.

Contrairement aux fonds de cartes des TJ et CA, la nomenclature des ressorts pénitentiaires dans GESTSR indiquait bien les liens entre les communes et leur secteur de rattachement au niveau des DISP et SPIP. Mais, les tables de cette base n'étaient pas suffisamment bien structurées pour permettre d'effectuer les travaux nécessaires en ce qui concernait les ASPIP. Pour pallier cette lacune, il a fallu créer une nouvelle base de données issue de certains redressements, qui prenait en compte toutes les dépendances des communes de façon générale. Cette initiative a permis de structurer les informations de manière plus cohérente et de faciliter la réalisation des analyses et des représentations cartographiques nécessaires.

### 3.4.1 Liens entre les ressorts et les communes

Chaque commune dépend d'un TJ, et chaque TJ est rattaché à une antenne SPIP, si elle existe. À son tour, chaque antenne SPIP est liée à un SPIP central, et enfin, chaque SPIP est rattaché à une DISP. C'est ce lien hiérarchique et organisationnel que nous voulons

mettre en évidence . Pour y parvenir , nous avons essayé de ranger les hiérarchies de un à quatre en fonction de la liaison mise en évidence .

Dans un premier temps, nous avons d'abord filtré dans la table des secteurs géographiques (**SEC\_GEO**) pour ne conserver que ceux dont l'identifiant du type de secteur correspond aux DISP. Puis, nous avons effectué une jointure entre les DISP filtrées et la table des regroupements des secteurs géographiques (**REG\_SEC\_GEO**), en utilisant l'identifiant **NU\_SGE** comme clé de jointure. Nous avons filtré les résultats pour ne conserver que les secteurs qui n'ont pas de date de fermeture. Ensuite, nous avons sélectionné les colonnes nécessaires pour notre analyse, incluant les identifiants des secteurs, les dates d'application et de fermeture. Nous avons effectué une nouvelle jointure avec la table des secteurs géographiques **SEC\_GEO** pour obtenir les noms des secteurs fils c'est à dire les SPIP qui sont rattachés à chaque DISP. Enfin, nous avons ajouté une colonne **RANG** avec la valeur 1 pour indiquer le premier rang de passage des DISP aux SPIP, et nous avons dupliqué la colonne **NU\_SGE** en **NU\_SGE\_PER** afin de connaître les numéros des sièges des rattachements des SPIP aux DISP pour faciliter les analyses ultérieures. On appellera cette base de données **REG\_SEC\_PREM\_RANG** qui nous donne les informations sur tous les numéros des sièges SPIP (**NU\_SGE\_FILS**) ainsi que leur DISP de rattachement (**NU\_SGE\_PER**).

Pour lier les SPIP aux ASPIP, nous avons effectué une démarche similaire dans l'ensemble, nous avons d'abord sélectionné et renommé les colonnes nécessaires dans la table **reg\_sec\_prem\_rang**, qui contient les informations sur les SPIP et leurs DISP de rattachement. Nous avons renommé **NU\_SGE\_FILS** en **NU\_SGE\_SPIP** pour indiquer qu'il s'agit des identifiants des SPIP. Nous avons également ajouté une nouvelle colonne **NU\_SGE\_PER** qui est une copie de **NU\_SGE\_SPIP** pour faciliter la jointure ultérieure.

Ensuite, nous avons effectué une jointure entre les données sélectionnées et la table **REG\_SEC\_GEO**, en utilisant **NU\_SGE\_PER** comme clé de jointure. Cela nous a permis d'obtenir les informations sur les secteurs géographiques (antennes) rattachés aux SPIP. Nous avons filtré les résultats pour ne conserver que les secteurs qui n'ont pas de date de fermeture et qui ont un secteur fils .

Nous avons ensuite effectué une nouvelle jointure avec la table **SEC\_GEO** pour obtenir les noms des secteurs fils (antennes) rattachés aux SPIP. La clé de jointure était **NU\_SGE\_FIL**. Enfin, nous avons ajouté une colonne **RANG** avec la valeur 2 pour indiquer le deuxième rang de passage des SPIP aux antennes.

Cette base de données résultante, **REG\_SEC\_DEUX\_RANG**, contient les informations sur tous les SPIP ainsi que leurs antennes de rattachement, facilitant ainsi les analyses ultérieures.

Pour lier les ASPIP aux TJ et les TJ aux instances plus basses, de façon itérative, nous sommes partis de la base de données précédente **REG\_SEC\_DEUX\_RANG** et avons procédé de la même manière pour son obtention, ce qui nous donne les bases de données **REG\_SEC\_TROIS\_RANG** pour la liaison des ASPIP aux TJ. Nous sommes répartis de cette dernière pour la liaison des TJ aux TI nommée **REG\_SEC\_QUATRE\_RANG** .Nous avons fait une union des quatres bases de données rangées qu'on appellera **REG\_SEC\_SM** .

Disposant de la hiérarchie totale des secteurs géographiques, nous voulons dans un second temps rattacher toutes les communes dépendantes de ces secteurs. Pour ce faire, avec la table **COM\_SEC** qui donne tous les secteurs géographiques rattachés à chaque commune, nous avons d'abord effectué une jointure entre **REG\_SEC\_SM** et la table **COM\_SEC** en uti-

lisant NU\_SGE\_FIL comme clé de jointure. Ensuite, nous avons effectué une nouvelle jointure par le code commune avec la table COMMUNE pour obtenir les noms des communes rattachées aux secteurs géographiques et enfin effectuer des filtres sur les dates d'application et de fermeture nommée COMMUNE\_ELE\_STR\_FILTRE. La figure ci-dessous résume les jointures clés effectuées.

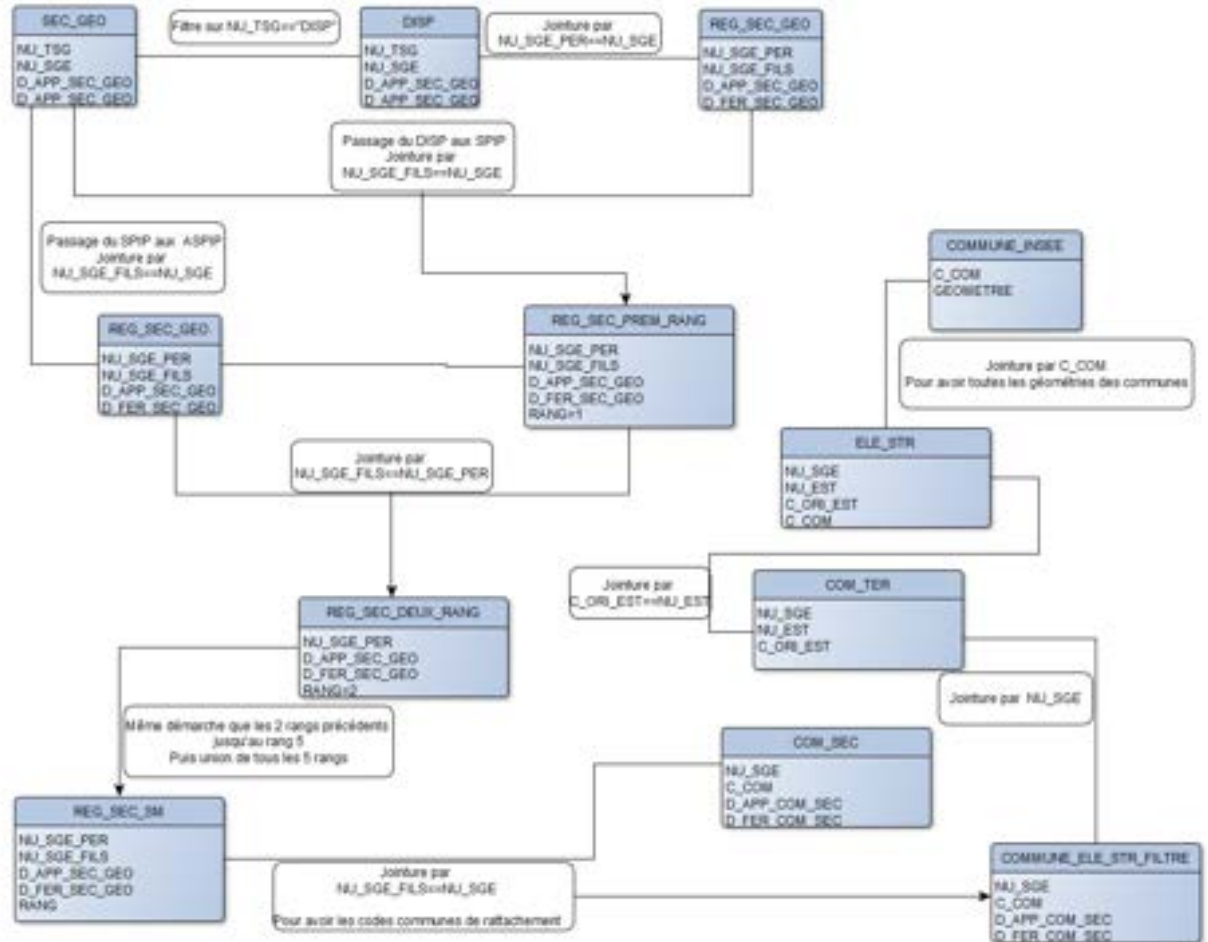


FIGURE 3.6 – Résumé des jointures clés effectuées

C'est cette base de données complète finale qui nous servira pour générer les fonds de cartes des ressorts.

### 3.4.2 Fonds de carte obtenus

Pour générer les fonds de carte à partir de la base de données finale obtenue, nous effectuons des jointures respectives avec les bases de données COM\_TER et ELE\_STR en utilisant simultanément les numéros de siège, le code d'origine de l'établissement et les numéros d'établissements. De plus, nous effectuons ces jointures en ajoutant le code de la commune rattachée, en faisant varier à chaque fois le type de ressort avant de finalement faire une jointure par code commune avec le fond de carte possédant toutes les géométries des communes et une union des géométries obtenues.



FIGURE 3.7 – Fond de carte des DISP en France hexagonale

La figure (3.7) illustre le fond de carte des DISP, à cette carte on ajoute la couche des SPIP (figure 3.8) .



FIGURE 3.8 – Fond de carte des SPIP en France hexagonale

Quant aux ASPIP dans la figure (3.9), nous avons mis en place un filtre permettant de faire une dichotomie du fait de la non présence des antennes car en réalité les antennes ne sont pas toujours présentes dans les SPIP. En effet, par exemple le SPIP de Nantes comprend les antennes de Nantes et celle de Saint Nazaire, tandis qu'au Mans par exemple, le SPIP n'en dispose pas .



FIGURE 3.9 – Fond de carte des ASPIP en France hexagonale

De plus près, en comparant les SPIP et ASPIP, nous constatons par exemple que l’absence d’ASPIP se lit facilement dans les villes du centre car les deux fonds de carte ont pas mal de ressemblance dans cette zone, mais pas que.

## 3.5 Géolocalisation des TJ et centres pénitentiaires

Après avoir créé le fond de carte des TJ, nous avons décidé d’essayer une autre approche. Cette fois, nous avons tenté de géolocaliser les TJ et les établissements pénitentiaires sur une carte interactive, similaire à Google Maps.

### 3.5.1 Cas des TJ

Pour ce faire, dans l’une des tables fournissant une vue augmentée de tous les éléments de structure de GESTSR, on filtre tous les TJ où la date d’application (D\_APP) est antérieure ou égale au 1er janvier 2024. De plus, nous avons conservé les enregistrements où la date de fermeture (D\_FER) est soit manquante, soit postérieure au 1er janvier 2024. Ensuite, nous avons sélectionné les variables pertinentes comme :

- **I\_ELST** : Le code de l’établissement
- **L\_ADR\_POS\_1** : L’adresse postale de l’établissement
- **C\_BDC** : le code bureau distributeur de la commune
- **C\_COM\_LOC** : Le code commune locative
- **L\_COM** : Le nom de la commune

Afin d’obtenir les coordonnées géographiques des TJ du type longitude et latitude, nous avons exporté ce data frame dans un fichier CSV vers l’API de géolocalisation de la Base Adresse Nationale (BAN) sur le site du gouvernement [6] en spécifiant les colonnes **L\_COM** et **L\_ADR\_POS\_1** pour la géolocalisation .

Cet API en retour nous renvoie notre CSV avec de nouvelles variables dont parmi elles les coordonnées géographiques . Elle nous rend aussi par exemple les vraies adresses

complètes qu'elle a pris en compte comme référence, et un score compris entre 0 et 1 indiquant la pertinence du résultat.

Enfin, après récupération des données renvoyées par l'API, nous avons utilisé la bibliothèque **leaflet** pour créer une carte interactive. Nous avons ajouté les tuiles de fond de carte et placé des marqueurs aux coordonnées géographiques (longitude et latitude) des établissements géolocalisés. Les popups des marqueurs affichent le nom de l'établissement ainsi que leur adresse, permettant ainsi une visualisation dynamique et interactive des données.



FIGURE 3.10 – Aperçu du fond de carte interactive des TJ en France entière et les COM

En zoomant de plus près sur l'image (figures 3.11 et 3.12), on aperçoit des détails supplémentaires qui étaient moins visibles à première vue.

L'API a globalement bien fonctionné, mais quelques petites anomalies ont été détectées.

Cinq d'entre les TJ ont été mal ou approximativement géolocalisés, à savoir ceux de Nice, Charleville-Mézières, Saint-Brieuc, Lons-le-Saunier et Nouméa. En effet, ce problème était lié aux adresses mal renseignées dans GESTSR et, pour certaines, non reconnues sur le site de data.gouv.fr.

Pour les TJ de Nouméa, Nice et Charleville-Mézières, les adresses respectives dans GESTSR étaient "2 Boulevard Extérieur", "Place du palais de justice" et "9 Esplanade du palais de justice", qui sont similaires à celles de Google Maps, mais en réalité, en se référant à la source des adresses nationales [7], elles n'y figurent pas et ne correspondent pas à celles référencées officiellement, qui sont respectivement "2 Boulevard Extérieur - Rue Auguste Mercier", "Place du palais" et "9 Esplanade Robert Badinter". L'API de la Base Adresse Nationale (BAN) étant très sensible et ces adresses n'étant pas reconnues dans ces villes, elles ont alors été géolocalisées respectivement à Flins-sur-Seine, Dunkerque et Saint-Quentin, car dans ces dernières il existe des adresses similaires à celles de GESTSR.

Quant au TJ de Lons-le-Saunier, il y avait une coquille dans le code postal, tandis qu'à Saint-Brieuc, le TJ se trouve en réalité dans un parc. C'est ce petit détail qui a causé une

erreur. Nous avons donc décidé de mettre l'adresse correspondante à celle de la sortie du parc.

L'une des limites de cette approche est qu'elle n'est qu'un correctif temporaire. Ces erreurs peuvent se reproduire si un statisticien non averti reprend les programmes sans appliquer les correctifs. La meilleure solution serait de modifier la source en collaboration avec les services responsables.



FIGURE 3.11 – Zoom sur les TJ de la France Hexagonale



FIGURE 3.12 – Zoom sur le TJ de Nantes

### 3.5.2 Géolocalisation des établissements pénitentiaires

Pour géolocaliser les centres pénitentiaires, le même procédé que celui utilisé pour les TJ a été appliqué avec mais avec quelques différences.

En effet, contrairement aux TJ, les adresses des centres pénitentiaires n'ont pas été bien renseignées dans le centre de données, ce qui ne permet pas d'utiliser l'API comme dans le cas précédent. Cependant, nous avons à disposition les codes des communes où les centres se situent, ce qui nous a permis de faire une jointure avec le fond de carte non rapproché généré à la figure (3.2) des communes de la France entière.

Disposant donc des géométries des communes où les centres se situent, nous avons, à l'aide des fonctions `st_coordinates` et `st_centroid` de *mapsf*, pu obtenir les coordonnées géographiques de ces communes, en prenant comme hypothèse le centre de la commune en référence. L'idée de prendre un fond de carte rapproché est bénéfique uniquement lorsqu'il s'agit de générer des fonds de cartes statiques, mais dès lors que nous faisons une approche dynamique, cela ne s'avère pas significatif. Dans notre cas, si on prend le fond de carte de base rapproché dès le début figure (3.3), les coordonnées des DROM obtenues ne seront pas bien calculées à cause des similitudes effectuées, ce qui justifie notre choix. Par la suite, nous avons procédé de la même manière que dans le cas des TJ pour obtenir la carte interactive en affichant cette fois, sur les marqueurs, les noms des centres et leur code d'établissement.

Bien évidemment, la limite de cette méthode est qu'elle donne une géolocalisation des établissements pénitentiaires de façon approximative, du fait d'avoir pris les coordonnées du centre des communes.



FIGURE 3.13 – Etablissement pour Mineurs d'Orvault

Dans la figure (3.13), il s'agit de l'établissement pour mineurs d'Orvault, qui a été géolocalisé par notre approche avec le marqueur bleu sur la figure tandis que le marqueur rouge correspond à l'emplacement réel de l'établissement, le trait sépare la distance entre les deux positions.



# Chapitre 4

## Création d'indicateurs spatiaux

Dans le contexte actuel de la surpopulation carcérale en France, la gestion des établissements pénitentiaires constitue un défi majeur qui affecte non seulement la sécurité et le bien-être des détenus, mais aussi l'efficacité du système judiciaire dans son ensemble.

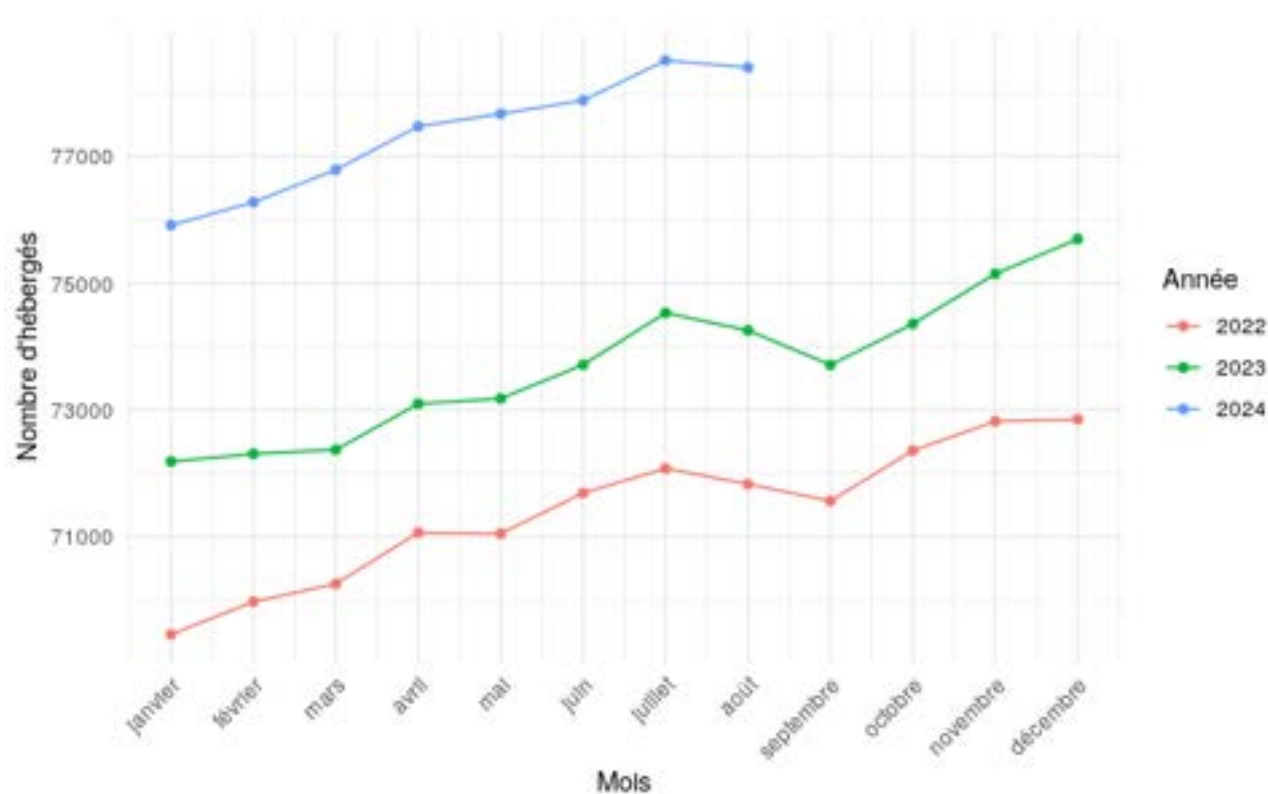


FIGURE 4.1 – Evolution du nombre de détenus depuis 2022

Le graphique présenté ci-dessus montre l'évolution du nombre de personnes incarcérées au premier de chaque mois, de janvier 2022 à août 2024. On constate une augmentation constante du nombre de personnes incarcérées chaque année. En 2022, le nombre d'hébergés en prison débute à 70 000 en janvier et progresse de manière stable pour atteindre environ 72 000 en décembre. Cette tendance haussière se poursuit en 2023, avec un

nombre de détenus commençant à 72 000 en janvier et, malgré quelques fluctuations, se stabilisant autour de 76 000 en décembre. L'année 2024 montre une augmentation encore plus prononcée, le nombre de détenus démarrant à 76 000 en janvier et atteignant environ 78 500 en août.

Cette croissance soutenue souligne une pression croissante sur le système pénitentiaire français et met en lumière l'urgence de mettre en place des mesures efficaces pour faire face aux défis liés à la surpopulation des établissements pénitentiaires mais plus précisément dans les maisons d'arrêt.

Ce chapitre se concentre sur la création d'indicateurs spatiaux pour caractériser la surpopulation des maisons d'arrêt en fonction des caractéristiques des établissements et des détenus. En analysant ces données, nous tenterons d'obtenir une compréhension plus fine des dynamiques démographiques et des conditions de détention.

Ensuite, comme indicateurs, nous utiliserons des techniques de prédiction pour estimer le nombre de détenus dans chaque maison d'arrêt. Ces prédictions seront basées sur des modèles statistiques. L'objectif est de développer des indicateurs précis qui permettront non seulement de quantifier la surpopulation, mais aussi de prédire l'évolution du nombre de détenus dans chaque maison d'arrêt. Ces indicateurs seront ensuite utilisés pour créer des cartes spatiales, permettant de visualiser et d'analyser les dynamiques de surpopulation à travers le territoire.

## 4.1 Présentation de GENESIS

GENESIS, acronyme de GEstion Nationale des personnes Écrouées pour le Suivi Individuel et la Sécurité, est une application indispensable utilisée dans les établissements pénitentiaires français. Elle est spécialement conçue pour la gestion et le suivi des personnes placées sous main de justice, souvent appelées écrouées, qui se trouvent en milieu fermé. Cette application couvre l'ensemble du territoire français, assurant ainsi une gestion cohérente et centralisée des informations relatives aux détenus. GENESIS permet de suivre chaque personne écrouée de manière individuelle, tout en garantissant un haut niveau de sécurité au sein des prisons. En centralisant les données, l'application facilite la coordination entre les différentes institutions pénitentiaires, améliorant ainsi l'efficacité et la transparence du système carcéral. Elle contribue également à une meilleure prise en charge des détenus, en offrant aux professionnels des outils précieux pour assurer leur bien-être et leur réinsertion future.

La figure (4.2) met en lumière les différentes composantes et interactions au sein de GENESIS. Tout d'abord, les personnes placées sous main de justice sont automatiquement mises sous écrou, ce qui signifie qu'elles sont enregistrées dans le système via un processus d'écrou. Une fois écrouées, leurs informations sont intégrées dans GENESIS, qui gère un total de 480 tables primaires. Ces tables primaires sont ensuite utilisées pour générer quatre tables individuelles.

Les tables individuelles alimentent différents stocks d'informations, notamment l'historique de la liste qui stocke toutes les informations relatives aux écroués et à leurs infractions. En outre, ces tables individuelles gèrent les flux entrants et sortants, assurant ainsi une gestion dynamique et continue des données. Cette mise à jour se fait de manière régulière, au premier de chaque mois, garantissant que les informations restent à jour et

précises.

Elles sont essentielles pour produire diverses statistiques et études. Elles permettent de générer les Statistiques Trimestrielles du Milieu Fermé (STMF), les statistiques de réponse à la demande (avec les tables primaires en complément si les données ne s’y trouvent pas).

Ce schéma permet de visualiser comment GENESIS centralise et organise les informations. En offrant une vue d’ensemble claire et détaillée des données, l’application contribue à améliorer la gestion des personnes écrouées et à assurer une meilleure prise en charge de leur situation.

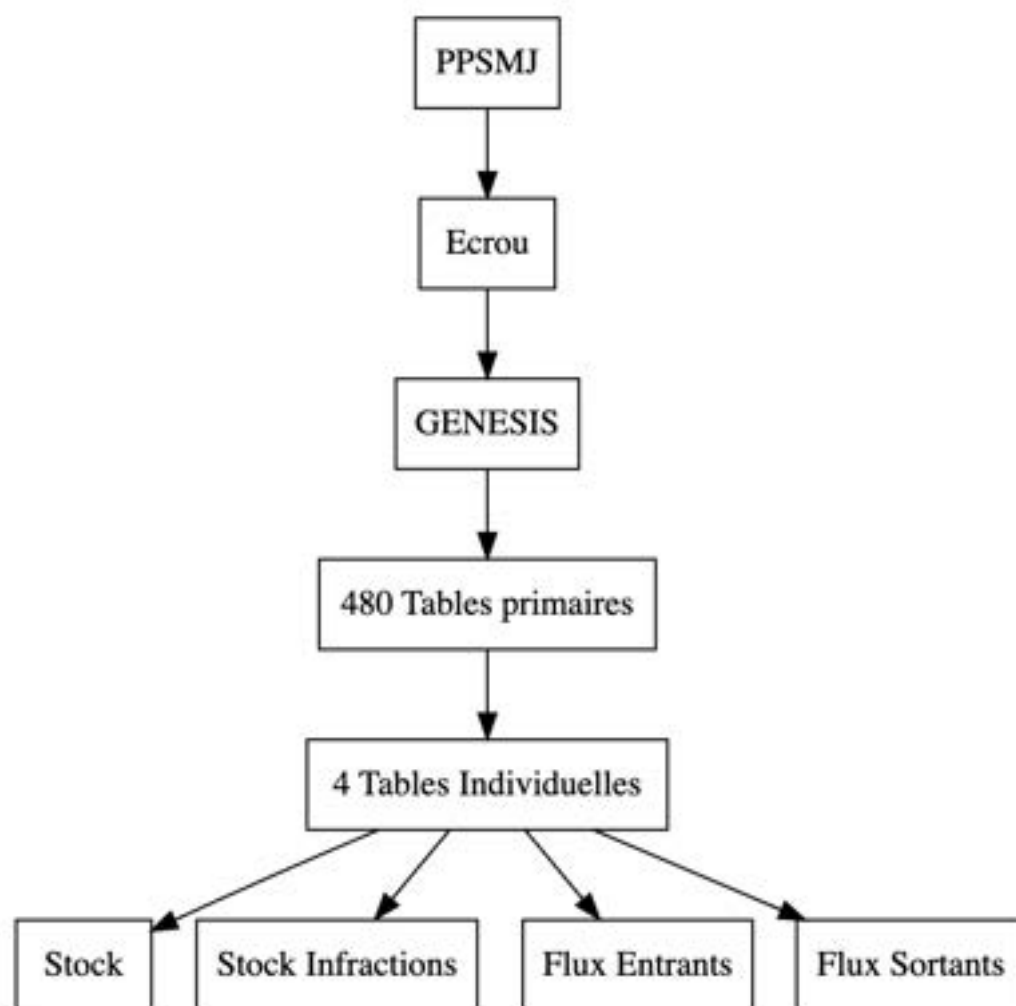


FIGURE 4.2 – Les tables issues de GENESIS

## 4.2 Présentation des données

Pour effectuer nos travaux, nous avons utilisé ces 4 tables individuelles et d’autres données que je citerai par la suite .

La table STOCK\_INFRACTIONS contient 20 variables, elle couvre les individus écroués et condamnés au 1er de chaque mois. L’unité de cette table est :

- **DATE\_OBSERVATION** : la date d'observation de la personne écrouée
- **PER\_ID** : l'identifiant de l'écrouée
- **AFFAIRE\_ID** : l'identifiant de l'affaire
- **INFRACTION\_ID** : l'identifiant de l'infraction

La table FLUX\_ENTRANT donne un effectif de nouveaux écroués. Elle couvre les individus avec mise sous écrou privative de liberté au cours d'un mois. L'unité de cette table est DATE\_OBSERVATION / PER\_ID, et elle contient 20 variables.

La table FLUX\_SORTANT quant à elle donne un nombre de levées d'écrou définitives. Elle couvre les levées d'écrou définitives au cours d'un mois. L'unité de cette table est DATE\_OBSERVATION / PER\_ID, et elle contient 22 variables. Il est important de noter qu'un individu peut avoir plusieurs levées d'écrou sur une même période.

Les tables STOCK\_INFRACTIONS et de FLUX n'existaient pas avant 2021 et sont mensuelles depuis 2022.

La table STOCK donne un effectif d'écroués ventilé selon certaines caractéristiques. Elle couvre les individus écroués au 1er de chaque mois, sur l'ensemble du territoire français. Les périodes de couverture sont annuelles de 2017 à 2020 , puis trimestrielles en 2021 et enfin mensuelles de 2022 à ce jour . L'unité de cette table est DATE\_OBSERVATION / PER\_ID, et elle contient 47 variables.

En raison du manque de renseignement sur les données mensuelles avant 2022 , nous avons décidé de mener nos travaux avec les données de janvier 2022 à août 2024.

Nous avons initialement commencé notre étude avec la table de STOCK. Puis, en utilisant les autres tables et données disponibles, nous avons enrichi puis créé notre base de données pour répondre aux besoins spécifiques de notre étude.

Variables	Description et modalités
DATE_OBSERVATION	Date d'observation de l'écroué Premier de chaque mois : 01-01-2022 à 01-08-2024
DATE_ECROU_FIRST	Date réelle du premier écrou
AGE_OBSERVATION	Âge de l'écroué
SEXE	Sexe de l'écroué
CATEGORIE_PENALE	Condamné : CO Prévenu : PR Condamné-prévenu :CPR
HEBERGEMENT	HEB : Si l'ecroué est hébergé NONHEB : sinon
ETAB_REEL_REGION	La région dans laquelle l'établissement est situé Bretagne , Pays de la loire etc...
NATAFF_CATEG_NATINFP	La nature de l'affaire Vol ,Viol et agression sexuelle etc...
I_ELST_ETAB_ADMIN	Code de l'établissement chargé de l'écroué
QTM_FERME_TACC	Quantum ferme toutes affaires confondues en jours
RELIQUAT	Peine restant à purger par catégories
QTM_FERME_TACC_CATEG	Quantum ferme toutes affaires confondues en catégories

TABLE 4.1 – Tableau des variables avec description et modalités

Après avoir filtré les données sur la période mentionnée précédemment et sélectionné parmi les 47 variables, notre base de données STOCK comprend 2 866 604 individus, c'est-à-dire les personnes écrouées, ainsi que 12 variables décrivant chaque écroué. Le tableau précédent décrit les variables de base qui ont été sélectionnées.

## 4.3 Pré-traitement des données

### 4.3.1 Les données manquantes

Lors de l'analyse préliminaire de notre base de données, nous avons constaté que de nombreuses données manquaient pour deux variables qui sont le quantum ferme toutes affaires confondues en jours ( QTM\_FERME\_TACC) avec la nature d'affaire (NATAFF\_CATEG\_NATINFP) qui avoisinaient l'ordre de 650.000. Ces valeurs manquantes peuvent avoir un impact significatif sur les résultats de notre étude et nécessitent donc une attention particulière. Pour mieux comprendre la distribution des données manquantes, nous avons utilisé le graphique de `gg_miss_var` de la librairie *naniar* (voir figure 4.3).

Aussi la variable du quantum catégorielle ( QTM\_FERME\_TACC\_CATEG), qui est liée à celle en jours, présente bien évidemment la même proportion de données manquantes que celle en jours mais est renseignée dans la base de données à travers une modalité pour spécifier qu'elle est manquante, ce qui fait qu'elle n'apparaît pas sur la figure comme ayant des données manquantes.

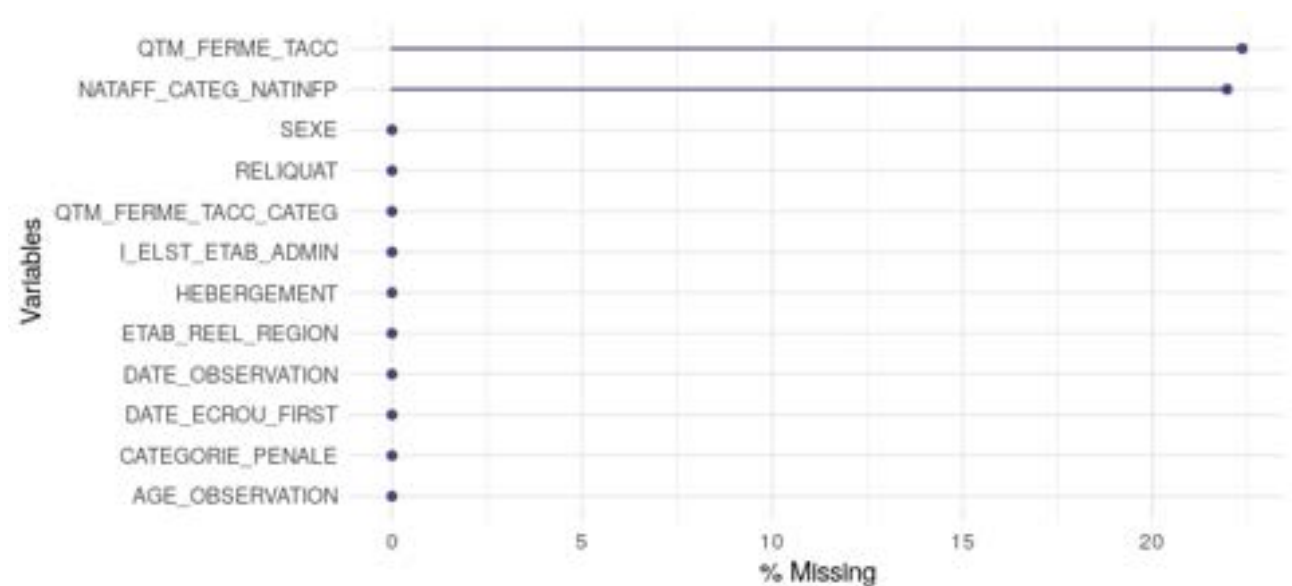


FIGURE 4.3 – Pourcentage des données manquantes des variables

Ces données manquantes sont liées aux catégories pénales, précisément aux écroués prévenus, c'est-à-dire aux personnes en attente de leur jugement. En effet, tant que le jugement n'a pas été rendu, certaines informations, telles que la nature de l'affaire et le quantum de la peine, ne sont pas encore disponibles. Cette situation est courante dans les systèmes judiciaires où les données ne sont complétées qu'après la finalisation des procédures judiciaires.

Compte tenu de l'importance des écroués prévenus dans notre base de données, représentant environ 22 % des individus, il n'est pas envisageable de mener notre étude sans inclure cette population. Exclure les prévenus de l'analyse pourrait biaiser nos résultats et ne pas refléter fidèlement la réalité du système judiciaire. Afin de traiter les données manquantes considérables associées à cette catégorie, nous avons décidé d'apporter des traitements spécifiques. Ces traitements incluent l'imputation des données manquantes.

En ce qui concerne la méthode d'imputation des données manquantes, nous avons décidé de remplacer la variable ( QTM\_FERME\_TACC) par le nombre de jours entre la date d'observation et la date réelle du premier écrou . Cette approche permet de fournir une estimation raisonnable de la durée de la peine pour les écroués prévenus, en attente de leur jugement. Pour la variable nature d'affaire (NATAFF\_CATEG\_NATINFP), nous avons remplacé les valeurs manquantes par une catégorie spécifique indiquant que l'écroué est un prévenu. Cette méthode permet de conserver l'information sur le statut des prévenus tout en minimisant l'impact des données manquantes sur nos analyses. La figure (4.4) représente le pourcentage des données manquantes après ces méthodes d'imputation.

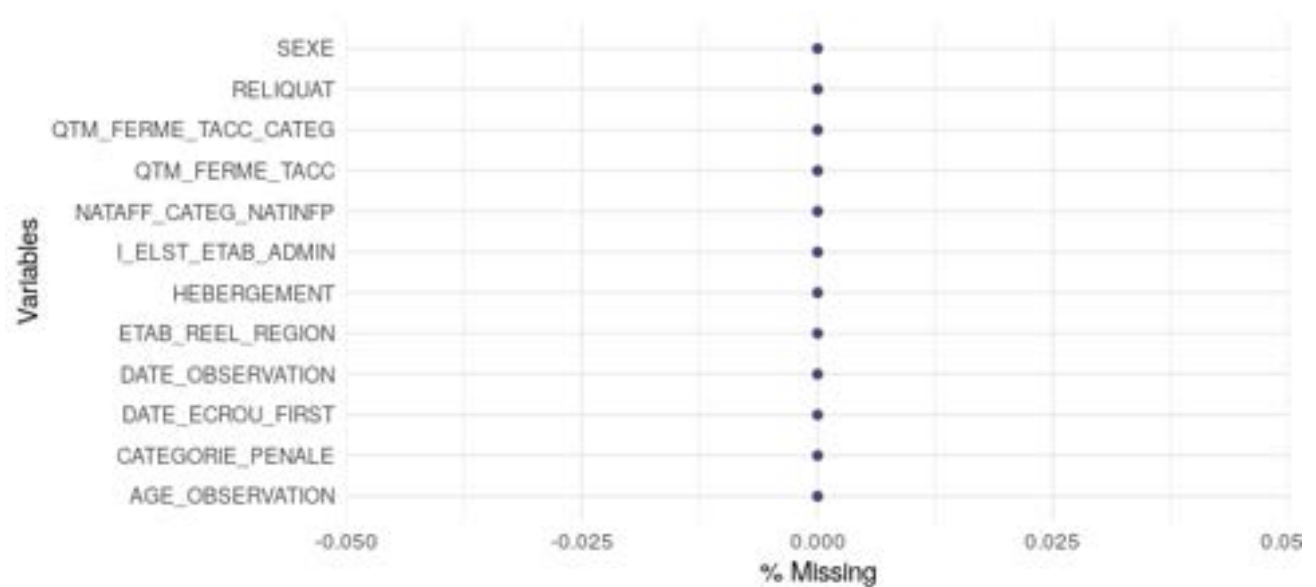


FIGURE 4.4 – Pourcentage des données manquantes après imputation

### 4.3.2 Restructuration de la base de données

Afin de caractériser les maisons d'arrêt, à partir des informations sur les écroués que nous avons à disposition, nous avons transformé notre base de données pour obtenir des statistiques agrégées par établissement et par date d'observation. Voici la démarche détaillée que nous avons suivie :

- **Groupeement des données** : Nous avons groupé les données par date d'observation et par le code de l'établissement administratif (I\_ELST\_ETAB\_ADMIN). Ce groupement nous permet de calculer des statistiques agrégées pour chaque établissement à chaque date d'observation.
- **Calcul des statistiques agrégées** :

- **Nombre total d'écroués** : Nous avons calculé le nombre total d'écroués (TOTAL\_ECROUES) pour chaque groupe c'est à dire pour chaque établissement à chaque premier du mois.
- **Proportion d'hommes et de femmes** : Nous avons calculé la proportion d'hommes (NB\_HOMMES) et de femmes (NB\_FEMMES) parmi les écroués.
- **Proportion de condamnés et de prévenus** : Nous avons calculé la proportion de condamnés (NB\_COND), de condamnés prévenus (NB\_COND\_PRV), de condamnés confondus c'est à dire les condamnés et condamnés prévenus, (NB\_COND\_CONF), et de prévenus (NB\_PRV).
- **Âge moyen** : Nous avons calculé l'âge moyen des écroués (AGE\_MOYEN).
- **Quantum moyen** : Nous avons calculé la moyenne des quantums de peine (QTM\_MOYEN).
- **Proportion d'hébergés et de non-hébergés** : Nous avons calculé la proportion d'écroués hébergés (HEB) et non-hébergés (NONHEB).
- **Proportion de différentes catégories d'infractions** : Nous avons calculé la proportion d'écroués hébergés pour différentes catégories d'infractions, telles que les viols et agressions sexuelles (NB\_VIOL\_AGRSX), les violences (NB\_VIOLIENCE), les vols (NB\_VOL), et les infractions liées aux stupéfiants et à la santé publique (NB\_STUP\_SPUB).
- **Proportion de quantums de peine inférieurs et supérieurs à 2 ans** : Nous avons calculé la proportion d'écroués hébergés avec des quantums de peine inférieurs à 2 ans (NB\_QTM\_INF\_2ANS) et supérieurs à 2 ans (NB\_QTM\_SUP\_2ANS), car les maisons d'arrêt sont spécifiques pour des peines maximales de 2 ans .
- **Catégories d'infractions les plus fréquentes pour les femmes et les hommes** : Nous avons identifié les catégories d'infractions les plus fréquentes pour les femmes (NATINF\_FEMMES) et les hommes (NATINF\_HOMMES).
- **Reliquat le plus fréquent** : Nous avons identifié le reliquat le plus fréquent (RELIQUAT\_FRQ).

Cette démarche nous permet de transformer les données individuelles des écroués en statistiques agrégées par établissement et par date d'observation, ce qui est essentiel pour caractériser les maisons d'arrêt dans notre étude. En suivant cette démarche , nous avons réduit drastiquement la taille de la base de données qui est passée de 2866604 individus à **5903 individus**, où cette fois ce sont tous les établissements pénitentiaires qui sont les individus c'est à dire les maisons d'arrêts, maisons centrales, les centres de détention et les centres de semi liberté .

Une autre table dans GENESIS nous a permis d'identifier parmi les établissements ceux qui étaient des maisons d'arrêt pour effectuer une jointure par le code établissement. Cette démarche nous a permis d'obtenir le libellé de l'établissement et la commune où il se localise.

Disposant alors des libellés des établissements, nous avons filtré toutes les maisons d'arrêt. Pour ce faire, nous avons utilisé une expression régulière pour détecter les libellés commençant par "MA", ce qui correspond aux maisons d'arrêt.

### 4.3.3 Ajout de variables supplémentaires

N'ayant pas suffisamment de variables qui caractérisaient les maisons d'arrêt, nous avons utilisé une autre table existante pour effectuer une jointure par le code de l'établissement et la date d'observation. En effectuant cette jointure, nous avons tenté d'obtenir des informations supplémentaires telles que la capacité théorique, la capacité opérationnelle et le nombre de lits dans chaque maison d'arrêt car ces nombres évoluent suivant certains mois, ainsi vous avons pu créer une variable qui mesurait le taux de surpopulation de l'établissement .

Toujours dans le but d'obtenir suffisamment de variables pour caractériser les maisons d'arrêt, nous avons tenté d'obtenir le nombre d'hébergés entrants et sortants chaque mois dans chaque établissement. Pour ce faire, nous avons utilisé les tables de FLUX\_ENTRANTS et SORTANTS. Ensuite, nous avons calculé un taux d'entrants par rapport aux sortants afin de savoir s'il y a eu un mois donné plus d'entrants que de sortants TAUX\_ENTR\_SORT avant d'effectuer une jointure par le code de l'établissement et la date d'observation pour obtenir une vue complète des taux d'entrants et de sortants dans chaque établissement.

Dans la même idée, nous avons pu obtenir des données provenant de **Cassiopée** (un autre infocentre interne) sur le nombre de décisions d'emprisonnement ferme (EPR) prononcées par les tribunaux judiciaires qui ne prend en compte que les exécutions provisoires, telles que les mandats de dépôt, les maintiens en détention et les ordres d'incarcération immédiate. Ayant ces données, nous avons calculé la moyenne annuelle du nombre de décisions d'emprisonnement ferme tous tribunaux confondus pour chaque année et avons calculé le taux d'EPR (TAUX\_EPR) en divisant le nombre de décisions d'emprisonnement ferme de chaque tribunal par la moyenne annuelle.

Disposant des codes des communes où les maisons d'arrêt se situent ainsi que des tribunaux judiciaires, nous avons effectué certaines jointures à partir des géofichiers des fonds de carte générés dans les chapitres précédents. D'abord, nous avons joint notre base de données des maisons d'arrêt au fichier du fond de carte des communes (figure 3.3). Ensuite, nous avons joint la base de données des EPR avec celui du fond de carte des tribunaux (figure 3.4). Cette démarche nous a permis d'obtenir les géométries des communes où chaque établissement est situé.

Nous avons ensuite tenté de joindre notre base de données avec celle des EPR en utilisant les données du tribunal judiciaire le plus proche de chaque maison d'arrêt. Cette approche est basée sur l'hypothèse que, dans les procédures judiciaires, pour réduire les risques d'évasion, on a tendance à incarcérer les détenus dans la maison d'arrêt la plus proche du tribunal où la peine a été prononcée.

Pour ce faire, nous avons suivi la démarche suivante :

- **Calcul des centroïdes** : A l'aide de la fonction `st_centroid` du package `sf` , nous avons calculé les centroïdes des géométries des communes de localisation des maisons d'arrêt et des tribunaux judiciaires. Cela nous a permis de représenter chaque établissement par le centre de la commune .
- **Détermination du tribunal le plus proche** : Nous avons utilisé la fonction `st_nearest_feature` pour déterminer le tribunal judiciaire le plus proche de chaque maison d'arrêt.



- **Ajout de l'identifiant du tribunal le plus proche** : Nous avons ajouté le code du tribunal judiciaire le plus proche à notre base de données des maisons d'arrêt.

Cette démarche nous a permis de joindre efficacement notre base de données des maisons d'arrêt avec celle des EPR, en tenant compte de la proximité géographique des tribunaux judiciaires et les dynamiques des procédures judiciaires.

Ainsi, après avoir supprimé certaines variables qui s'avéraient finalement pas pertinentes, nous avons obtenu notre base de données définitive composée de 2365 individus et 16 variables. Notons bien qu'il existait 75 maisons d'arrêt en France hexagonale, et nous avons pris les données du premier de chaque mois entre le 01-01-2022 et le 01-08-2024, ce qui équivaut à 32 mois. Logiquement, nous devrions donc avoir une base de données composée de 2400 individus ( $75 \times 32$ ), mais certaines maisons d'arrêt ont fermé à partir de certains mois, d'où le nombre de 2365 individus.

## 4.4 Analyses exploratoires des données

### 4.4.1 Analyse en composantes principales

Nous allons maintenant procéder à une analyse en composantes principales (ACP). Cette méthode est essentielle pour réduire la dimensionnalité de notre base de données et identifier les structures sous-jacentes. L'objectif est de simplifier les données tout en conservant le maximum d'informations, ce qui permet de visualiser les relations entre les variables et les individus de manière plus claire. Pour ce faire, notre base de données comprenant les maisons d'arrêt à chaque mois, nous avons calculé une moyenne par maison d'arrêt. Cela nous permet d'obtenir une base de données composée des moyennes des variables regroupées par maison d'arrêt. Ainsi, nous avons à présent 75 individus avec le même nombre de variables, ce qui facilite l'analyse et la comparaison entre les différentes maisons d'arrêt.

À l'aide de la fonction **PCA** du package **FactoMineR**, nous avons réalisé notre ACP centrée et normée. Afin de déterminer le nombre de composantes principales optimales, nous avons tracé un barplot représentant la variance expliquée par chaque composante (figure 4.5). Ce graphique nous aide à visualiser la contribution de chaque composante à la variance totale et à identifier le nombre de composantes nécessaires pour capturer une proportion significative de la variance des données.

Afin de conserver le maximum d'informations, nous avons utilisé la règle du coude (WSS) pour déterminer le nombre optimal de composantes principales. Selon cette règle, nous avons choisi de retenir trois composantes principales, qui permettent d'expliquer environ 74% de l'inertie totale des données. Cependant, pour simplifier l'analyse et la visualisation, nous nous concentrerons principalement sur les deux premières composantes principales dans la suite de notre étude. Ces deux composantes captent une grande partie de la variance des données et offrent une représentation suffisamment informative pour notre analyse.

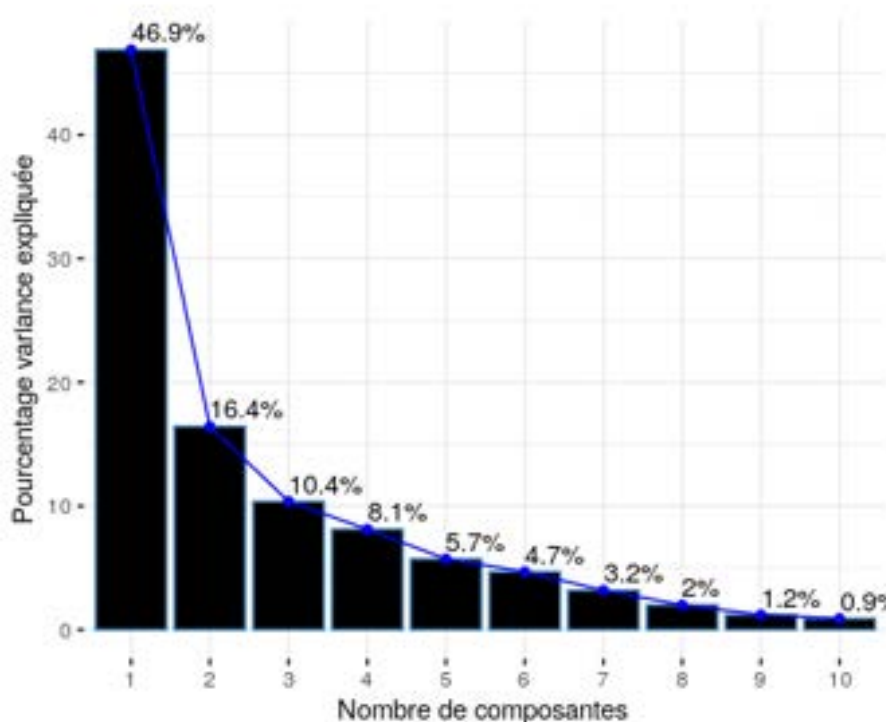


FIGURE 4.5 – Pourcentage de variance expliquée par les composantes principales

#### 4.4.1.1 Analyse des corrélations des variables

Le cercle de corrélations suivant (figure 4.6) est tracé grâce à la fonction **fviz\_pca\_var** de la librairie **factoextra** en ajoutant le taux de surpopulation comme variable supplémentaire. Les flèches du cercle ont été dégradées en fonction du  $\cos^2$ , qui correspond à la qualité de représentation des variables. Selon les deux axes, les variables ont été majoritairement plutôt bien représentées.

Les variables quantum moyen des écroués dans l'établissement (QTM\_MOYEN), quantum supérieur à 2 ans (QTM\_sup\_2ans), nombre de lits (NOMBRE\_LITS), le nombre total d'écroués (TOTAL\_ECROUES) et la capacité théorique de la maison d'arrêt CAPACITES\_THEORIQUES sont corrélées positivement à l'axe 1, tandis que les variables quantum inférieur à 2 ans (QTM\_inf\_2ans) et taux entrants-sortants (TAUX\_ENTR\_SORT) sont corrélées négativement avec l'axe 1.

Quant au deuxième axe, les variables taux de vol (TAUX\_VOL), taux de femmes (TAUX\_FEMMES), taux d'emprisonnement ferme prononcé par tribunal (TAUX\_EPR), et le taux des condamnés confondus c'est à dire les condamnés et condamnés-prévenus (TAUX\_COND\_CONF) sont corrélées positivement à cet axe, tandis que les variables âge moyen (AGE\_MOYEN), taux des prévenus (TAUX\_PRV) et taux\_surpopulation sont corrélées négativement avec celui-ci. Par contre, les variables des taux d'hébergés (HEB) et non hébergés (NONHEB) sont plutôt semi-corrélées aux deux axes. La variable HEB est semi-corrélée positivement avec l'axe 1 et semi-corrélée négativement avec l'axe 2, tandis que la variable NONHEB suit le même scénario mais à l'inverse.

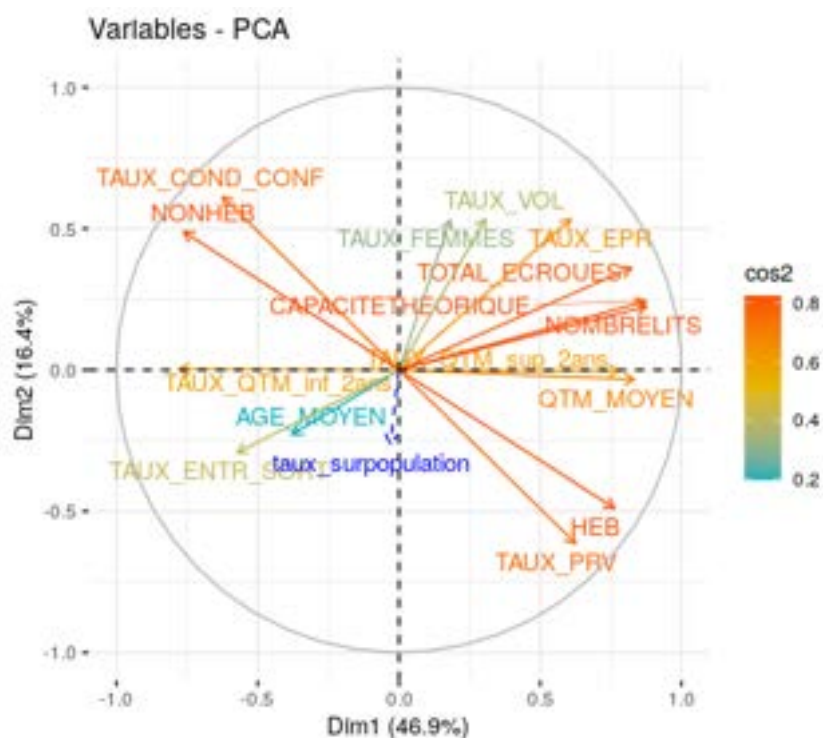


FIGURE 4.6 – Cercle de corrélations des variables

#### 4.4.1.2 Analyse des contributions des variables

Ayant une idée des corrélations des variables avec chaque axe, nous pouvons analyser la contribution de chaque variable sur chaque axe. Le tableau suivant représente la contribution de chaque variable sur chaque axe.

Variables	Axe 1		Axe 2	
	+	-	+	-
QTM_MOYEN	+			
QTM_sup_2ans	+			
NOMBRE_LITS	+			
TOTAL_ECROUES	+			
CAPACITES_THEORIQUES	+			
QTM_inf_2ans		-		
TAUX_VOL			+	
TAUX_FEMMES			+	
TAUX_EPR			+	
TAUX_COND_CONF			+	
TAUX_ENTR_SORT		-		
TAUX_PRV				-
AGE_MOYEN				-
taux_surpopulation				-
HEB	+			-
NONHEB		-	+	

TABLE 4.2 – Contribution de chaque variable sur chaque axe

#### 4.4.1.3 Projection des individus

Les individus sont majoritairement tous regroupés ensemble dans la figure (4.7) ce qui forme des clusters sauf quelques individus qui sont plutôt des outliers. Les outliers, définis comme des individus qui se distinguent nettement des autres en raison de leurs caractéristiques atypiques, sont les individus 59, 70 et 73 qui sont respectivement les maisons d'arrêt de Versailles, Lyon et Nancy. En effet, cela est lié au fait que la maison d'arrêt de Versailles est la seule qui héberge uniquement des femmes, tandis que les maisons d'arrêt de Lyon et Nancy sont celles qui disposent du plus grand nombre de lits et dont la capacité théorique de leurs établissements est assez grande .

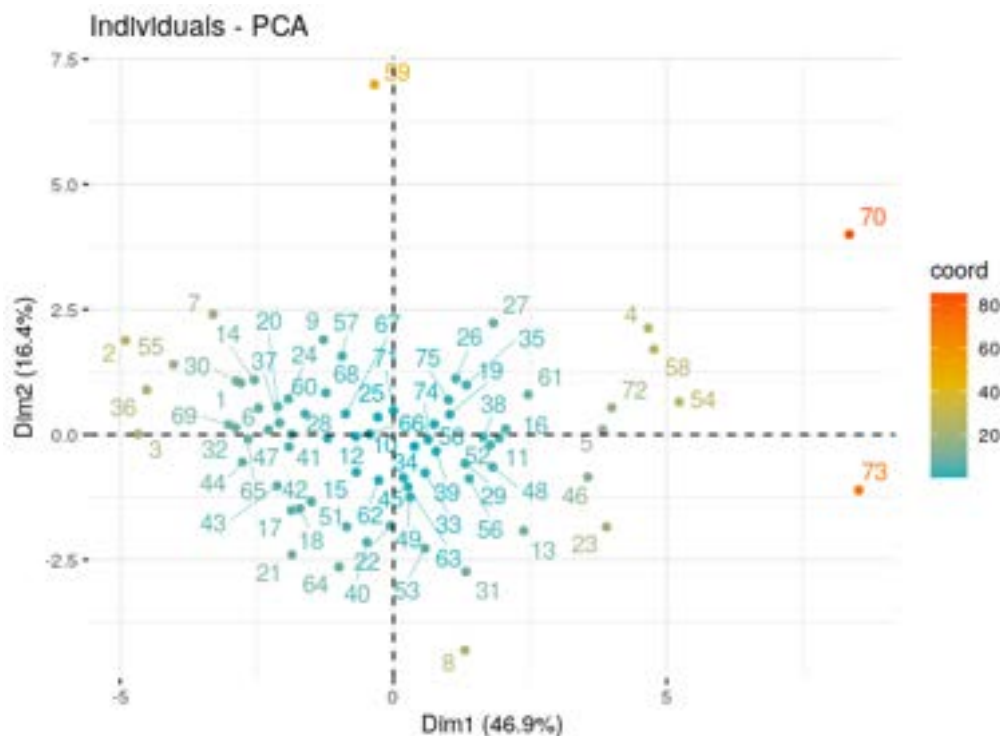


FIGURE 4.7 – Coordonnées des individus dans le plan 1-2

La présence de clusters indique que les individus partagent des caractéristiques communes, ce qui justifie une classification. En effet, les individus situés dans le même cluster ont des profils similaires en termes de variables étudiées. Cette observation est renforcée par la proximité des points dans l'espace des composantes principales, ce qui montre que les individus au sein d'un même cluster sont plus proches les uns des autres que des individus appartenant à d'autres clusters.

#### 4.4.2 Classification des maisons d'arrêt

Dans la figure (4.8), à l'aide des fonctions **kmeans** et **fviz\_nbclust** de la librairie **cluster**, et la règle du coude, nous avons effectué une classification par partition en 3 classes en utilisant les coordonnées des individus avec différents algorithmes tels que Hartigan-Wong, MacQueen et Lloyd. Cependant, les résultats étaient plutôt similaires dans l'ensemble. Par conséquent, nous avons fait le choix d'utiliser l'algorithme Hartigan-Wong.

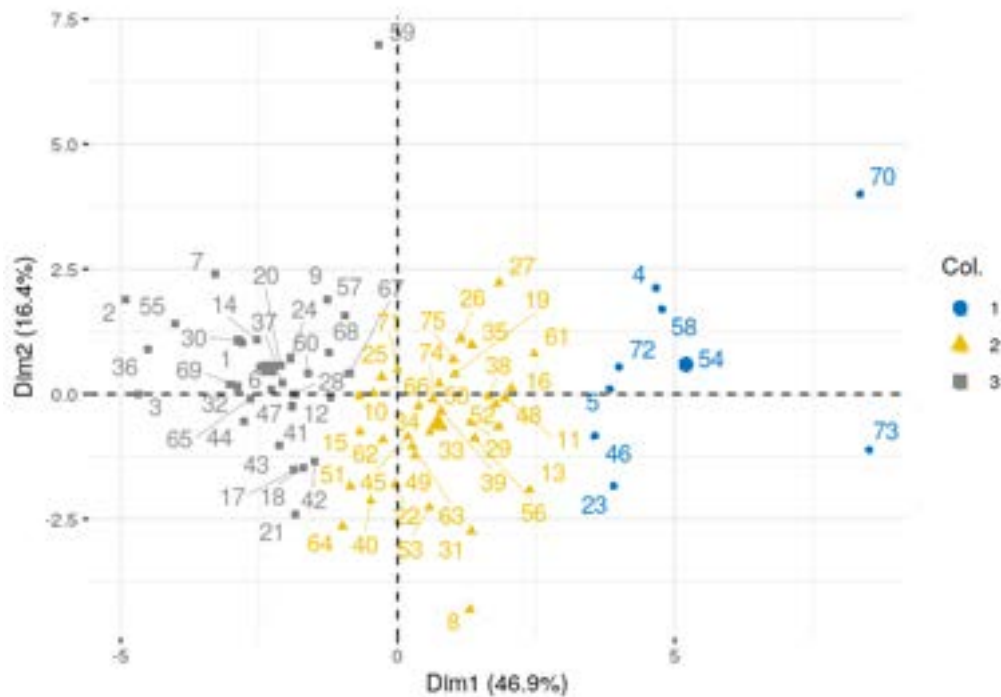


FIGURE 4.8 – Plan factoriel des individus avec clusters 3-means

Le tableau suivant présente quelques individus dans chaque classe .

Cluster en bleu	Cluster en gris	Cluster en jaune
MA de Lyon (70)	MA de Digne les bains (2)	MA de Caen (11)
MA de Nancy (73)	MA de Vesoul (55)	MA d'Angoulême (13)
MA de Rouen (58)	MA de Gap (3)	MA de Bourges (16)
MA de Strasbourg (54)	MA de Cherbourg (36)	MA de Chambéry (56)
MA de Grasse (4)	MA Charleville-Mézières(7)	MA de Mende (34)
MA de Nice (5)	MA de Lons-le-Saunier (30)	MA d'Amiens (61)
MA de Draguignan (72)	MA de Saint-Brieuc (20)	MA de Nîmes (27)

TABLE 4.3 – Quelques individus de chaque classe avec le libellé des maisons d'arrêt

Le cluster en bleu, composé des individus correspondant aux maisons d'arrêt, est caractérisé par le plus grand nombre d'écroués. Ces établissements ont une capacité d'accueil et un nombre de lits très grand. Il est à noter que les maisons d'arrêt de Strasbourg et Besançon sont les seules où tous les écroués sont hébergés.

Le cluster en gris, quant à lui, correspond aux établissements qui ont un très fort taux de condamnés confondus, c'est-à-dire les condamnés et les condamnés-prévenus tous réunis.

Enfin, le dernier cluster en jaune correspond aux établissements où le quantum de peine moyen est élevé.

La suite de nos travaux consistera à développer des modèles d'apprentissage statistique afin d'obtenir des indicateurs des taux de surpopulation des maisons d'arrêt.

## 4.5 Création d'un modèle de prédiction du taux de surpopulation des maisons d'arrêt

### 4.5.1 Méthodes d'apprentissage statistique

Nous partons de notre base de données composée de 2365 individus et 16 variables mentionnées plus haut, qui comporte les données des maisons d'arrêt du premier de chaque mois entre le 01-01-2022 et le 01-08-2024 inclus. Dans cette partie, le but sera d'entraîner des modèles sur les données jusqu'au 01-07-2024 et de prendre les données du 01-08-2024 comme échantillon de test.

La méthode générale consiste à construire des modèles de machine learning pour expliquer le taux de surpopulation en fonction des autres variables. Pour ce faire, nous utilisons les données historiques disponibles pour entraîner les modèles. L'entraînement des modèles implique l'utilisation d'algorithmes qui apprennent à partir des données passées pour identifier les relations et les tendances entre les différentes variables et le taux de surpopulation. Une fois les modèles entraînés, nous les testons sur un échantillon de données distinct, en l'occurrence les données du 01-08-2024, pour évaluer leur performance et leur capacité à prédire le taux de surpopulation de manière précise. Cette approche permet de valider la robustesse des modèles et de s'assurer qu'ils peuvent être généralisés efficacement à de nouvelles données.

#### 4.5.1.1 Régression linéaire

La régression MCO est une méthode de régression linéaire classique qui utilise la technique des moindres carrés pour estimer les coefficients des variables indépendantes. Elle prédit la valeur de la variable cible en utilisant une équation linéaire à l'aide de la méthode **lm** de la fonction **train** du package R *caret*. Les coefficients de régression sont estimés en minimisant la somme des carrés des différences entre les valeurs observées et les valeurs prédites.

#### 4.5.1.2 Régression par arbre de décision

La régression par arbre de décision est une méthode d'apprentissage automatique qui utilise un arbre de décision pour prédire une variable continue. L'arbre est construit en divisant les données en sous-ensembles basés sur les valeurs des variables.

Pour prédire la valeur de la variable dépendante, la régression par arbre de décision suit un chemin de décisions basé sur les valeurs des variables indépendantes. Chaque nœud de l'arbre représente une décision basée sur une variable indépendante, et chaque feuille de l'arbre représente une prédiction. La prédiction finale est obtenue en suivant les décisions de la racine de l'arbre jusqu'à une feuille. La méthode **rpart** du package *rpart* nous a permis à réaliser cette régression.

#### 4.5.1.3 Régression par la méthode Xgboost

XGBoost (Extreme Gradient Boosting) est elle aussi une méthode d'apprentissage automatique qui utilise un ensemble de modèles de régression par arbre de décision pour prédire une variable continue. Elle repose sur le principe du boosting, où chaque nouveau modèle est ajouté pour corriger les erreurs des modèles précédents. XGBoost est réputé

pour sa rapidité et sa précision, et il utilise des techniques avancées comme le gradient boosting et la régularisation pour améliorer la précision et éviter le surapprentissage et nous l'avons réalisé grâce à la méthode **XGbTree** du package **XGBoost**.

XGBoost prédit la valeur de la variable dépendante en combinant les prédictions de plusieurs arbres de décision. Chaque arbre est construit pour minimiser l'erreur résiduelle des arbres précédents. La prédiction finale est obtenue en additionnant les prédictions de tous les arbres, pondérées par leurs coefficients respectifs. XGBoost utilise des techniques avancées comme le gradient boosting et la régularisation pour améliorer la précision et éviter le surapprentissage.

## 4.5.2 Comparaison des résultats

En toute logique, il serait judicieux de créer un échantillon de validation pour évaluer la performance de nos modèles. Cependant, avec une petite base de données de 2365 individus, la division des données en ensembles d'entraînement et de validation peut réduire encore plus la taille de chaque ensemble, entraînant une perte d'information et rendant l'entraînement des modèles moins efficace. Les modèles comme XGBoost et les arbres de décision sont complexes et nécessitent souvent un grand nombre de données pour être entraînés de manière efficace. Avec une petite base de données, il peut être difficile de capturer toutes les relations et les tendances dans les données. De plus, la validation croisée, une technique courante pour évaluer la performance des modèles, peut être difficile à appliquer avec des petites bases de données. Par exemple, la validation croisée k-fold peut diviser les données en k sous-ensembles, rendant chaque sous-ensemble encore plus petit et moins représentatif.

### 4.5.2.1 Qualité d'ajustement

Le tableau ci-dessous représente les  $R^2$  des trois modèles ajustés.

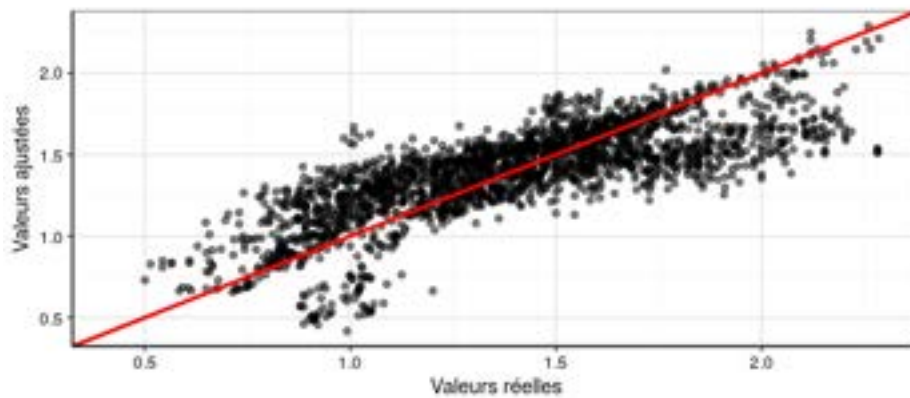
Modèle	$R^2_{\text{train}}$
Modèle MCO	0.61
Modèle rpart	0.46
Modèle XGBoost	0.98

TABLE 4.4 – Récapitulatif des scores  $R^2$  d'ajustement des modèles

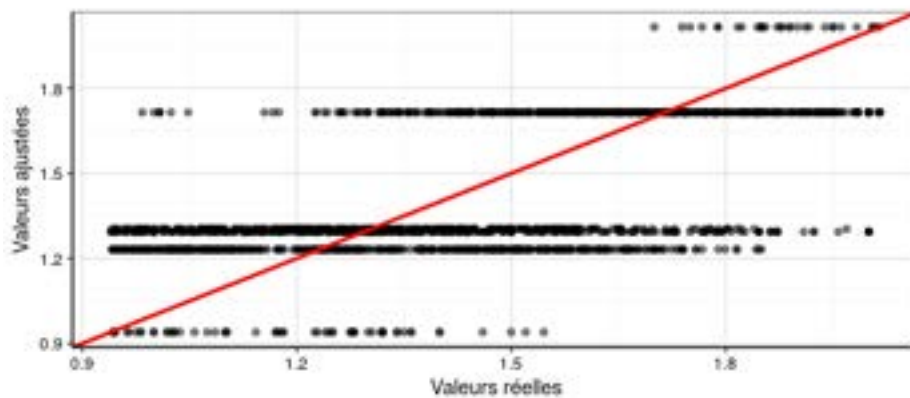
Le score  $R^2$  indique la proportion de la variance de la variable dépendante qui est expliquée par le modèle. Un score  $R^2$  de 0.61 pour le modèle MCO signifie que 61% de la variance de la variable dépendante est expliquée par ce modèle. Bien que ce score soit raisonnable, il laisse encore une part significative de la variance inexpliquée. Le modèle avec les arbres de décision ou rpart, avec un score  $R^2$  de 0.46, explique 46% de la variance, ce qui est relativement faible et suggère que ce modèle pourrait ne pas capturer toutes les relations importantes dans les données. En revanche, le modèle XGBoost présente un score  $R^2$  de 0.98, ce qui est exceptionnellement élevé et indique que 98% de la variance de la variable dépendante est expliquée par ce modèle. Cela suggère que le modèle XGBoost est particulièrement efficace pour capturer les relations complexes dans les données et offre une performance très élevée sur l'ensemble d'entraînement. Cependant, il est important de noter que des scores  $R^2$  très élevés peuvent également indiquer un risque de

surapprentissage, et il est donc crucial de valider ces résultats sur un ensemble de données de validation pour s'assurer de la généralisation du modèle , mais pour les raisons citées plus haut , nous allons pas poursuivre cette démarche .

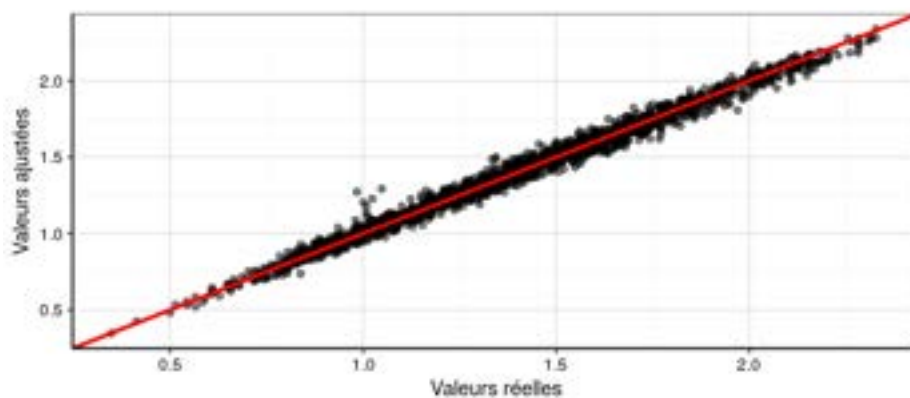
La figure suivante présente les valeurs ajustées d'entraînement en fonction des vraies valeurs .On observe bien le phénomène lié aux  $R^2$  précédents.



(a) MCO



(b) Arbre de décision



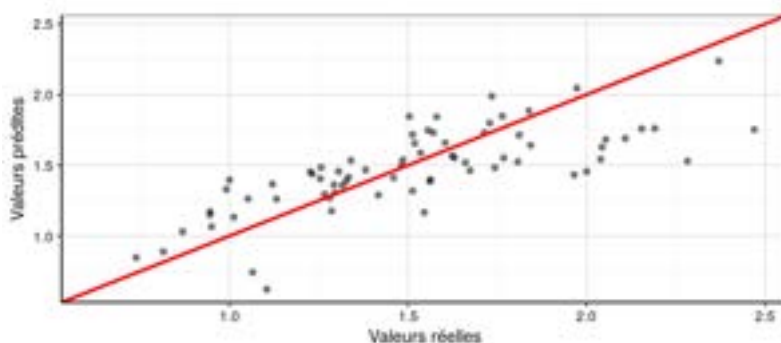
(c) XGBoost

FIGURE 4.9 – Valeurs d'entrainement ajustées par chaque modèle

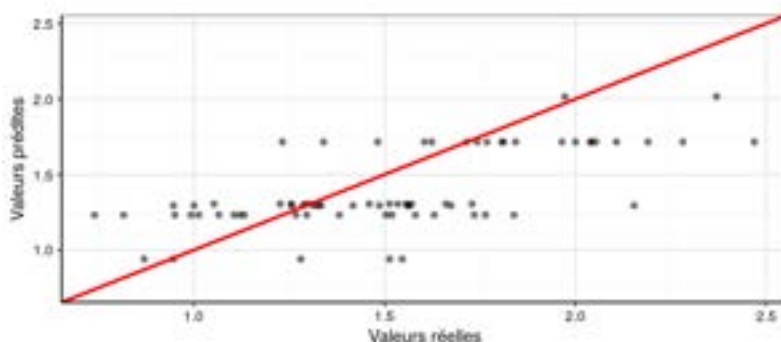


#### 4.5.2.2 Qualité de prédiction

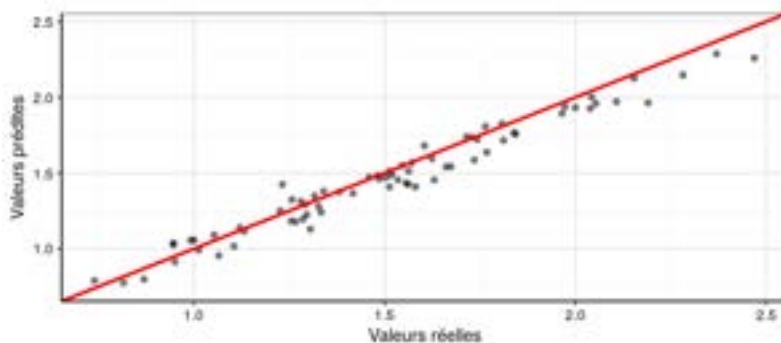
La figure suivante présente les valeurs prédites sur les données de test en fonction des vraies valeurs. La droite rouge correspond à un outil visuel pour évaluer la performance du modèle de régression. Elle permet de voir rapidement à quel point les prédictions du modèle sont proches des vraies valeurs. Plus les points sont proches de cette droite, meilleure est la performance du modèle. En observant ces nuages de points, nous constatons que c'est le modèle XGBoost qui prédit le mieux, en raison de l'écart minimal entre les points et la droite rouge. Le modèle MCO, quant à lui, prédit de manière moyennement satisfaisante. Cependant, le modèle rpart demeure celui qui prédit le moins bien, ce qui est peut-être dû à la taille de nos données.



(a) MCO



(b) Arbre de décision



(c) XGBoost

FIGURE 4.10 – Valeurs prédites par chaque modèle

Pour confirmer nos propos, nous avons calculé le RMSE, un outil utilisé pour évaluer la précision des prédictions et obtenir une idée du taux d'erreur. Il en résulte que le modèle XGBoost est celui qui se trompe le moins, suivi par le modèle MCO, et enfin par le modèle basé sur les arbres. En effet, leurs taux d'erreur respectifs sont de 9%, 0.26% et 0.31%. Ces valeurs prédites correspondent aux taux de surpopulation des maisons d'arrêt au 01-08-2024 qu'on considérera comme nos indicateurs spatiaux .

Du fait de ne pas avoir pu réaliser une procédure de validation croisée en raison de la dimension des données, j'ai adopté une approche naïve en supposant que les données du mois de juillet et d'août étaient similaires. Cette approche a servi de point de référence ou de benchmark. Les résultats des deux approches se sont révélés quasiment identiques, avec une différence de seulement 4 % de taux d'erreur en faveur du meilleur modèle de la démarche naïve. Cela indique que les modèles sont robustes et que les prédictions sont fiables.

Cette similitude montre que, dans certains cas, une approche naïve peut fournir des résultats très proches de ceux obtenus par des méthodes plus complexes. Cela souligne l'importance de tester des hypothèses simples après avoir exploré des méthodes plus sophistiquées, car elles peuvent parfois offrir des solutions efficaces et moins coûteuses en termes de temps et de ressources.

## 4.6 Représentation spatiale des taux de surpopulation des Maisons d'Arrêt

Maintenant que nous avons obtenu des indicateurs, nous allons les mettre en valeur sur le fond de carte des Directions Interrégionales des Services Pénitentiaires créé dans le chapitre 3 (figure 3.5). Pour ce faire, nous allons dans un premier temps associer ces prédictions aux données qui correspondent au mois d'août. Ensuite, disposant dans le fichier de fond de carte de DISP créé, de la variable qui correspond au code d'établissement de la direction interrégionale, et par ailleurs, des codes d'établissement (I\_ELST\_ETAB\_ADMIN) des maisons d'arrêt , nous allons effectuer une jointure par le code de rattachement. Cela nous permettra de regrouper chaque maison d'arrêt à sa DISP de rattachement pour enfin calculer une moyenne des prédictions des taux de surpopulation regroupés par directions interrégionales .

Le choix du fond de carte Directions Interrégionales des Services Pénitentiaires dans la figure (4.11) est particulièrement judicieux pour notre analyse. En effet, les établissements pénitentiaires sont organisés et gérés au niveau des directions interrégionales, ce qui nous permet de mieux comprendre et gérer les données à un niveau régional. En utilisant ce fond de carte, nous pouvons visualiser les taux de surpopulation des maisons d'arrêt dans un contexte géographique plus large, ce qui est essentiel pour identifier les tendances et les disparités régionales. De plus, ce fond de carte nous permet de regrouper les données par région administrative, facilitant ainsi les comparaisons entre différentes zones géographiques.

Dans la figure (4.11), nous avons créé une carte choroplèthe pour obtenir un aperçu géographique du taux de surpopulation. En se basant sur ce graphique, nous constatons que toutes les directions interrégionales des services pénitentiaires sont surpeuplées. Cependant, il ressort que les DISP les plus surpeuplées sont celles de Rennes et de Toulouse,

tandis que les moins surpeuplées sont celles de Strasbourg et de Paris.



FIGURE 4.11 – Prédiction du taux de surpopulation moyenne des MA par DISP au mois d'Aout 2024

# Chapitre 5

## Conclusion et appréciation de l'expérience du stage

La création de fonds de carte s'est révélée être un outil d'analyse extrêmement utile, offrant une visualisation claire et précise des données géographiques. Cependant, pour garantir l'exactitude et la pertinence de ces analyses, il est crucial de prêter une attention particulière au choix du projecteur. Le projecteur détermine la manière dont les données géographiques sont représentées sur une surface plane, et un mauvais choix peut entraîner des distorsions significatives. Ces distorsions peuvent affecter les distances, les angles, les aires ou les formes des objets représentés, ce qui peut fausser les conclusions tirées des analyses. Par conséquent, il est essentiel de sélectionner le projecteur le plus adapté aux besoins spécifiques de chaque étude pour assurer la fiabilité et la précision des résultats. De plus, il est important de choisir le bon type de carte en fonction de ce que l'on veut mettre en valeur, car chaque catégorie de carte est adaptée à un cas précis.

L'étude sur les maisons d'arrêt, visant à obtenir des indicateurs, a également montré que l'on pouvait interpréter les résultats d'un point de vue géographique en les mettant en valeur sur une carte. La collaboration avec les services statistiques du ministère a été enrichissante et a permis de mettre en lumière l'importance de la dimension géographique dans l'analyse des données judiciaires. Les fonds de carte créés à partir des données de GESTSR et les indicateurs sur la surpopulation des maisons d'arrêt, élaborés grâce à GENESIS, ont démontré leur utilité pour une analyse territoriale précise et détaillée.

Pendant ce stage, j'ai eu l'occasion d'animer un atelier de cartographie, qui a permis aux membres de l'équipe de s'initier et de s'intéresser à la cartographie. J'ai également participé à une multitude de réunions, ce qui m'a permis d'apprendre à mieux m'exprimer devant un public.

Ce stage m'a également permis de développer des compétences techniques en cartographie et en analyse de données, tout en renforçant ma capacité à travailler en équipe et à m'adapter à des environnements professionnels exigeants. L'expérience acquise au sein du ministère de la Justice a été extrêmement formatrice et m'a offert une vision approfondie des défis et des opportunités dans le domaine de la justice pénale.

En conclusion, ce stage de fin d'études a été une expérience enrichissante et formatrice, tant sur le plan professionnel que personnel. Les résultats obtenus montrent que la mise à disposition de fonds de carte spécifiques et la création d'indicateurs spatiaux sont des

outils essentiels pour améliorer l'efficacité et la précision des analyses judiciaires. Je suis reconnaissant pour cette opportunité et je suis convaincu que les compétences et les connaissances acquises durant ce stage me seront précieuses dans ma future carrière.

# Bibliographie

- [1] Nicolas LAMBERT et Christine ZANIN. *Chapitre 1. Le fond de carte*. Sous la dir. de Nicolas LAMBERT et Christine ZANIN. Cursus. Paris : Armand Colin, 2016, p. 29-51. URL : <https://www.cairn.info/manuel-de-cartographie--9782200612856-page-29.htm>.
- [2] Arthur H ROBINSON et al. *Elements of Cartography*. 6th. New York : John Wiley & Sons, 1995. URL : <https://www.nhbs.com/elements-of-cartography-book>.
- [3] P. BOLSTAD. *GIS Fundamentals : A First Text on Geographic Information Systems*. XanEdu, 2019. ISBN : 9781593995522. URL : <https://books.google.fr/books?id=fUU7yQEACAAJ>.
- [4] CRÉACARTES. *Fonds de cartes commune France Entière 2024*. 2024. URL : <https://creacartes.s3.eu-west-3.amazonaws.com/2024/franceentiere/index.html>.
- [5] INSEE. *Information*. 2023. URL : <https://www.insee.fr/fr/information/6800675>.
- [6] DATA.GOUV. *API Adresse*. URL : <https://adresse.data.gouv.fr/api-doc/adresse>.
- [7] DATA.GOUV. *Base Adresse Nationale*. URL : <https://adresse.data.gouv.fr/#17.75/51.03135/2.37595>.