Centre d'épidémiologie sur les causes médicales de décès – Institut national de la santé et de la recherche médicale

Fiche méthodologique n°3

Anonymisation des données en *open data* du CépiDc

Yann Aubineau
Fanny Godet
Inserm-CépiDc

Version n°1 – Janvier 2026

Cette fiche méthodologique de travail ne reflète pas la position de l'Inserm et n'engagent que ses auteurs.

Ce qu'il faut retenir

La base de données mise à disposition sur le site d'*open data* du CépiDc a fait l'objet d'une procédure d'anonymisation visant à empêcher le dévoilement de l'identité des personnes décédées, et donc de leur cause de décès, tout en conservant la qualité de l'information diffusée.

Pour ce faire, le département de résidence de 2 % des individus qui composent la base de données a été modifié. Dans l'extrême majorité des cas, le département d'échange se trouvait dans la même région. **Cela signifie qu'au niveau régional, les effectifs et taux sur les causes de décès, même au croisement le plus fin proposé** (par exemple, les effectifs de tumeurs du pancréas des femmes d'Occitanie de plus de 85 ans en 1980), **sont les mêmes que si la base de données n'avait pas été anonymisée.** Au niveau départemental, **des variations d'effectifs et de taux minimes** existent. Dans 99 % des croisements fins possibles avec la base de données, la différence d'effectifs en valeur absolue est inférieure à 2 ; dans 75 % des cas, elle est nulle.

De même, aucun croisement entre l'âge, le sexe et la cause de décès n'a été effacé. **Autrement dit, tous les décès d'une cause rare ou fréquente ont été maintenus, au bon âge, pour le bon sexe.** L'unique différence concerne la géographie de résidence.

Les données disponibles en *open data* sont donc toujours utilisables et riches d'informations pour la recherche médicale et scientifique, nous vous invitons cependant à la vigilance sur le commentaire de croisement fin, d'autant plus si les effectifs sont réduits (<20).

Ce document explicite les raisons poussant à anonymiser la base de données, il en décrit et illustre les conséquences minimes sur les données. Il propose également une présentation de la méthode utilisée ainsi que son implémentation concrète par le CépiDc.

Cette note méthodologique reprend en grande partie l'article présenté aux journées de la méthodologie statistique de l'Insee 2025 (Godet et Aubineau 2025).

Key points to remember

The database made available on the CépiDc open data website is anonymized to prevent the disclosure of the identities of deceased individuals and their causes of death, while maintaining the quality of the information provided.

To do this, the department of residence of 2 % of individuals in the database was changed. In the vast majority of cases, the department of exchange is part of the same region (residents of Guadeloupe, Martinique, Réunion, French Guiana, and Mayotte are exchanged among themselves). **This means that at the regional level, the numbers and rates for causes of death, even at the most detailed query** (for example, the number of pancreatic tumors for women in Occitanie over the age of 85 in 1980)**, are the same as if the database had not been anonymized.**
At the departmental level**, there are minimal variations in numbers and rates**. In 99 % of the possible detailed query on the database, the difference in numbers in absolute terms is less than 2. In 75 % of cases, the difference is null.

Similarly, no interaction between age, sex, and cause of death is deleted. In other words, all deaths from a rare or frequent cause was retained, at the correct age and for the correct sex. The only difference is on the department (and for the overseas departments, the region).

The data available in open data is therefore still usable and rich in information for medical and scientific research. However, we advice caution when interpreting contingency tables including the departement, especially if the sample size is small (<20).

This document explains the reasons for anonymizing the database and describes and illustrates the minimal consequences on the data. It also presents the method used and its practical implementation by CépiDc.

Table des matières

Introduction : pourquoi anonymiser les données disponibles en *open data* ?

Rappel sur le Centre d'épidémiologie sur les causes médicales de Décès (CépiDc)

Le Centre d'épidémiologie sur les causes médicales de décès (CépiDc) est le service de l'Inserm en charge de produire la statistique nationale des causes médicales de décès (Coudin et Robert 2024). Cette statistique repose sur l'analyse des textes écrits par les médecins et les infirmiers sur les certificats de décès. Leur collecte s'articule avec celle de l'état civil, ce qui assure l'exhaustivité de la collecte. Chaque décès sur le territoire donne lieu à la rédaction d'un certificat par professionnel de santé, document nécessaire pour fermer le cercueil et procéder à l'inhumation. Le texte inscrit est ensuite classé et codé d'après la classification internationale des maladies (CIM) de l'Organisation mondiale de la santé (OMS). La cause initiale de décès est ensuite déterminée en appliquant les règles de codage décrite dans la CIM.

Les données sont *in fine* diffusées à un niveau individuel dans le Système national des données de santé (SNDS) et tabulées sur le site internet du CépiDc. Les décès sont classifiés suivant l'année de décès, la cause initiale de décès décrite dans la shortlist européenne des causes de décès (European Commission. Statistical Office of the European Union. 2013) et suivant les caractéristiques du défunt (sexe, âge au décès et lieu de domicile). Des données agrégées au niveau national servent de support aux publications annuelles (Godet *et al.* 2025; Aubineau *et al.* 2025) qui accompagnent la diffusion des données.

Le cadre législatif encadrant les données sur les causes de décès

La diffusion des données sur les causes de décès en *open data* est régie par le règlement 223/2009 qui garantit, par le secret statistique, la protection des données confidentielles et interdit leur divulgation illicite. La diffusion des données sur les causes de décès en *open data* est aussi régie par l'article L2223-42 du Code Général des Collectivités Territoriales qui indique que les données du volet médical ont pour finalité une utilisation statistique pour des motifs de santé publique. **Le CépiDc n'a donc pas le droit de diffuser en *open data* des informations individuelles et réidentifiantes sur les causes de décès aux proches, aux médecins ou à tout autre public.**

Les risques portant sur la réidentification

Si l'on reprend la typologie des risques de divulgation proposée par (Bergeat 2016), il existe deux types de risques de divulgations pour les fichiers en *open data* : la divulgation d'identité et la divulgation d'attributs. Dans notre cas, nous pouvons en distinguer une de plus, qui est la divulgation par appariement, proche de la divulgation d'identité.

Divulgation d'identité et/ou d'appariement

« Il y a divulgation d'identité lorsqu'un individu statistique (entreprise, ménage, ou personne) peut être retrouvé dans un fichier, en retrouvant la ligne correspondante » (Bergeat 2016). Dans notre cas, cela signifie qu'un croisement année x sexe x âge ne concerne qu'un seul individu. Plusieurs cas de figure amènent à une divulgation de l'identité : l'individu isolé peut être retrouvé par un proche/voisin

ou l'individu isolé est suffisamment connu pour être identifié par n'importe qui. Cependant, du fait de la libre diffusion par l'Institut National de la Statistique et des Études Économiques (Insee) du « Fichier des personnes décédées », issu des données de l'état civil et comportant notamment nom, prénom, date de naissance, date de décès et commune de décès, le risque d'identification est élevé pour toutes les personnes de la base.

De plus, certaines informations sur les causes de décès relèvent du point 3 de l'article L311-6 du Code des relations entre le public et l'administration : « Ne sont communicables qu'à l'intéressé les documents administratifs : […] 3° Faisant apparaître le comportement d'une personne, dès lors que la divulgation de ce comportement pourrait lui porter préjudice ». Par exemple, les suicides peuvent rendre caduque une assurance-décès. Il faut donc également être en mesure d'empêcher des appariements avec des bases de données privées et très précises.

Les divulgations d'attributs

« Il y a divulgation d'attributs lorsque de l'information sensible sur un individu est révélée suite à la publication d'un fichier. […] En particulier, si un fichier de données individuelles contient des informations sensibles non perturbées, toute divulgation d'identité conduit à de la divulgation d'attributs pour ces variables. Il est possible qu'il y ait divulgation d'attributs sans qu'il y ait divulgation d'identité » (Bergeat 2016). L'application d'*open data* du CépiDc, sans méthode d'anonymisation, est hautement à risque de divulgation d'attributs dans plusieurs situations :

- Un croisement par année x classe d'âge décennal x sexe sur un département à faible mortalité donne facilement des effectifs de 1 ou 2 individus, permettant la réidentification et donc conduire au dévoilement de la cause de décès à une tierce personne ;
- Un croisement par année x classe d'âge décennal x sexe sur un département peut être constitué à plus de 80 % de la même grande cause de décès, ce qui permet, indépendamment de l'effectif du croisement, d'imputer avec quasi-certitude la cause de décès des individus connus et appartenant à ce croisement.

La méthode d'anonymisation utilisée par le CépiDc pour se prémunir des risques de divulgation lui permet de répondre à la double exigence d'anonymiser et de conserver l'utilité de la base de données publiée en *open data*. À ces deux exigences se rajoute celle de la cohérence : le CépiDc publie annuellement des effectifs de décès et des taux standardisés de mortalité au niveau national par causes de décès, sexe et trois grandes classes d'âge. Une requête sur le même champ, pour le même croisement, doit toujours donner le même résultat.

Quelle(s) conséquence(s) de l'anonymisation sur les résultats de vos recherches ?

Tout l'objectif et l'intérêt de l'algorithme utilisé est **d'anonymiser les données tout en gardant leur utilité**, c'est-à-dire que les données mises à disposition sont floutées (donc différentes de la vraie valeur) tout en restant proches des valeurs réelles. On précise par la suite les situations ou la différence est nulle, et des exemples réels parmi ceux où les données ont le plus de risque de s'écarter.

Le niveau de floutage dépend uniquement du niveau de précision géographique appliqué aux données.

Au niveau national (France métropolitaine et DROM, France métropolitaine, DROM)

Il n'y a aucune différence entre les tableaux que vous pourrez produire, quelle qu'en soit la précision en âge, sexe et cause de décès, si la variable géographique est « France métropolitaine et DROM » ou « France métropolitaine » ou « DROM ».

Au niveau régional

Si vous travaillez exclusivement sur les régions de France métropolitaine, il n'y a aucune différence entre les tableaux que vous pourrez produire quelle que soit la précision en âge, sexe et cause de décès.

Si vous travaillez sur les données de la Guadeloupe, la Martinique, la Guyane, la Réunion ou Mayotte, leur traitement a été le même que les départements métropolitains, la sous-partie suivante s'applique à eux.

Au niveau départemental, pour les requêtes les plus précises possibles

Si l'on considère l'ensemble des effectifs disponibles en *open data*, c'est-à-dire l'ensemble des croisements Année de décès × Département de domicile × Sexe × Classe d'âge décennale × Cause de décès au niveau le plus fin, **la distribution des différences absolues d'effectifs a une moyenne de 0,24 décès** (tableau 1). Pour 99 % des croisements la différence absolue est inférieure à 2 et le maximum est de 28 décès[1]. La différence est donc extrêmement faible et, les exemples ci-dessous l'illustrent, les variations annuelles sont préservées.

[1] Pour ce cas précis de 1982, il s'agit du chapitre des causes externes qui représentait 80% des plus de 70 décès d'un département × hommes × 1-24 ans. Tous les individus présents dans ce croisement avec ce chapitre ont été permutés, les causes de décès précises à l'intérieur du chapitre « causes externes » (homicides, accidents, suicide) ont été permutées pour ce département. En somme, pour un certain croisement sexe × âge × année, la répartition des causes de décès/âge pour un chapitre de cause a été « lissé » sur l'ensemble de la région, plutôt que de laisser la possibilité de dévoilement de la cause de décès de ces jeunes.

Tableau 1 - Distribution des différences absolues d'effectifs entre effectifs réels et effectifs anonymisés, par croisement fin entre 1979 et 2023

Moyenne	3ème quartile	9ème décile	99ème centile	Maximum
0,24	0	1	2	28

Lecture : Pour 75 % des croisements, la différence d'effectifs entre les données originales et les données anonymisées est nulle.
Champ : Personnes décédées entre 1979 et 2023, résident en France métropolitaine et DROM.
Source : Inserm-CépiDc.

Exemples réels de données parmi les plus permutées

Nous traitons de deux exemples[2] caractérisés par un niveau de permutation élevé : le département le plus permuté d'abord, puis les âges les plus sujets à la permutation (moins d'un an, 1-24 ans) en région de Provence-Alpes-Côte d'Azur. **Ces exemples représentent notamment des cas où la variation est la plus importante.**

Le département le plus permuté de France : les Hautes-Alpes

On s'intéresse donc au département avec le plus haut taux de permutation observé hors DROM, afin de vérifier que ces données restent exploitables même à un niveau de détail fin. Ces exemples permettront de répondre à la question : **est-ce que les données du département le plus permuté de France métropolitaine (en proportion) sont toujours utilisables ?**

Les Hautes-Alpes sont un département dont le territoire est entièrement situé en zone de montagne ; c'est le 4ème département le moins peuplé de France, représentant environ 3 % des décès et de la population de l'entièreté de la région PACA (tableau 2). Dans notre cas, ce déséquilibre démographique important au sein de la région PACA conduit à une forte élévation du taux de permutation des Hautes-Alpes par le mécanisme expliqué en 0.

[2] L'ensemble des données des graphiques sont présentées en moyenne glissante de 2 ans, afin de ne pas diffuser les données réelles précises.

Tableau 2 – Effectifs de population vivante et de décès annuels pour les Hautes-Alpes et la région Provence-Alpes-Côte d'Azur

	Hautes-Alpes		Provence-Alpes-Côte d'Azur	
Année	Population	Décès	Population	Décès
2015	140 916	1 410	5 007 977	49 948
2016	141 107	1 456	5 021 928	49 791
2017	141 284	1 388	5 030 890	51 076
2018	140 698	1 421	5 052 832	50 782
2019	141 220	1 467	5 081 101	51 515
2020	140 605	1 671	5 098 666	55 630
2021	(r) 140 976	1 572	(r) 5 127 840	58 179
2022	(r) 141 677	1 522	(r) 5 170 312	57 056
2023	(p)(r) 141 826	1 519	(p)(r) 5 194 349	53 713

(p) : Données provisoires — (r) : Données révisées

Source : CépiDc — Insee, Estimations de population Hautes-Alpes et Provence-Alpes-Côte d'Azur.

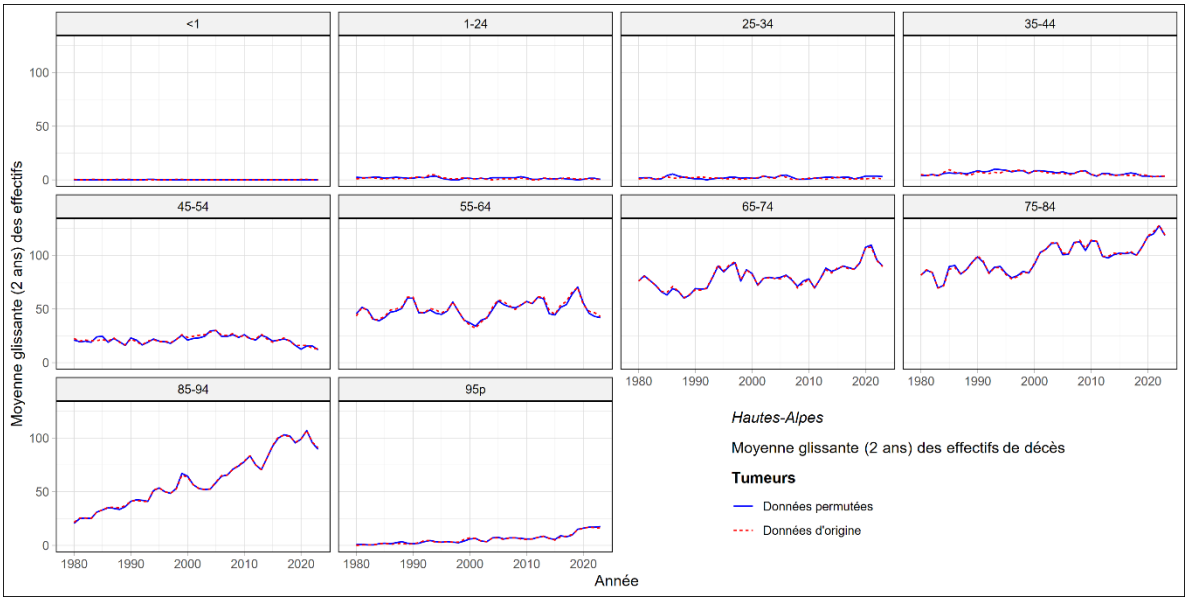
Champ : Personnes résident dans le département des Hautes-Alpes et dans la région Provence-Alpes-Côte d'Azur.

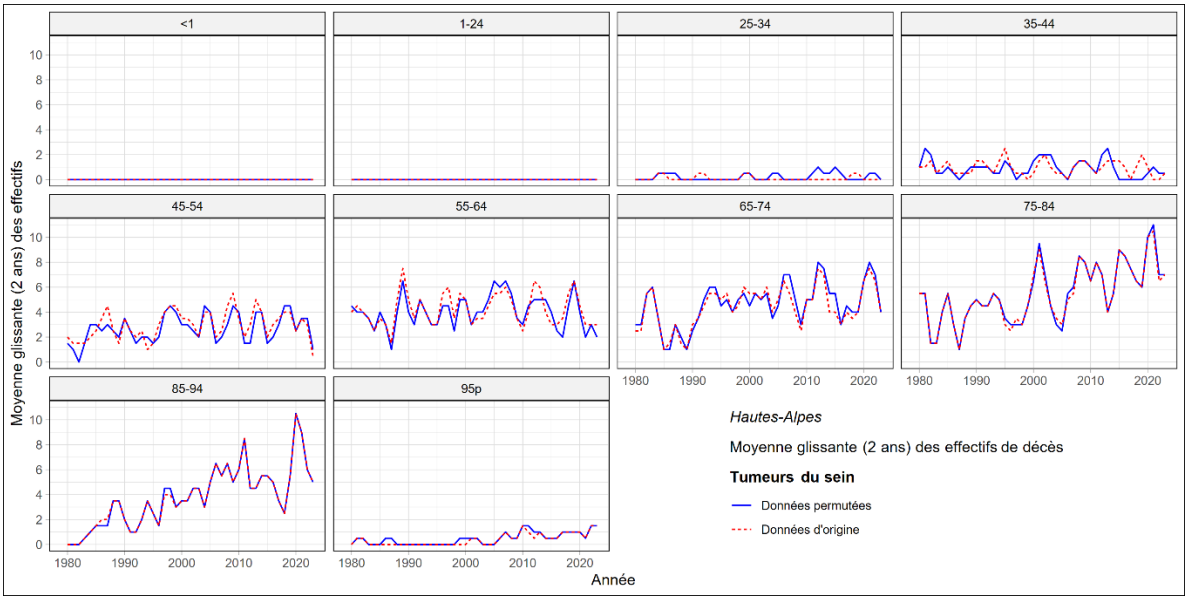
Lecture : En 2015, la région PACA compte 5 007 977 habitants dont 140 916 en Hautes-Alpes. Cette même année, 49 948 décès de résidents en PACA ont été recensés, dont 1 410 de résidents en Hautes-Alpes.

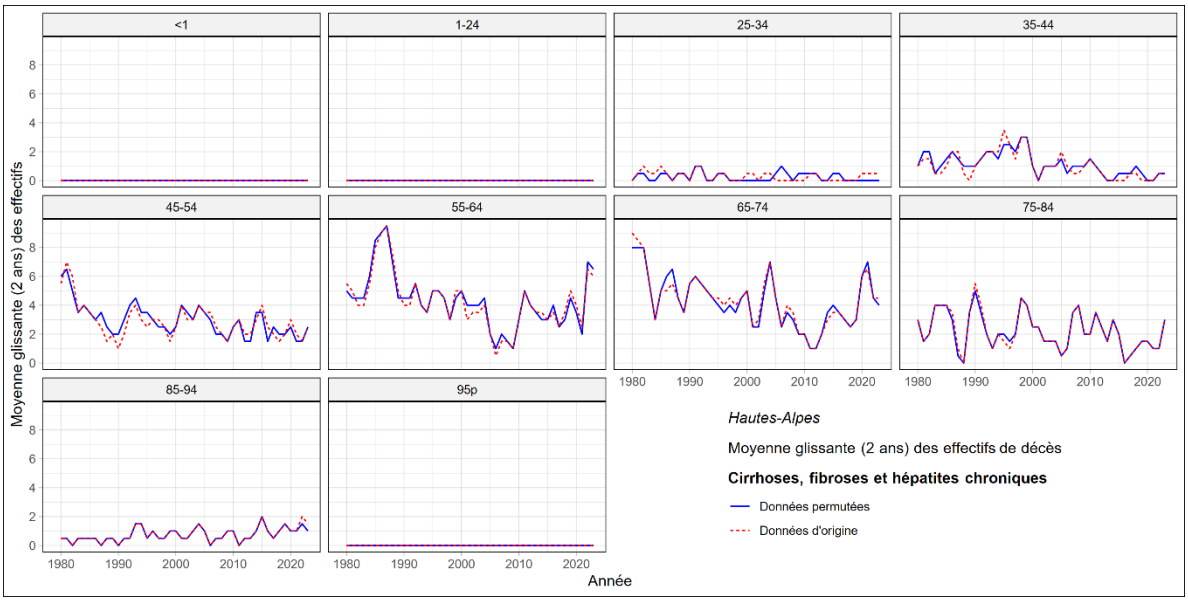
Pour les causes « médicales » (tumeurs, tumeurs du sein, cirrhoses et hépatites chroniques), la permutation introduit du bruit presque imperceptible malgré des effectifs parfois très faibles (graphique 1). À l'opposé, pour les causes de décès qui ne sont pas d'origine pathologique (appelées « causes externes » dans la classification internationale), le floutage est plus visible, tout en n'étant pas en décrochage par rapport aux variations des données d'origine. **Plus généralement, les données mises à disposition en *open data* ne permettent plus de suivre au niveau départemental les clusters de décès par une cause spécifique (accident de bus, concentration de cancers pédiatriques, terrorisme, ...) des plus jeunes**. C'est une bonne chose du point de vue de l'anonymisation mais constitue une contrainte pour les usagers cherchant à établir des statistiques d'accidents au niveau départemental[3].

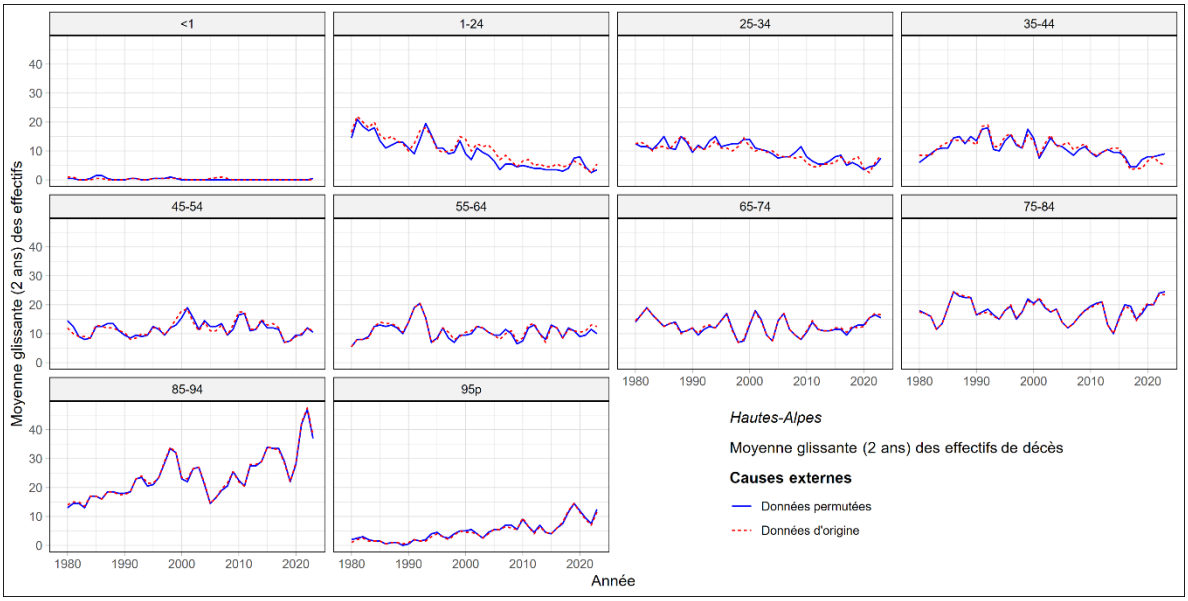
[3] Sur les causes externes, les données du CépiDc sont à mettre au regard d'autres données produites par des observatoires nationaux comme l'Observatoire national interministériel de la sécurité routière (ONISR) ou les services statistiques ministériels comme le Service statistique ministériel de la sécurité intérieure (SSMSI).

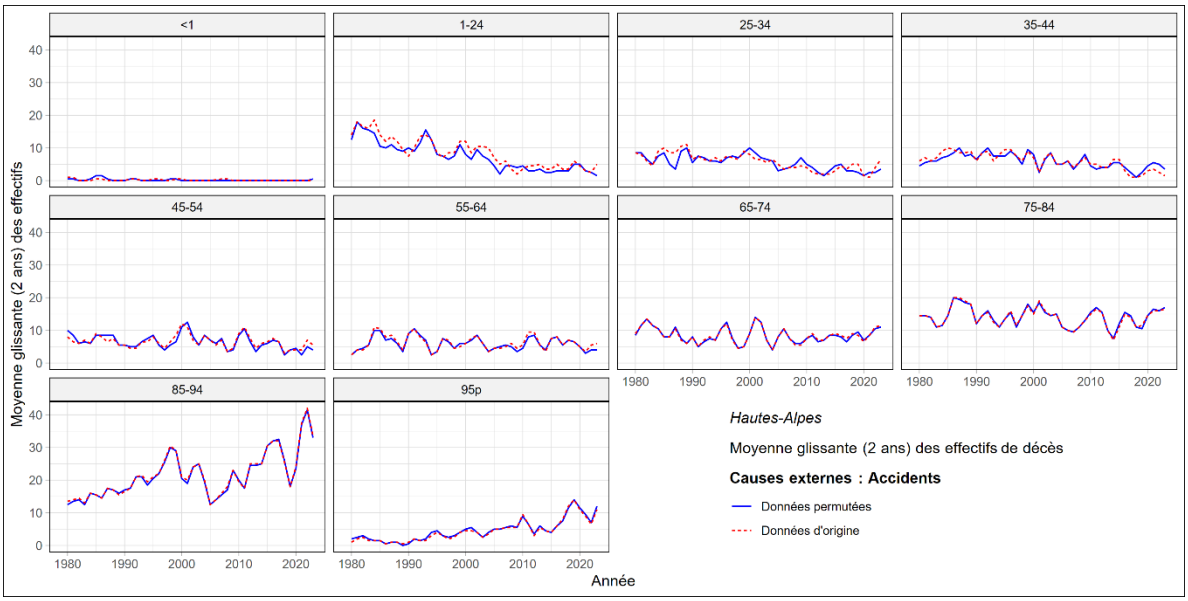
Graphique 1 – Moyenne glissante (2 ans) des effectifs par âge selon différentes causes pour les données d'origine et les données permutées dans les Hautes-Alpes

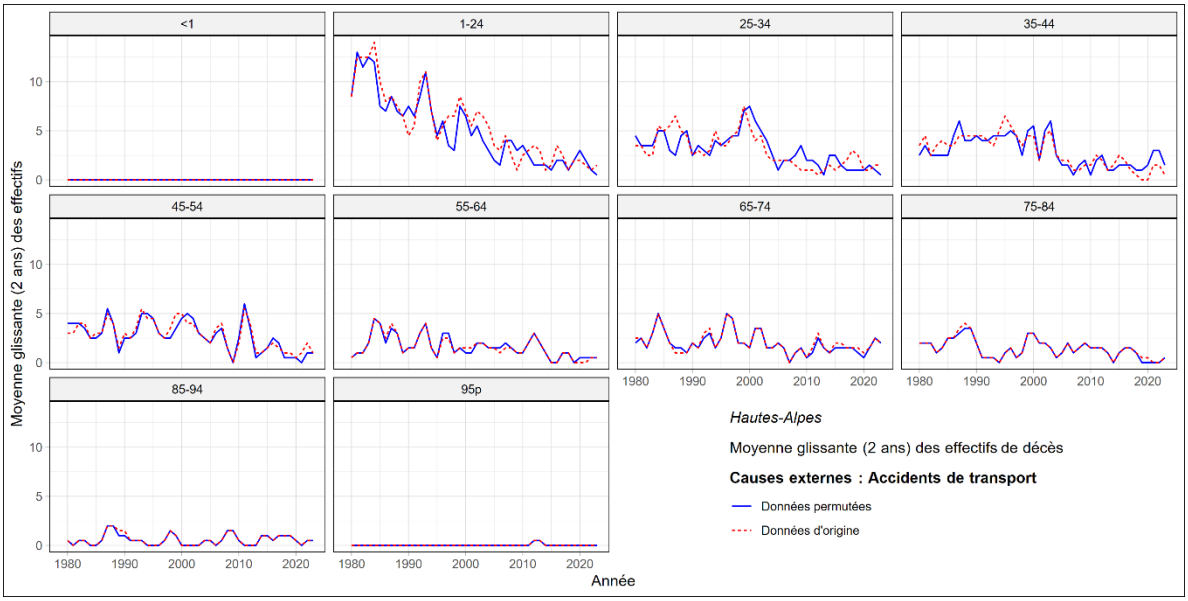





<1
1-24
25-34
35-44
45-54
55-64
65-74
75-84
85-94
95p
Moyenne glissante (2 ans) des effectifs
Année
Hautes-Alpes
Moyenne glissante (2 ans) des effectifs de décès
Cirrhoses, fibroses et hépatites chroniques
Données permutées
Données d'origine


<1
1-24
25-34
35-44
45-54
55-64
65-74
75-84
85-94
95p
Moyenne glissante (2 ans) des effectifs
Année
Hautes-Alpes
Moyenne glissante (2 ans) des effectifs de décès
Causes externes
Données permutées
Données d'origine





Les défunts jeunes dans la région Provence-Alpes-Côte d'Azur

Les décès des enfants de moins d'un an sont ventilés en trois principales causes de décès (Aubineau *et al.* 2025). Pour ces trois causes, malgré le niveau de précision des données (Année × Âge fin × Département × Grande cause de décès), les résultats obtenus à partir des données permutées sont systématiquement proches des données réelles (graphique 2). **Plus spécifiquement, la méthode de permutation ne créé pas d'artefacts comme des évolutions ou des tendances absentes des données d'origine**. Seules deux observations font exception, toutes les deux pour le département des Bouches-du-Rhône : on observe en 2022 pour les décès dus à une affection périnatale et en 1991 pour la mort subite du nourrisson un écart d'effectifs significatif entre donnée permutée et donnée d'origine. Cependant, dans les deux cas, la permutation reproduit la baisse observée dans les données réelles, mais dans une moindre ampleur.

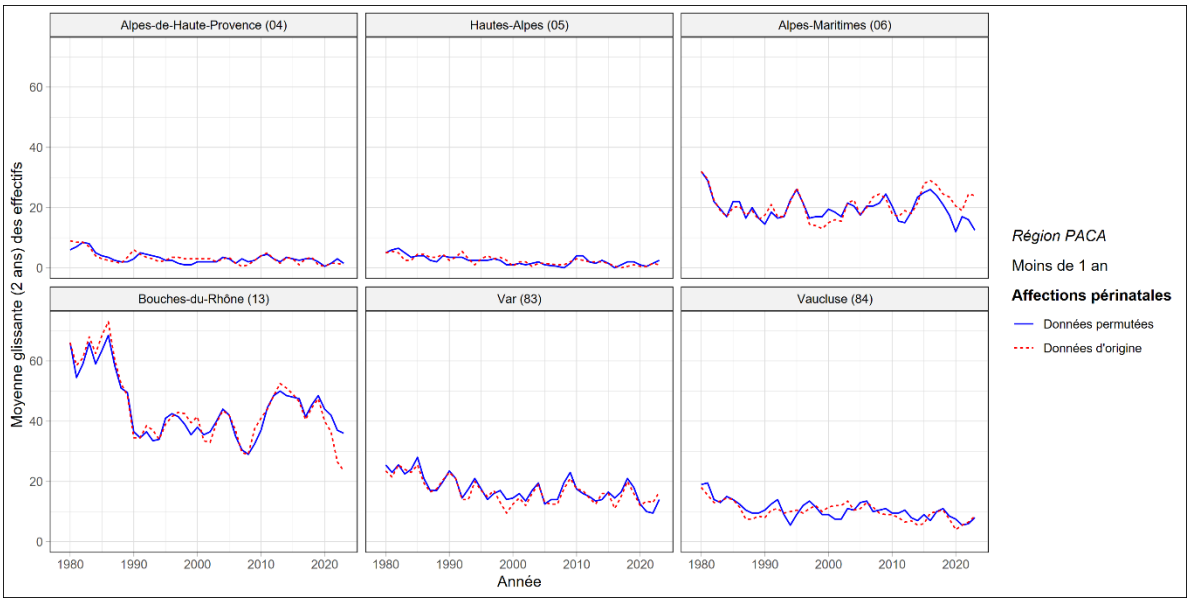
Pour les moins d'un an, catégorie d'âge la plus concernée par les permutations, le maintien des tendances sur plus de 40 ans, ainsi que la quasi-totale absence de différences significatives en effectifs

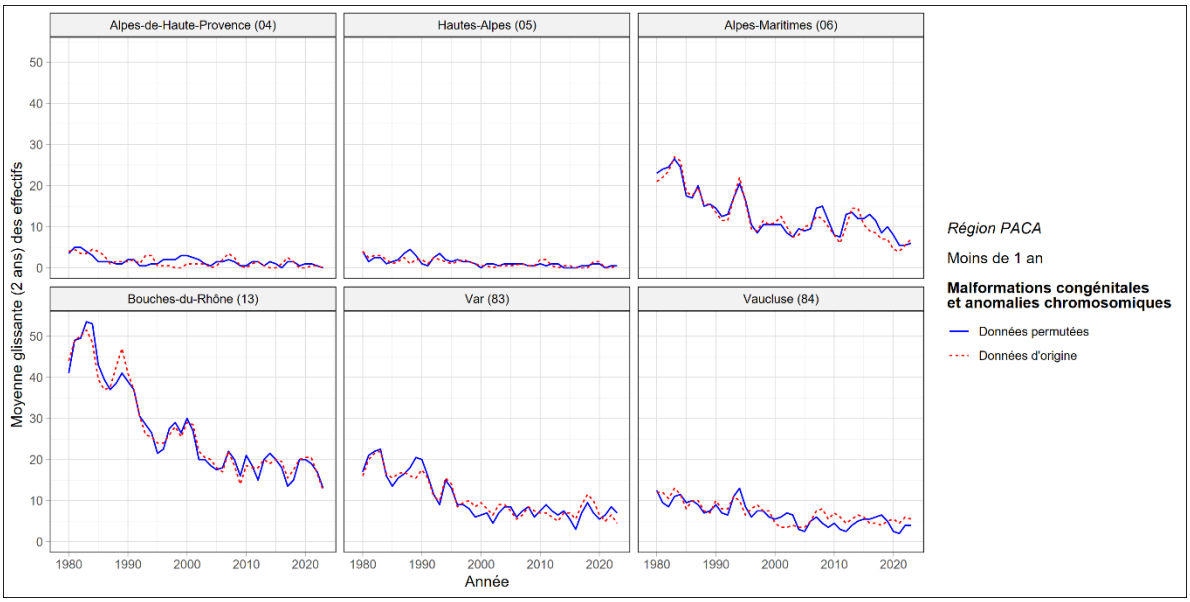
permettent de valider le fonctionnement de la méthode ainsi que le respect du critère d'utilité des données permutées.

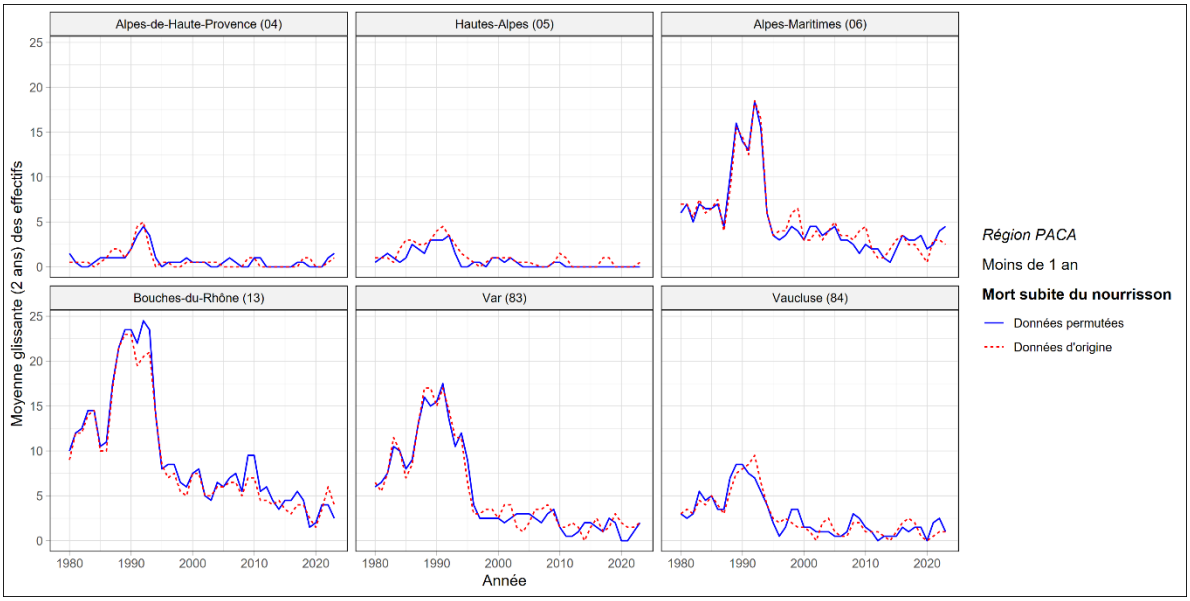
Concernant les personnes âgées d'entre 1 et 24 ans lors de leur décès, les causes choisies ici sont des causes appartenant aux deux chapitres les plus fréquents sur cette classe d'âge (graphique 2), elles sont à vocation illustrative.
Pour l'ensemble des causes présentées (tumeurs malignes, accidents de transport, suicide), les données permutées partagent les mêmes variations que les données d'origine, avec sporadiquement des différences plus marquées sur le niveau. **En particulier, on remarquera le suivi de tendance des données permutées sur le suicide pour les Bouches-du-Rhône, où la tendance à la baisse est parfaitement suivie de 1979 à 2023.**
On observe également une légère surestimation des décès par accidents de transport en Alpes-Maritimes, où la courbe des données permutées est systématiquement supérieure ou égale à celle des données d'origine sur l'ensemble de la période diffusée. Cela caractérise une situation où la répartition des causes de décès fines est inégalement répartie dans les départements d'une région. Dans ce cas précis, les décès par accidents de transport sont en général moins fréquents en Alpes-Maritimes que dans les autres départements de PACA. Comme attendu, la méthode lisse marginalement les différentes intrarégionales.

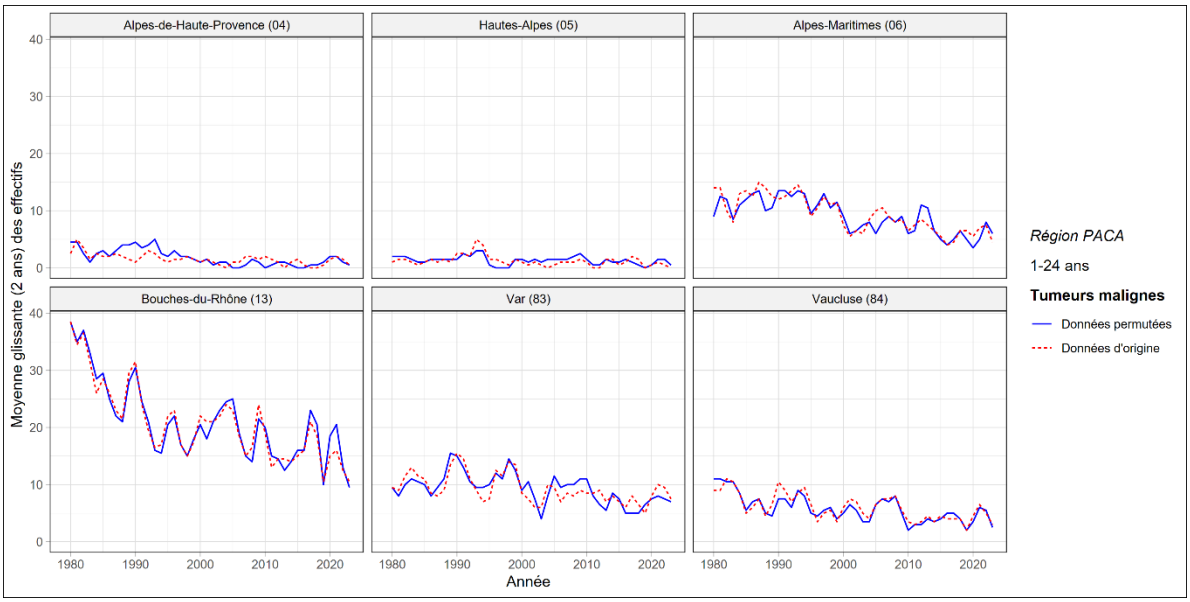
Pour les personnes âgées de plus d'un an et de moins de 24 ans, deuxième catégorie d'âge la plus touchée par la procédure d'anonymisation, le maintien des tendances sur plus de 40 ans ainsi que la quasi-totale absence de différences significatives en effectifs permettent de valider le fonctionnement de la méthode ainsi que le respect du critère d'utilité des données permutées.

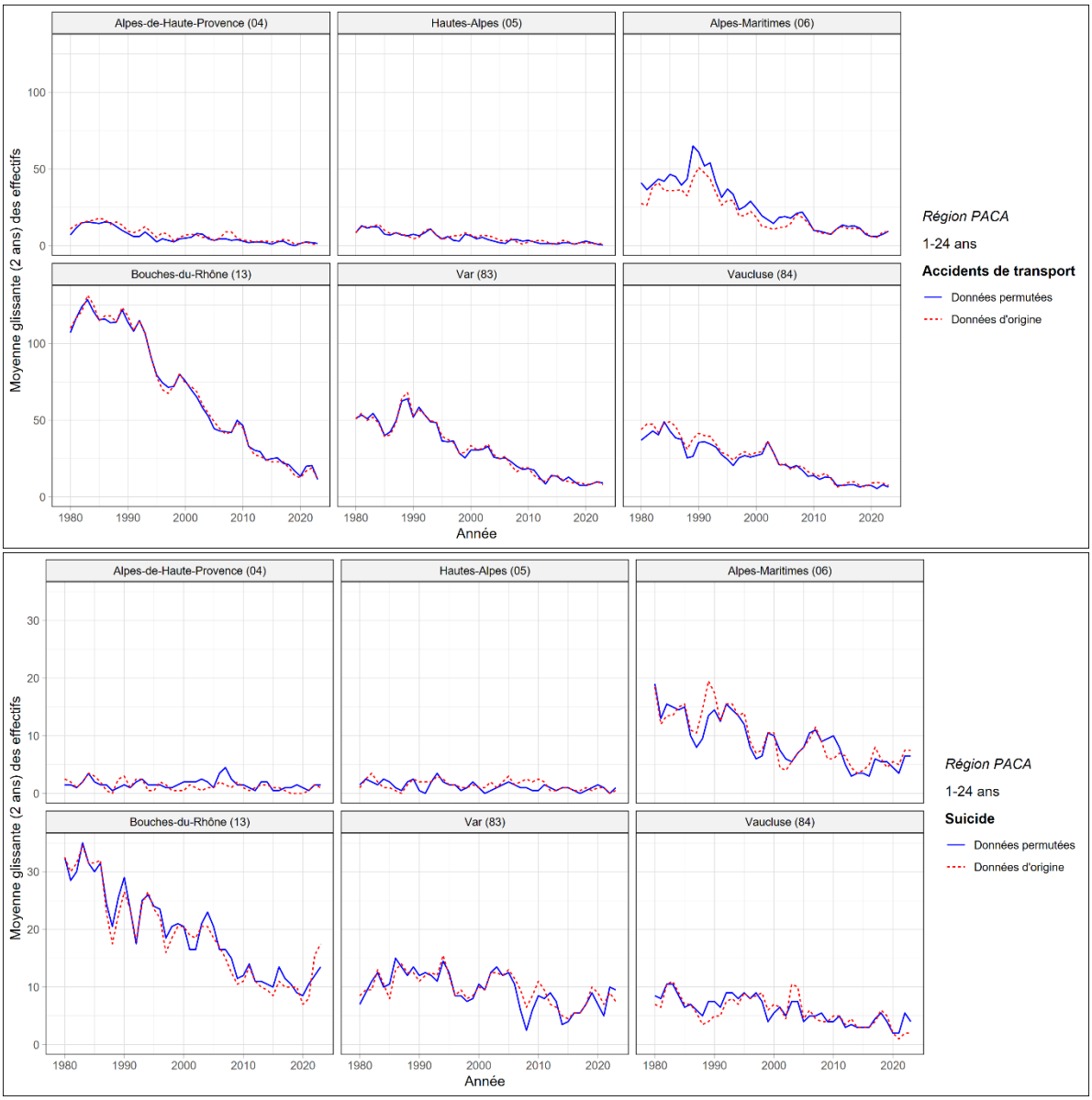
Graphique 2 – Moyenne glissante (2 ans) des effectifs des décès des moins d'un an par département de PACA selon différentes causes pour les données d'origine et données permutées




Alpes-de-Haute-Provence (04)
Hautes-Alpes (05)
Alpes-Maritimes (06)
Bouches-du-Rhône (13)
Var (83)
Vaucluse (84)
Moyenne glissante (2 ans) des effectifs
Année
Région PACA
Moins de 1 an
Malformations congénitales et anomalies chromosomiques
Données permutées
Données d'origine


Alpes-de-Haute-Provence (04)
Hautes-Alpes (05)
Alpes-Maritimes (06)
Bouches-du-Rhône (13)
Var (83)
Vaucluse (84)
Moyenne glissante (2 ans) des effectifs
Année
Région PACA
Moins de 1 an
Mort subite du nourrisson
Données permutées
Données d'origine


Alpes-de-Haute-Provence (04)
Hautes-Alpes (05)
Alpes-Maritimes (06)
Bouches-du-Rhône (13)
Var (83)
Vaucluse (84)
Moyenne glissante (2 ans) des effectifs
Année
Région PACA
1-24 ans
Tumeurs malignes
Données permutées
Données d'origine



Quels croisements sont les plus modifiés ?

Cas général : requête sur les plus de 24 ans ou tous âges

La méthode utilisée et décrite **Erreur ! Référence non valide pour un signet.** a pour objectif d'empêcher une réidentification des personnes décédées en échangeant leurs informations avec des personnes décédées la même année mais dans un autre département. Sont ciblées en priorité les individus particulièrement exposés du fait de leurs caractéristiques démographiques (par exemple, si un seul homme de 25-34 ans est décédé dans le Lot en 2023, le risque de la découverte de son identité, ainsi que de sa cause de décès est particulièrement important) ou du fait de la concentration des causes de décès pour leur catégorie (par exemple, si parmi les 10 décès des femmes de 55-64 ans du Jura en 2022, 9 étaient décédées d'une tumeur du sein, il est possible de déduire la cause de décès à 90 % de chance d'une femme décédée dans le Jura pour cette catégorie d'âge en 2022).

Si on retient ces deux exemples, qu'est-il arrivé à ces individus lors de l'anonymisation ?

L'homme isolé du Lot a été permuté avec un homme décédé la même année, âgé entre 35 et 44 ans (la contrainte étant entre 25 et 64 ans), possédant la même grande cause de décès (mais pas nécessairement le même sous-chapitre) et résidant dans le Gers (la contrainte étant de résider dans la même région, l'Occitanie). **Ainsi, tous les croisements au niveau régional sont restés inchangés,** car les échanges ont eu lieu à l'intérieur même de la région. Au niveau départemental cependant, les effectifs par âge ont pu être modifiés (à l'intérieur de la contrainte d'appartenance à la même grande classe d'âge : <1 an, 1-24, 25-64,65-84,85>), ainsi que les effectifs des causes de décès fines : si cet homme est décédé d'une tumeur du pancréas, l'effectif du chapitre « Tumeurs » est resté le même, mais il a pu être échangé avec un homme décédé d'une leucémie. **Ainsi, au niveau départemental, les données sont différentes pour les effectifs et taux par âge et/ou par cause de décès fine. Cela signifie qu'une requête par année x département x sexe x grande cause de décès donne des résultats inchangés.**

Pour ce qui est des 9 femmes âgées de 55-64 ans et décédées d'une tumeur du sein dans le Jura en 2022, elles sont échangées avec des femmes décédées la même année d'une tumeur dans la région Bourgogne-Franche-Comté. Ainsi, quelle que soit la raison pour laquelle une permutation a été réalisée, **tous les croisements au niveau régional restent inchangés. Au niveau départemental, les données sont différentes pour les effectifs et taux par âge et/ou par cause de décès fine. Les tableaux ne comportant ni l'une ni l'autre sont inchangés.**

Cas spécifique des jeunes : les moins de 1 an et les 1-24 ans

Certaines limites sont apparues lors du traitement du cas général, en lien notamment avec les causes de décès des moins de 1 an et des 1-24 ans qui sont très spécifiques. Les laisser être permutées avec celles de personnes plus âgées aurait pu vider un département des décès des plus jeunes, et donc de certaines causes de décès précises. Pour s'assurer du maintien d'une distribution des causes de décès (et des âges) par département, les moins de 1 an ne peuvent être permutés qu'entre eux, et les 1-24 ans qu'entre eux. Mettre une telle contrainte sur des âges avec un faible effectif de décès peut empêcher de trouver une observation dans la même région du même sexe, de la même tranche d'âge, et qui partagerait la même grande cause de décès. De plus, les causes de décès parmi les jeunes sont très homogènes et peu détaillées[4] donc forcer à rester dans la même grande cause de décès conduirait à ne pas changer les causes de décès. Aussi, pour ces deux classes d'âges, il est accepté de pouvoir permuter avec une observation du même sexe, de la même tranche d'âge, de la même région mais sans contrainte sur la cause de décès.

Concrètement, cela signifie que les causes de décès spécifiques aux moins de 1 an et aux 1-24 ans sont plus fluctuantes que pour les autres âges au niveau départemental. Cependant les causes de décès des moins de 1 an mises en *open data* sont peu précises, aussi peu d'information est perdue. **Pour les 1-24 ans, cette méthode conduit à lisser sur l'ensemble de la région les causes de décès**

[4] Aucune cause fine pour « 14. Certaines affections dont l'origine se situe dans la période périnatale », ni pour « 15. Malformations congénitales et anomalies chromosomiques ». Par définition « 16.1 Syndrome de la mort subite du nourrisson » offre peu de détail.

Cas spécifique des DROM

La Martinique, la Guyane, la Guadeloupe, la Réunion et Mayotte sont des régions composées d'un seul département. En clair, il n'est pas possible de permuter des individus d'un département avec ceux d'un autre département de la même région. Les résidents de la Martinique, la Guyane, la Guadeloupe, la Réunion et Mayotte décédés sont donc permutés à l'intérieur de la sur-région « DROM ». Cela signifie donc que le niveau géographique « DROM » possède les mêmes caractéristiques que dans le **Erreur ! Source du renvoi introuvable.** : **tous les croisements au niveau « DROM » restent inchangés. Au niveau départemental/régional, les données sont différentes pour les effectifs et taux par âge et/ou par cause de décès fine. Les tableaux ne comportant ni l'une ni l'autre sont inchangés.**

Méthode utilisée : Targeted Record Swapping

Le *Targeted Record Swapping* (permutation ciblée d'observations, TRS) est une méthode standard d'anonymisation appliquée à des données individuelles pour protéger les tableaux issus de ces données (Hundepool *et al.* 2024). En particulier, elle est utilisée lorsque les données diffusées ont des caractéristiques géographiques et que le nombre de tableaux produits est élevé ou hors du contrôle du producteur de données.

L'algorithme a été implémenté initialement en SAS par l'*Office for National Statistics* puis il a été recodé en C++ et interfacé en R avec la fonction `recordSwap()` du package `sdcMicro` (Templ *et al.* 2015). C'est cette fonction qui a été utilisée pour réaliser l'anonymisation.

Algorithme d'échanges d'observations

L'objectif de l'algorithme de *target record swapping* est d'identifier des groupes d'individus à risque de réidentification (critère de k-anonymité), de permuter l'emplacement géographique de l'ensemble d'entre eux avec d'autres individus partageant un certain nombre de caractéristiques, puis d'atteindre un niveau global de permutation défini en permutant des individus qui n'appartiennent pas à des cases à risque. Le fait de permuter des observations qui ne sont pas à risque ajoute un niveau de sécurité supplémentaire, car on ne peut déduire les groupes à risque d'une information externe : la différence observée peut-être dû à du risque de réidentification ou du hasard.

L'algorithme s'applique nécessairement à des données « hiérarchisées géographiquement », c'est-à-dire que l'observation appartient à des niveaux imbriqués dans l'espace (dans notre cas, Département ⊆ Région ⊆ Appartenir à la France Métropolitaine ou aux DROM[5]). Le résultat obtenu est une base de données permutée où les variables géographiques d'une partie des observations ont été échangées. Dans notre cas, les défunts d'un département *x* sont déplacés dans un département *y*. La méthode comprend quatre étapes, détaillées ci-dessous.

Étape 1 : Ciblage des observations à risque

La première étape consiste à identifier les observations ayant le plus besoin d'être permutées, ici les défunts ayant de grandes chances d'être réidentifiés.

Le critère d'anonymité utilisé par la fonction `recordSwap()` est la k-anonymité : pour chaque niveau géographique[6], si le croisement de caractéristiques quasi-identifiantes (niveau géographique × {sexe, âge, cause de décès, ...}) a un effectif total inférieur au seuil de k-anonymité (par exemple, moins de 5), alors toutes les observations de ce croisement sont à risque. Elles composent donc un premier groupe de cibles pour la permutation : 100 % des observations de ce groupe seront permutées. L'algorithme tentera alors de permuter ces observations avec d'autres au même niveau géographique. Ainsi, si un croisement à risque est identifié dans le Gard, l'algorithme cherchera à trouver des paires aux observations concernées dans un département de l'Occitanie ou, en cas d'échec, en France métropolitaine[7]. De même, si un croisement à risque est identifié en Guyane, l'algorithme cherchera à trouver des paires dans une autre région d'Outre-mer.

[5] Lorsqu'une référence est faite à une variable géographique, nous nous référons uniquement au lieu de domicile du défunt, jamais au lieu de décès.

[6] En commençant par le plus haut niveau hiérarchique. On identifie donc d'abord s'il y a des croisements à risque au niveau de l'appartenance à la France Métropolitaine, puis au niveau de la région, puis au niveau du département.

[7] S'il échoue (le croisement n'existe pas dans un autre département de l'Occitanie), il montera au niveau géographique supérieur : une paire sera cherchée dans une autre région de France métropolitaine.

La fonction `recordSwap()` n'accepte pas nativement d'autres critères que la k-anonymité pour identifier ces groupes à risque. Néanmoins

Étape 2 : Appariement (*matching*)

Le principe de l'étape d'appariement (appelé aussi *matching*) est de trouver, pour chaque observation du groupe ciblé, une correspondance (une observation « donneuse ») au même niveau géographique mais pas au même emplacement[8], qui présente des caractéristiques similaires[9], en privilégiant les autres défunts à risque. Une fois le donneur trouvé, il est sorti de la liste des donneurs potentiels. Il n'y a donc jamais deux permutations pour la même observation. Si aucune observation ne convient au niveau géographique recherché, l'observation ciblée n'est pas permutée.

Étape 3 : ciblage complémentaire

Une fois tous les croisements à risque identifiés ainsi que les observations « donneuses », l'algorithme vérifie si le taux global de permutation demandé est atteint *via* les effectifs appariés.
Soit s le taux de permutation global, N_{risk} l'effectif d'observations des cases à risque, $N_{donneur}$ l'effectif d'observations utilisées comme « donneuses »[10], N_{total} l'effectif total, alors l'algorithme s'arrête si :

$$\frac{N_{donneur} + N_{risk}}{N_{total}} \geq s \quad (1)$$

Généralement le taux global de permutation n'est pas atteint. L'algorithme vérifie alors le niveau de permutation par niveau hiérarchique décroissant, jusqu'à arriver au niveau géographique le plus précis, ici le département. Il ciblera alors aléatoirement des observations à permuter et recherchera une observation « donneuse », jusqu'à ce que (1) soit vérifié pour le niveau géographique supérieur, ici la région.
Si dans notre cas, une région a un taux de permutation attendu supérieur à s, avant ciblage complémentaire, cette étape supplémentaire ne sera pas réalisée pour cette région. Le niveau de permutation ne sera donc pas nécessairement égal à s pour chaque département, ni n'en sera nécessairement proche. Cela dépend essentiellement de la répartition des décès à l'intérieur de la région. Le tableau 3 présente l'exemple de la région Hauts-de-France en 2023 avec $s = 2\ \%$.

[8] Si la case dans laquelle se trouve l'observation ne respecte pas la k-anonymité au niveau du département, l'observation « donneuse » sera d'abord cherchée dans un département de la même région. Si la recherche est infructueuse, elle sera cherchée en France métropolitaine pour les régions de France métropolitaine
[9] Plusieurs contraintes de similarités peuvent être données, avec un ordre de priorité : si la recherche de donneur pour la première contrainte échoue, la suivante est utilisée, ainsi de suite jusqu'à ce qu'un donneur soit trouvé, ou que la recherche ait échouée.
[10] Ne sont pas comptées dans $N_{donneur}$ les observations qui font partie d'une case à risque et qui seront « donneuses » pour une observation d'une case à risque.

Tableau 3 – Répartition des décès en Hauts-de-France et pourcentage de permutation en 2023

Région/ Département	Part des décès (%)	Pourcentage de permutation (%)
Hauts-de-France	100	2,0
Aisne (02)	10,1	3,0
Nord (59)	40,7	1,5
Oise (60)	12,4	2,6
Pas-de-Calais (62)	26,5	1,8
Somme (80)	10,2	2,9

Lecture : En 2023 dans les Hauts-de-France, la Somme représente 10,2 % des décès et parmi ses défunts, 2,9 % ont été permutés.
Champ : Personnes décédées en 2023, résident en Hauts-de-France.
Source : Inserm-CépiDc

Cet exemple permet d'illustrer le fait que plus un département concentre les décès d'une région, moins son taux de permutation est élevé. Le Nord représentant 40,7 % des décès de sa région, un tirage aléatoire des défunts à permuter va mécaniquement beaucoup cibler ce département, qui va devoir trouver des paires ailleurs. Or ajouter une permutation à l'Aisne augmente beaucoup plus sa part de permutés que celle du Nord.

Étape 4 : Permutation (*swapping*)

Enfin, les informations géographiques sont permutées entre les individus appariés. Une observation étant remplacée par une autre, les tableaux uni-dimensionnels d'effectifs (sexe, âge, département, ...) sont donc toujours équivalents, avant et après permutation. Les variations apparaissent en augmentant le nombre de variables, en fonction des contraintes de similarité. Par exemple, si la variable sexe est présente chaque contrainte de similarité, les permutations se feront toujours à sexe constant, un tableau sexe × département sera le même avant et après permutation.

Implémentation par le CépiDc

L'algorithme est appliqué pour chaque année séparément. Ainsi de nouvelles données annuelles peuvent être mises à disposition sans modifier les données déjà en ligne. De plus, cela assure facilement le respect du pourcentage global d'anonymisation sur l'ensemble des années.

Définition des variables à risque (quasi-identifiants)

L'application de mise à disposition en *open data* du CépiDc permet de croiser pour chaque année les données sur les causes de décès avec le sexe, l'âge et le lieu de résidence géographique du défunt. Les trois formes les plus fines de ces variables sont désignées comme variables à risque : le sexe, la classe d'âge décennale[11] et le département de domicile. À ces trois variables démographiques s'ajoute une variable *ad hoc* assurant le traitement de la L-diversité.

Intégrer par la petite porte la L-diversité
Bien que la k-diversité soit la seule mesure d'anonymité proposée par la fonction `recordSwap()`, il est possible de tenir compte d'un critère de L-diversité. Pour cela il faut créer et introduire dans les variables à risque un identificateur d'individus appartenant à des cases homogènes. L'astuce ici est de remplir une variable (`risque_homogeneite` si l'observation est dans une case à risque d'homogénéité (Sexe × Catégorie d'âge décennale × Département de domicile), de s'assurer que

[11] Moins de 1 an, 1-24, 25-35, ..., 85-94, 95 ans et plus.

l'effectif de chaque modalité de cette variable soit inférieur à 4 par case à risque, puis d'introduire dans un autre département de la même région au moins 5 fois la même modalité sur une case similaire (ainsi, l'algorithme identifiera au niveau départemental[12] des observations à permuter à permuter car Année × Département de domicile × Catégorie d'âge décennale × Sexe × `risque_homogeneite` ≤ 4, alors que dans l'autre département qui a servi de pair, l'effectif est égal ou supérieur à 5).

Ainsi, en ajoutant la variable `risque_homogeneite` dans les variables quasi-identifiantes définies plus haut, on s'assure que les observations qui ne respectent pas une L-diversité à 80 % sont également permutées.

Définition des observations similaires

Un candidat à la permutation doit pouvoir trouver sa paire, sa « donneuse ». Les observations doivent partager, en ordre de hiérarchie de similarité :

1. Grande cause regroupée[13] × Région de domicile × Grande classe d'âge 13 × Sexe
2. Si aucune observation ne correspond alors : Grande cause regroupée × Indicatrice d'appartenance aux DROM × Grande classe d'âge × Sexe
3. Sinon : Région de domicile × Grande classe d'âge× Sexe
4. Sinon : Indicatrice d'appartenance aux DROM × Grande classe d'âge × Sexe

Sinon, le décès n'est pas échangé. Cette situation n'existe pas, l'ensemble des paires sont trouvées par le cas 1. et 2.

Ainsi, seule une paire partageant le même sexe peut être permutée. Cela permet de conserver la répartition sexe × cause de décès et de prévenir la création de décès chez les hommes dus à des complications de grossesse. De même, la similarité 1. ne peut jamais être prise par une observation résidant en DROM : toutes les régions d'outre-mer étant composées que d'un seul département, à l'intérieur de la région de domicile l'algorithme ne peut trouver une paire de permutation qui appartienne à un autre département. Pour illustrer, un défunt de la Guadeloupe ne peut donc être permuté qu'avec un habitant de la Martinique, la Guyane, Mayotte ou la Réunion.

Paramètres numériques

Pour chaque année disponible, les paramètres numériques en entrée de l'algorithme sont donc :

- Niveau de k-anonymité : 5. Les croisements des variables à risque qui comportent 4 observations ou moins ont un taux de permutation de 100 %, car ils ne respectent pas le critère de k-anonymité.

[12] Au niveau France métropolitaine/DROM ou au niveau de la région de domicile, il y a plus de 5 observations par case Niveau Géographique × Catégorie d'âge décennale × Sexe × `risque_homogeneite` (somme de la modalité dans la case à risque et dans le département « paire »).

[13] Correspondant aux chapitres de la shortlist d'Eurostat (18 chapitres : Maladies infectieuses et parasitaires, Tumeurs, ...) à deux exceptions à près :

i) **Les causes avec le moins de décès ont été regroupées dans une catégorie « autres »**, les chapitres suivants peuvent donc être échangés entre eux : Maladies du sang, Maladies de la peau, Maladies du système ostéoarticulaire, Maladies du système génito-urinaire, Complications de grossesse et d'accouchement, Certaines affections dont l'origine se situe dans la période périnatale, Malformations congénitales et anomalies chromosomiques.

ii) **Pour les décès d'enfants de moins d'un an**, dont les causes sont très concentrées sur quelques chapitres, ainsi que pour les décès de jeunes de 1-24 ans, on autorise l'échange sans contrainte sur les chapitres. Cela permet d'assurer, pour ces deux catégories d'âge, conjointement avec la règle d'appartenance à la même grande classe d'âge, que la distribution des âges et des causes de décès de décès fine reste similaire pour chaque département, et s'assurer de trouver une paire.

- Niveau de permutation sur l'ensemble des données : on choisit le niveau de 2 %. Ce niveau de floutage nous permet nécessaire pour introduire un doute raisonnable dans l'esprit des utilisateurs sur la possibilité ou non d'identifier les causes de décès d'individus. Mais aller au-delà de 2 % aurait conduit à des niveaux d'échanges considéré comme trop élevé dans les petits territoires.
- « Faux paramètre » — Niveau de L-diversité : 80 %. La variable `risque_homogene` est remplie pour les observations appartenant à un croisement dont plus de 80 % des décès concerne la même cause de décès.

Référence

Aubineau, Yann, A. Fouillet, Fanny Godet, Vianney Costemalle, et E. Coudin. 2025. « Grandes causes de mortalité en France en 2023 et tendances récentes ». *Journal of Epidemiology and Population Health*.

Bergeat, Maxime. 2016. *La gestion de la confidentialité pour les données individuelles*. Document de travail M2016/07. Insee.

Coudin, Elise, et Aude Robert. 2024. « Les statistiques sur les causes de décès ». *Courrier des statistiques*, 27.

European Commission. Statistical Office of the European Union. 2013. *Revision of the European Standard Population: Report of Eurostat's Task Force : 2013 Edition.* Publications Office. https://data.europa.eu/doi/10.2785/11470.

Godet, Fanny, et Yann Aubineau. 2025. « Anonymiser des données diffusées en open data par le CépiDc - Ou comment déplacer des cercueils ». *Journées de la méthodologie statistique de l'Insee*.

Godet, Fanny, Vianney Costemalle, Yann Aubineau, Anne Fouillet, et Élise Coudin. 2025. « Causes de décès en France en 2023 : des disparités territoriales ». *Études et Résultats*, n° 1342.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, et al. 2024. « Handbook on Statistical Disclosure Control ».

Templ, Matthias, Alexander Kowarik, et Bernhard Meindl. 2015. « Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro ». *Journal of Statistical Software* 67 (octobre): 1-36. https://doi.org/10.18637/jss.v067.i04.