PROBABILITES, STATISTIQUE ET CITOYENNETE : INSCRIRE LE DEVELOPPEMENT DU JUGEMENT CRITIQUE DU FUTUR CITOYEN DANS LE CADRE DES PROGRAMMES DE MATHEMATIQUES DE L'ENSEIGNEMENT SECONDAIRE

Philippe **DUTARTE**
IA-IPR de mathématiques
Académie de Créteil
philippe.dutarte@ac-creteil.fr

Résumé
Nous décrivons la demande institutionnelle des programmes de mathématiques du collège et des lycées en matière d'éducation du futur citoyen et son évolution ces dernières années. L'accent est notamment porté sur le développement du jugement critique, auquel l'enseignement des mathématiques doit participer et singulièrement celui des probabilités et de la statistique.
À l'appui de cet objectif nous prenons cinq illustrations particulièrement emblématiques, dans des situations expérimentées en classe.
– Peut-on croire un sondage ? Depuis la présidentielle française de 2002 jusqu'à celle des Etats-Unis en 2016, la fiabilité des sondages est interrogée mais ceux-ci demeurent incontournables.
– Cas de leucémies à Woburn : hasard ou pollution ? Un exemple de santé publique où la statistique joue le rôle de « lanceur d'alerte ».
– Une « preuve statistique » de discrimination : l'affaire Castaneda contre Partida où les probabilités s'invitent au tribunal.
– Coïncidences et pseudo-sciences : le cas de la « psychogénéalogie ». Des connaissances en probabilités et en algorithmique permettent de démasquer des impostures.
– Exploration de données massives : Python, avec sa bibliothèque Pandas, permet, dès la classe de seconde, le traitement de données assez massives. Il s'agit notamment, pour le futur citoyen, de pouvoir mieux comprendre le monde, comme celui d'*Airbnb*, ou d'assurer une vigilance active, comme pour l'analyse des « *Paradise Papers* ».

Mots clés
Statistique, Probabilités, Citoyenneté, Esprit critique, Sondage, Discrimination, Coïncidences, Pseudosciences, Big data, Python, Pandas.

1. LA DEMANDE INSTITUTIONNELLE ET SON EVOLUTION

La demande institutionnelle des programmes, dont ceux de mathématiques, pour une éducation du futur citoyen va croissant.
Parmi les programmes en vigueur à la rentrée 2017, ceux du lycée professionnel[1] sont particulièrement explicites. La première phrase des programmes de mathématiques, sciences physiques et chimiques est la suivante :

> « L'enseignement des mathématiques et des sciences physiques et chimiques concourt à la formation intellectuelle, professionnelle et citoyenne des élèves. »

Le domaine « statistique et probabilités » du programme contribue spécifiquement à cet objectif, notamment par son caractère interdisciplinaire. On lit ainsi dans le programme de la classe de seconde professionnelle :

> « Ce domaine [statistique et probabilités] constitue un enjeu essentiel de formation du citoyen. Il s'agit de fournir des outils pour comprendre le monde, décider et agir dans la vie quotidienne. (…). Leur enseignement facilite, souvent de façon privilégiée, les interactions entre diverses parties du programme de mathématiques (traitements numériques et graphiques) et les liaisons entre les enseignements de différentes disciplines. »

Le programme de mathématiques de la classe de seconde générale et technologique[2] indique les finalités suivantes, au premier rang desquelles un objectif plutôt citoyen :

> « Le programme de mathématiques a pour fonction :
> – de conforter l'acquisition par chaque élève de la culture mathématique nécessaire à la vie en société et à la compréhension du monde ;
> – d'assurer et de consolider les bases de mathématiques nécessaires aux poursuites d'étude du lycée ;
> – d'aider l'élève à construire son parcours de formation. »

Au collège, le socle commun de connaissances, de compétences et de culture[3], s'inscrit dans le cadre de la loi d'orientation du 8 juillet 2013 qui, en son article 13, pose le principe du socle commun :

> « La scolarité obligatoire doit garantir à chaque élève les moyens nécessaires à l'acquisition d'un socle commun de connaissances, de compétences et de culture, auquel contribue l'ensemble des enseignements dispensés au cours de la scolarité. Le socle doit permettre la poursuite d'études, la construction d'un avenir personnel et professionnel et préparer à l'exercice de la citoyenneté. Les éléments de ce socle commun et les modalités de son acquisition progressive sont fixés par décret, après avis du Conseil supérieur des programmes. »

L'article 4 précise :

> « (la formation scolaire) développe les connaissances, les compétences et la culture nécessaires à l'exercice de la citoyenneté dans la société contemporaine de l'information et de la communication. »

[1] BO spécial n° 2 du 19/02/2009.
[2] BO 30 du 23/07/2009.
[3] BO n°17 du 23/04/2015.

Le socle donne à la scolarité obligatoire l'objectif suivant :

> « la scolarité obligatoire poursuit un double objectif de formation et de socialisation. Elle donne aux élèves une culture commune, fondée sur les connaissances et compétences indispensables, qui leur permettra de s'épanouir personnellement, de développer leur sociabilité, de réussir la suite de leur parcours de formation, de s'insérer dans la société où ils vivront et de participer, comme citoyens, à son évolution. »

Le domaine 3 du socle est celui de « la formation de la personne et du citoyen ». Il y est affirmé que « *L'École a une responsabilité particulière dans la formation de l'élève en tant que personne et futur citoyen.* ». La prise en compte à parts égales des 8 composantes du socle dans l'évaluation du contrôle continu pour le DNB (Diplôme National du Brevet) fait que ce domaine 3 du socle représente 1/8 des points de contrôle continu pour l'examen, ce qui est assez considérable.

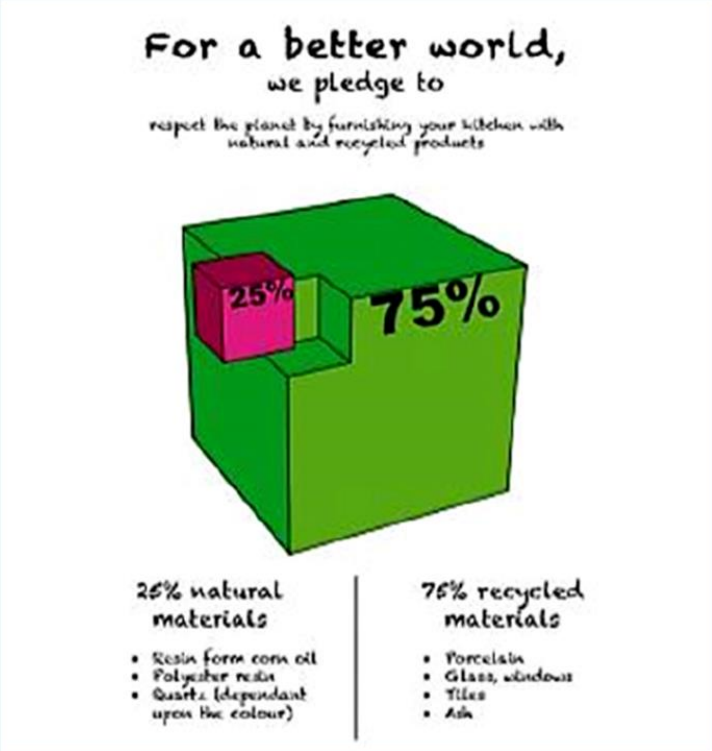
Le document d'accompagnement « Éléments pour l'appréciation du niveau de maîtrise satisfaisant en fin de cycle 4 », paru sur le site Eduscol, indique notamment comme « élément signifiant » du domaine 3 du socle, « *exercer son esprit critique, faire preuve de réflexion et de discernement* ».

> « En fin de cycle 4, l'élève qui a une maîtrise satisfaisante parvient notamment à utiliser les médias et l'information de manière raisonnée et responsable, à distinguer ce qui relève d'une croyance ou d'une opinion et ce qui constitue un savoir (ou un fait) scientifique. »

La figure 1 donne un exemple d'évaluation de l'esprit critique en mathématiques en fin de cycle 4.

ÉNONCÉ

La publicité ci-contre vise à exprimer que la proportion de produits naturels est égale à 25% de la production totale, et que celle de produits recyclés est égale à 75% de la production totale.



1. Quelle démarche a pu aboutir à cette représentation ?
2. Est-elle conforme à l'objectif visé ?

Figure 1 : exemple d'évaluation du domaine 3 du socle en mathématiques (Eduscol)

2. PEUT-ON CROIRE UN SONDAGE ?

Les sondages politiques constituent un domaine récurrent d'exercice de la culture statistique du citoyen et c'est un secteur que doit investir l'enseignement. L'un des « chocs » dans le domaine est l'exemple, déjà historique, du second tour de l'élection présidentielle française de 2002, dont on peut penser qu'il a joué un rôle dans l'évolution des programmes d'enseignement (voir Dutarte, 2011).

L'élection présidentielle française de 2002

Après la considération, en classe de seconde, des fluctuations des fréquences d'un caractère obtenues sur des échantillons aléatoires de taille *n*, on peut mettre en place la notion de « fourchette de sondage » et l'illustrer à propos de l'exemple suivant.
Lors du premier tour des élections présidentielles de 2002, le dernier sondage publié par l'institut B.V.A. , effectué sur 1 000 électeurs le vendredi 19/04/02, prévoyait :

Jacques Chirac	**19 %**
Lionel Jospin	**18 %**
Jean-Marie Le Pen	**14 %**

La surprise a été grande le dimanche 21/04/02 au vu des résultats, puisque Jean-Marie Le Pen figurait au second tour :

Jacques Chirac	**19,88 %**
Lionel Jospin	**16,18 %**
Jean-Marie Le Pen	**16,86 %**

Doit-on considérer que le dernier sondage B.V.A. était « faux » ?
Cet exemple a fait l'objet d'une analyse détaillée dans Dutarte et al. (2007).

L'élection présidentielle américaine de 2016

Plus près de nous, prenons l'exemple de la victoire, pour beaucoup inattendue, de Donald Trump à l'élection présidentielle américaine de 2016.
Un tweet malheureux du *Huffington Post* le 7 novembre 2016, veille de l'élection, annonçait, selon leur « modèle », la victoire d'Hillary Clinton avec une probabilité de 98,1 % (admirons la précision).

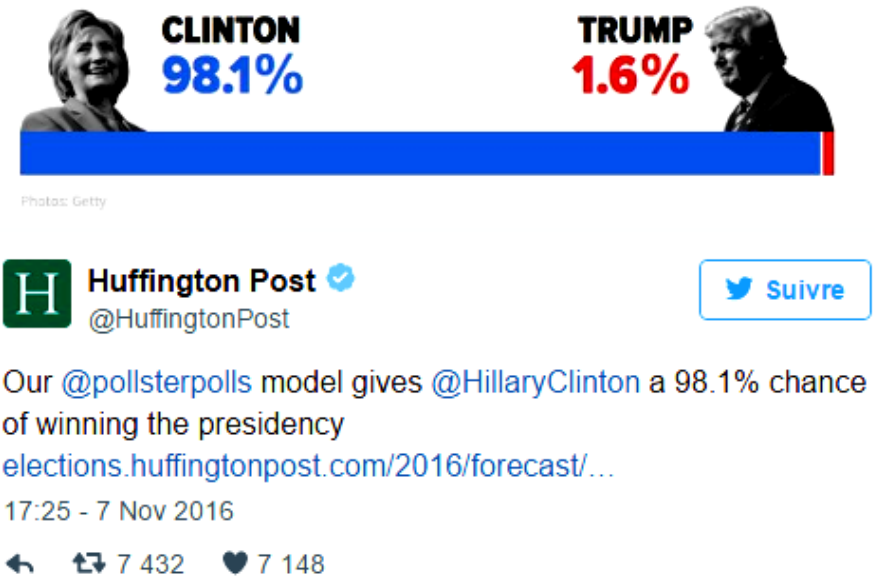


Figure 2 : tweet du Huffington Post annonçant une victoire écrasante d'Hillary Clinton

Une réponse d'un internaute dépité le 9 novembre : « Hé, les gars ! Peut-être que ce n'est pas un travail pour vous. »

Figure 3 : réponse au tweet de la fig. 2

La tendance des médias a été assez générale, même si certains ont été un peu plus prudents, comme ci-dessous le site FiveThirtyEight.com.

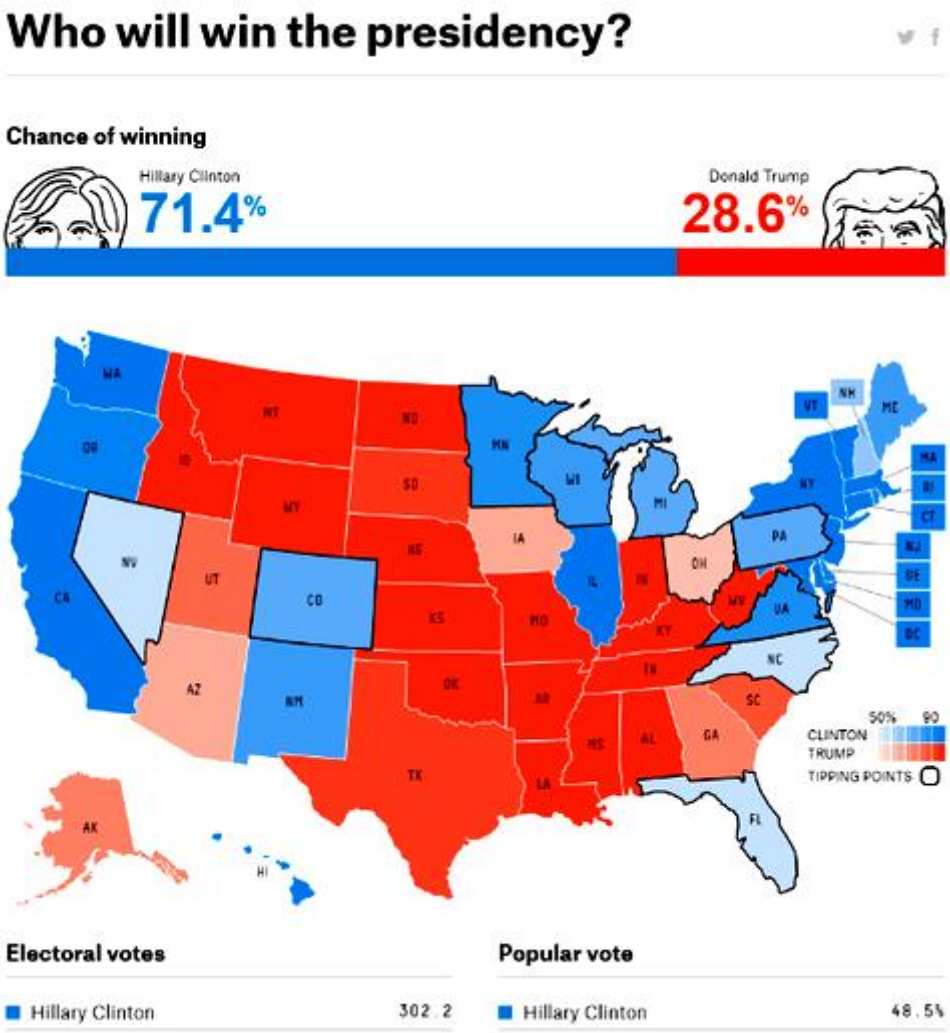


Figure 3 : prévisions de FiveThirtyEight.com

Si l'on prend cet exemple (figure 3) et que l'on le compare aux résultats du scrutin, on s'aperçoit que cela ne diffère que dans assez peu de cas.
La prise en compte des marges d'erreurs permettait de relativiser la vision des sondages comme le montre la figure 4.

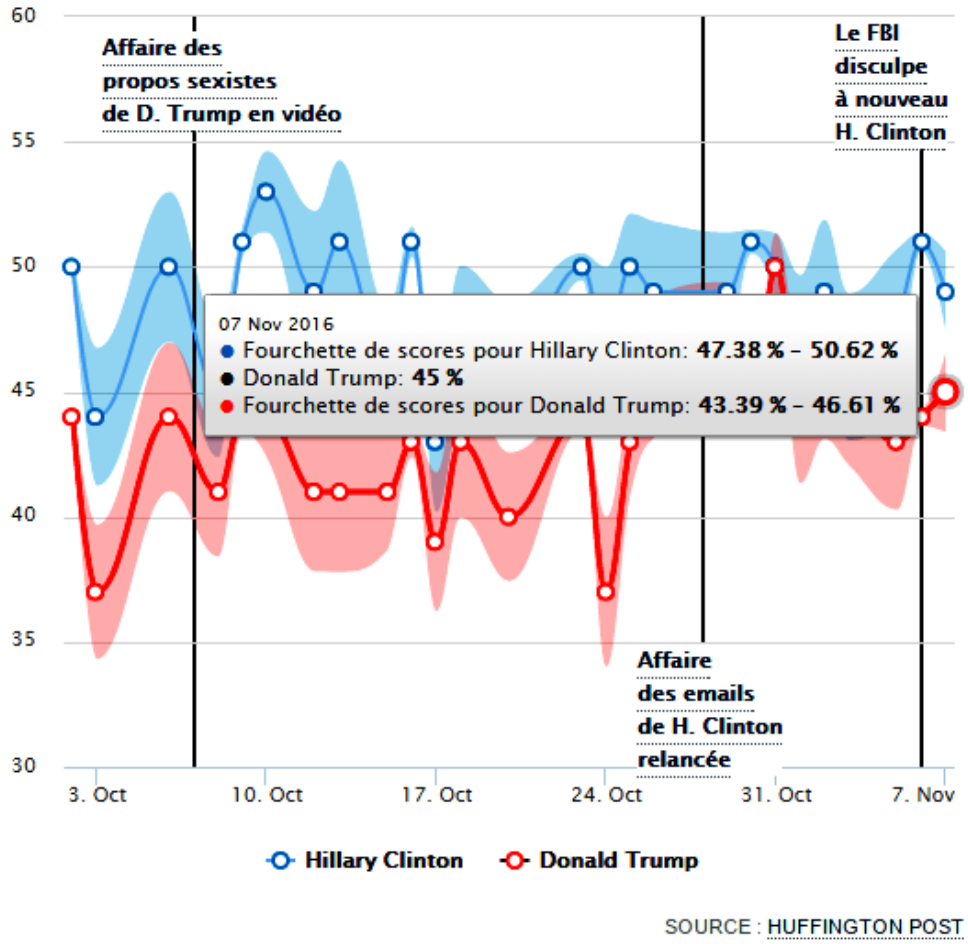


Figure 4 : fourchettes de sondage du 7/11/16 source Le Monde

Sur le nombre de votes, les sondages ont peu failli puisqu'Hillary Clinton a obtenu la majorité des suffrages. La difficulté principale de l'estimation tient au système électoral américain et aux fameux « swing states » (états-bascules) dont le basculement d'un côté ou de l'autre peut modifier, par le jeu des grands électeurs, le résultat de l'élection. Pour cette petite dizaine d'états, les sondages ont été assez nombreux et montraient bien une tendance très serrée la dernière semaine. On peut par exemple consulter les résultats des sondages, début novembre 2016, de quatre de ces états-bascules : la Caroline du Nord, la Floride, la Pennsylvanie et le Wisconsin (source : https://fr.m.wikipedia.org/wiki/Liste_de_sondages_sur_l'%C3%A9lection_pr%C3%A9sidentielle_am%C3%A9ricaine_de_2016).

Les sondages de la dizaine d'états-bascules de 2016 montrent que les deux candidats étaient très proches lors de la dernière semaine, avec quasiment une chance sur deux, pour chacun, de l'emporter. On a pu tenir le raisonnement selon lequel il était hautement improbable qu'un candidat remporte l'ensemble des états-bascules puisque $\left(\frac{1}{2}\right)^{10} \approx 0{,}001$. Un tel raisonnement pourrait expliquer les modèles prévoyant la victoire d'Hillary Clinton avec une très forte probabilité. C'était supposer qu'il y a indépendance de ces 10 événements, ce qui n'est bien sûr pas le cas.

Au Canada, les sondages sont publiés en mentionnant leur « marge d'erreur » et en indiquant que cette dernière ne vaut que 19 fois sur 20. Une façon assez pédagogique d'évoquer un niveau de confiance de 95 % dont on ferait bien de s'inspirer.

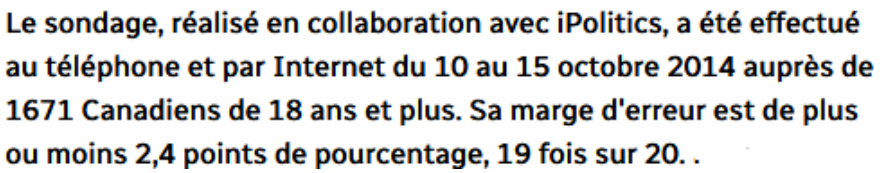


Figure 5 : mentions lors d'un sondage au Canada

Dès la classe de Seconde, voire en Troisième, on peut simuler les fourchettes de sondage avec un tableur. En Terminale, on peut indiquer que la marge d'erreur est voisine de celle donnée par la formule $1{,}96\sqrt{\frac{f(1-f)}{n}}$.

3. CAS DE LEUCEMIES A WOBURN : HASARD OU POLLUTION ?

La santé et l'environnement sont des domaines essentiels de la citoyenneté et l'exemple suivant, tiré de faits réels, a été expérimenté de nombreuses fois en classe de Seconde, en utilisant la simulation sur tableur (voir par exemple Dutarte et al., 2007).
Woburn est une petite ville industrielle du Massachusetts, au Nord-Est des Etats-Unis. Du milieu à la fin des années 1970, la communauté locale s'émeut d'un grand nombre de leucémies infantiles survenant dans certains quartiers de la ville. Les familles se lancent alors dans l'exploration des causes et constatent la présence de décharges et de friches industrielles ainsi que l'existence de polluants. Dans un premier temps, les experts gouvernementaux concluent qu'il n'y a rien d'étrange. Mais les familles s'obstinent et saisissent leurs propres experts.
Une étude statistique montre qu'il se passe sans doute quelque chose « d'étrange ».
Le tableau suivant résume les données statistiques concernant les enfants de Woburn de moins de 15 ans, pour la période 1969-1979 (Source : *Massachusetts Department of Public Health* et *Harvard University*).

Enfants entre 0 et 14 ans	Population de Woburn selon le recencement de 1970 n	Nombre de cas de leucémie infantile observés à Woburn entre 1969 et 1979	Fréquence des leucémies à Woburn f	Fréquence des leucémies aux Etats-Unis p
Garçons	5 969	9	0,001 51	0,000 52
Filles	5 779	3	0,000 52	0,000 38
Total	11 748	12	0,001 02	0,000 45

Compte-tenu de ces données, le hasard seul peut-il raisonnablement expliquer les fréquences observées à Woburn, considérées comme résultant d'un échantillon prélevé dans la population américaine ?

4. UNE « PREUVE STATISTIQUE » DE DISCRIMINATION : L'AFFAIRE CASTANEDA CONTRE PARTIDA

L'exemple suivant a été proposé dans Dutarte et al. (2007) et se situe dans le contexte de la discrimination raciale aux Etats-Unis, auquel les élèves sont particulièrement sensibles et qui constitue un objet important d'éducation à la citoyenneté. L'activité est ici présentée sous une forme assez « ouverte » d'analyse d'un texte juridique au niveau de la Terminale, mais peut être abordée, notamment par simulation, dès la classe de Troisième.
En 1976 au Texas, un accusé d'origine mexicaine conteste le jugement du tribunal au motif que la désignation des jurés est discriminatoire envers les Américains d'origine mexicaine. On analyse ici les arguments statistiques et probabilistes qui apparaissent dans l'attendu de la Cour Suprême des États-Unis.

> Attendu de la Cour Suprême des Etats-Unis (affaire Castaneda contre Partida)[4] :
> « Si les jurés étaient tirés au hasard dans l'ensemble de la population, le nombre d'américains mexicains dans l'échantillon pourrait alors être modélisé par une **distribution binomiale**… Etant donné que **79,1 %** de la population est mexico-américaine, le nombre attendu d'américains mexicains parmi les **870** personnes convoquées en tant que grands jurés pendant la période de 11 ans est approximativement **688**. Le nombre observé est **339**. Bien sûr, dans n'importe quel tirage considéré, une certaine fluctuation par rapport au nombre attendu est prévisible. Le point essentiel cependant, est que le modèle statistique montre que les résultats d'un tirage au sort tombent vraisemblablement dans le voisinage de la valeur attendue… La mesure des fluctuations prévues par rapport à la valeur attendue est l'**écart type**, défini pour la distribution binomiale comme la racine carrée de la taille de l'échantillon (ici 870) multiplié par la probabilité de sélectionner un américain mexicain (ici 0,791) et par la probabilité de sélectionner un non américain mexicain (ici 0,209)… Ainsi, dans ce cas, l'écart type est approximativement de **12**. En règle générale pour de si grands échantillons, si la différence entre la valeur attendue et le nombre observé est plus grand que deux ou trois écarts types, alors l'hypothèse que le tirage du jury était au hasard serait suspecte à un spécialiste des sciences humaines. Les données sur 11 années reflètent ici une différence d'environ **29** écarts types. Un calcul détaillé révèle qu'un éloignement aussi important de la valeur attendue se produirait avec moins d'**une chance sur 10^{140}**. »

La constitution des jurys est-elle faite au hasard ?
Signalons qu'au-delà des arguments mathématiques, peut être abordée la question du mode de constitution des jurys aux États-Unis.

5. COÏNCIDENCES ET PSEUDO-SCIENCES :

Les mathématiques, et singulièrement la statistique et les probabilités, constituent des atouts décisifs pour exercer sa rationalité, notamment pour se prémunir des pseudo-sciences. L'exemple suivant, inspiré d'un ouvrage de Jean-Paul Delahaye et Nicolas Gauvrit, a été présenté lors du séminaire « Sciences et jugement critique » de novembre 2017 de l'académie

[4] Source : *Prove It with Figures (Statistics for Social Science and Behavioural Sciences)* - Hans Zeisel, D. H. et D. Kaye - Springer 2006.

de Créteil (documentation sur le site de l'académie de Créteil) et travaillé dans le cadre du projet de recherche « Les lois du hasard » d'Alain Bernard et Caroline Ehrhardt (Bernard & Ehrhardt, 2017). Voici une présentation possible de cette activité en classe de Seconde, ainsi que des éléments de réponse.

> « Dans les années 1970, la psychologue Anne Ancelin Schützenberger développa une théorie d'inspiration psychanalytique nommée «psychogénéalogie». Selon cette théorie, un inconscient familial travaille si bien dans l'ombre qu'on peut contracter des maladies ou des troubles mentaux à certaines dates parce qu'un de nos ancêtres aurait lui aussi vécu quelque chose de remarquable à cette date. Disons tout de suite que l'idée présentée comme cela n'est pas absurde : on peut imaginer que quelqu'un commence une dépression le jour anniversaire de la mort de ses parents, par exemple. En revanche dans la théorie de Schützenberger, il peut s'agir de cas bien plus mystérieux. Ainsi, elle imagine qu'on peut déclarer un cancer le jour anniversaire de l'accident d'un grand-oncle, et cela même si nous ne savons pas qu'un tel accident a eu lieu.
> L'argument massue de la psychogénéalogie est nommé le « syndrome des anniversaires » : Schützenberger a en effet noté que, si l'on cherche bien, on finit souvent par retrouver des coïncidences de dates, bref des anniversaires communs entre événements.
> La thérapie psychogénéalogique consiste à rechercher au moyen d'une enquête généalogique les dates importantes concernant nos ancêtres (naissance, majorité, premier amour, maladie, accident, mort, etc.), en remontant aussi loin qu'il le faut pour qu'une de ces dates se trouve être celle d'un événement important pour nous (accident, début d'une dépression, déclaration d'une maladie, etc.). Le nombre de dates recueillies lors de l'enquête dépasse bien souvent la cinquantaine et parfois la centaine, ce qui laisse planer un doute sur la nature improbable des coïncidences. […]
> La question qu'on doit se poser est celle-ci : si nous prenons deux listes de dates (disons *n* et *m* dates), quelle est la probabilité qu'une date de la première liste soit la même qu'une date de la seconde ? Avec une centaine de dates concernant les ancêtres, et une dizaine nous concernant, la probabilité de collision est alors de 0,96 environ. Que deux dates coïncident est en réalité beaucoup moins étonnant que l'événement inverse. »
>
> Jean-Paul Delahaye, Nicolas Gauvrit – *Comme par hasard !* book-e-book 2012.

1. Implanter les fonctions suivantes sur Python.

```
import random
import matplotlib.pyplot as plt

def liste_dates(n) :
    # Liste aleatoire de n dates sans remise
    dates = random.sample(range(1,366),n)
    return dates
def coincidence(n, m) :
    # Recherche d'au moins une coincidence entre deux listes de n dates et m dates
    dates_moi = liste_dates(n)
    print(dates_moi)
    dates_ancetres = liste_dates(m)
    print(dates_ancetres)
    double = 0
```

```
    for d in dates_moi :
        if d in dates_ancetres :
            print(d)
            double = 1
    return double
```

2. Effectuer quelques expériences de recherche de coïncidences entre deux listes aléatoires de 10 dates et de 100 dates, en imprimant les listes.
Qu'observe-t-on ?
3. Représenter l'évolution de la fréquence de l'événement E : « il existe au moins une coïncidence entre une liste aléatoire de 10 dates et une liste aléatoire de 100 dates » lorsque l'on répète l'expérience du choix aléatoire des listes de dates.
Vérifie-t-on l'affirmation du texte ?
4. Combien de fois suffit-il de répéter l'expérience pour obtenir une estimation de la probabilité de E à 10^{-2} près au seuil de 95 % ? (On admet que compte-tenu de la fluctuation d'échantillonnage, la fréquence obtenue après la répétition de n expériences fournit dans environ 95 % des cas une estimation de la probabilité de E à $\frac{1}{\sqrt{n}}$ près.)

Nous fournissons ici des éléments de réponse montrant l'intérêt de cette activité dont l'originalité est qu'elle permet de faire intervenir des éléments d'algorithmique et de programmation en situation de développement de l'esprit critique.

2. Exemple d'exécution de l'expérience :

```
>>> coincidence(10, 100)
[309, 304, 8, 82, 4, 180, 97, 157, 217, 60]
[302, 207, 216, 307, 322, 342, 85, 245, 315, 231, 98, 79, 110
, 349, 318, 21, 3, 240, 287, 226, 41, 25, 182, 169, 305, 185,
153, 212, 177, 144, 183, 148, 159, 102, 71, 155, 10, 74, 95,
101, 156, 152, 272, 108, 129, 301, 97, 286, 244, 336, 361, 35
2, 180, 267, 221, 277, 4, 350, 242, 354, 170, 117, 264, 323,
14, 8, 55, 265, 122, 70, 67, 288, 294, 248, 88, 33, 18, 198,
124, 171, 306, 316, 268, 186, 2, 320, 58, 76, 356, 247, 321,
201, 127, 42, 49, 131, 120, 218, 78, 69]
8
4
180
97
1
```

Figure 6 : exemple d'exécution de la fonction coincidence

En renouvelant l'expérience on constate que la coïncidence est extrêmement fréquente.
3. On peut produire le graphique suivant pour 10 000 répétitions de l'expérience. Cela confirme bien une estimation de la probabilité de l'événement E à 0,96, comme annoncé dans le texte.

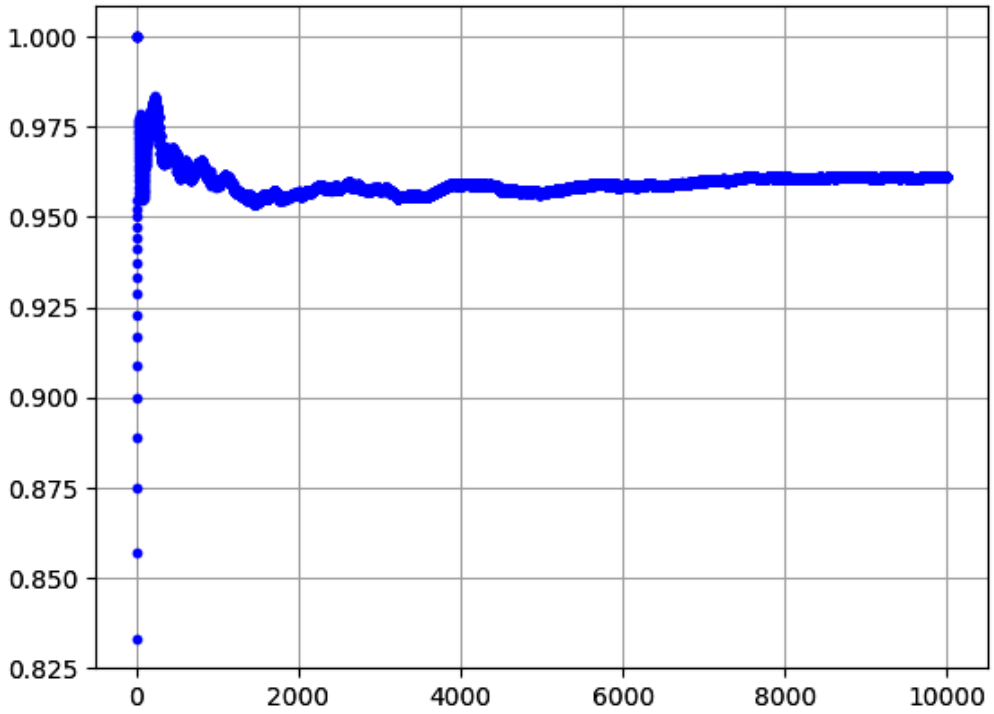


Figure 7 : stabilisation de la fréquence après 10 000 exécutions de l'expérience

4. Il suffit de répéter n fois avec $\frac{1}{\sqrt{n}} \leq 10^{-2}$ c'est-à-dire $n \geq 10\,000$.

6. EXPLORATION DE DONNEES MASSIVES : AIRBNB ET PARADISE PAPERS

Chaque jour, nous générons 2,5 milliards de milliards d'octets de données et plus de 90 % des données existantes ont été crées ces deux dernières années. L'exploitation de ces « big data », dont le potentiel économique est gigantesque, nécessite des techniques statistiques, mathématiques et informatiques en pleine évolution constituant un pôle important de recherche lié à la notion d'intelligence artificielle. Cependant plusieurs types de risques d'atteinte à la vie privée ou aux droits fondamentaux sont cités, notamment après les révélations d'Edward Snowden en 2013. Ainsi, environ 80 % des données personnelles mondiales seraient détenues par les GAFA (Google, Apple, Facebook, Amazon). On comprend le sentiment de défiance que peuvent provoquer ces technologies à l'égard des algorithmes et de l'intelligence artificielle comme en témoignent ces affiches photographiées à Londres fin 2017 (figure 8).

Figure 8 : dans les rues de Londres

Une attitude plus constructive consiste plutôt à renforcer l'éducation des futurs citoyens en matière d'analyse des données massives, permettant ainsi de mieux voir et appréhender le monde dans lequel on vit, comme dans l'exemple des locations Airbnb, voire de participer au contrôle de la vie citoyenne, comme dans le cas de l'analyse des « Paradise papers ». Un aperçu d'exploitation en classe de lycée de ces deux exemples est donné ici à l'aide du module Pandas de Python.

Airbnb

Le site insideairbnb.com[5] pose la question suivante : « Comment Airbnb est-il réellement utilisé et affecte-t-il les quartiers de votre ville ? ». Pour répondre à cette question, il est possible d'y télécharger les données Airbnb de nombreuses villes dans le monde dont Paris. On obtient pour Paris (avril 2017) un fichier csv de 56 450 lignes, correspondant chacune à une location, pour 12 variables étudiées, dont le prix, la disponibilité, le nombre d'avis et les coordonnées géographiques.
L'étude de la disponibilité des locations, en jours par an, montre par exemple que 64 % des locations sont disponibles plus de 120 jours par an.

[5] Ce site a été créé par Murray Cox, écrivain et informaticien indépendant se qualifiant de « data activiste » (« activiste des données »).

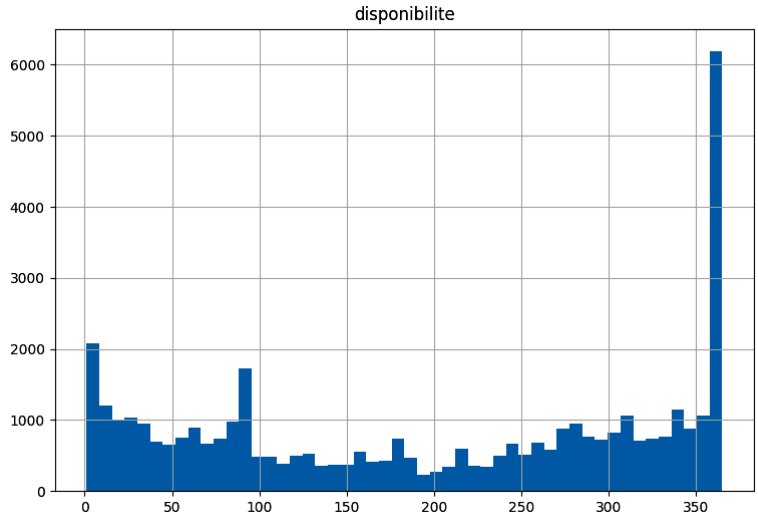


Figure 9 : disponibilité, en jours par an, des locations Airbnb de Paris (2017)

Une visualisation des données par le module Folium de Python (modèle « carte de chaleur ») permet d'illustrer la concentration des locations dans certains quartiers.

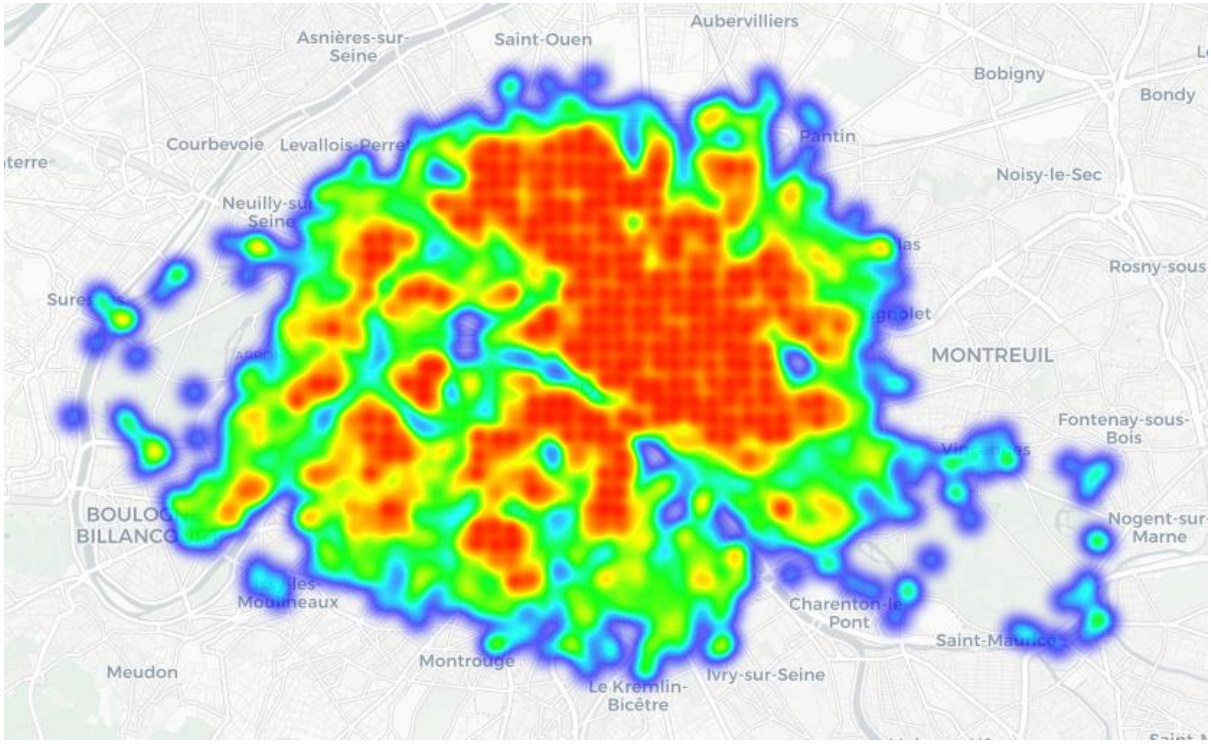


Figure 10 : concentration des locations Airbnb de Paris (2017)

Paradise Papers

Les données des « Paradise Papers », publiées par le Consortium international des journalistes d'investigation en novembre 2017, concernent des investissements « offshore » (i.e. un investissement de capital dans un pays fiscalement intéressant)[6]. On peut obtenir sur le site

[6] On peut à ce propos consulter les articles suivants du journal Le Monde : 05/11/2017 *Les « Paradise Papers » : nouvelles révélations sur les milliards cachés de l'évasion fiscale* ou 14/02/2018 *« Paradise Papers » : des dizaines de milliers de sociétés offshore rendues publiques dans la « Offshore Leaks Data Base »*.

kaggle.com[7] des fichiers csv correspondant à ces données. Nous avons exploré avec Python quatre fichiers nommés entity, officer, address et edges. Le fichier entity.csv, de dimension 24 957 × 18, correspond aux compagnies offshore jouant le rôle d'écran dans un paradis fiscal. On constate que, pour l'essentiel, les paradis fiscaux représentés dans les « Paradise Papers » sont les îles Caïmans et les Bermudes.

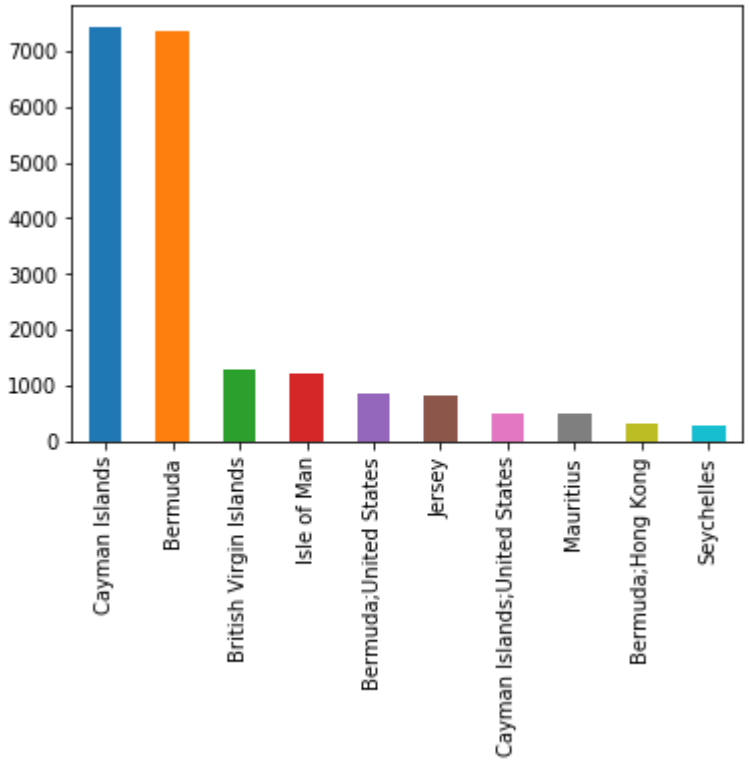


Figure 11 : répartition des paradis fiscaux des « Paradise papers »

Le fichier officer.csv, de dimension 77 012 × 18, correspond aux exécuteurs, c'est-à-dire aux entreprises ou particuliers donneurs d'ordre pour des « clients » souhaitant échapper au fisc de leur pays. À la ligne 1185 apparaît « The Duchy of Lancaster », le domaine privé de la reine d'Angleterre. La France apparaît en vingtième position des pays exécuteurs les plus cités.

[7] Site d'une start-up californienne organisant des compétitions en science des données.

```
>>> pays_executeurs[:20]
United States                  17303
United Kingdom                  4298
Hong Kong                       3262
China                           3050
Bermuda                         2955
Cayman Islands                  1925
Bermuda;United Kingdom          1362
Canada                          1226
China;Hong Kong                  981
Switzerland                      969
British Virgin Islands           919
Singapore                        789
Jersey                           651
Taiwan                           645
Australia                        640
Japan                            570
Isle of Man;United Kingdom       508
Ireland                          497
Isle of Man                      489
France                           422
Name: n.countries, dtype: int64
```

Figure 12 : principaux pays exécuteurs dans les Paradise papers

Le fichier address.csv, de dimension 59 228 × 18, donne la localisation des « clients », c'est-à-dire des commanditaires, ceux à qui profite la fraude. La France apparaît ici en dix-huitième position.

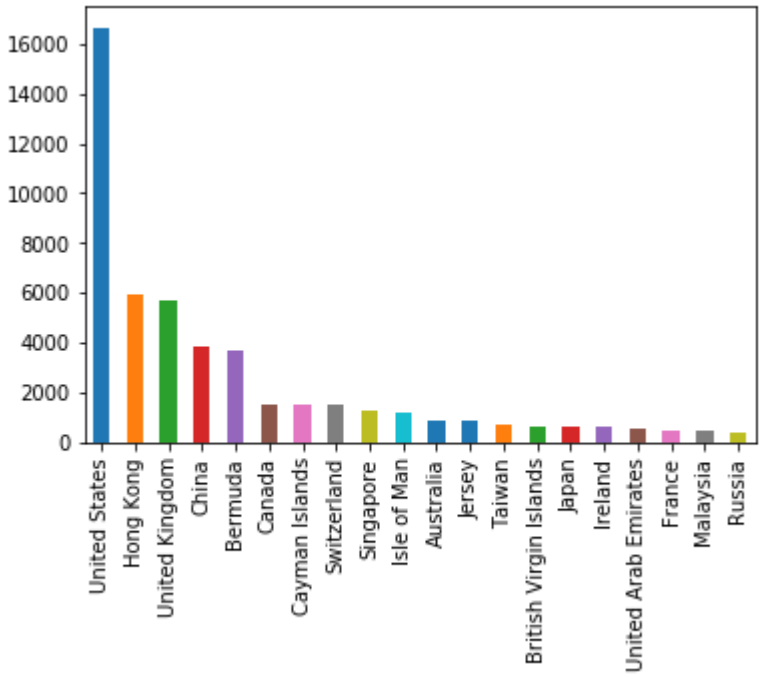


Figure 13 : principaux pays des clients des Paradise papers

On peut, en utilisant le module Folium, représenter les principaux pays impliqués par un cercle de rayon[8] proportionnel au nombre de fois qu'ils sont cités dans les trois fichiers des entités, des exécuteurs et des clients (figure 14).

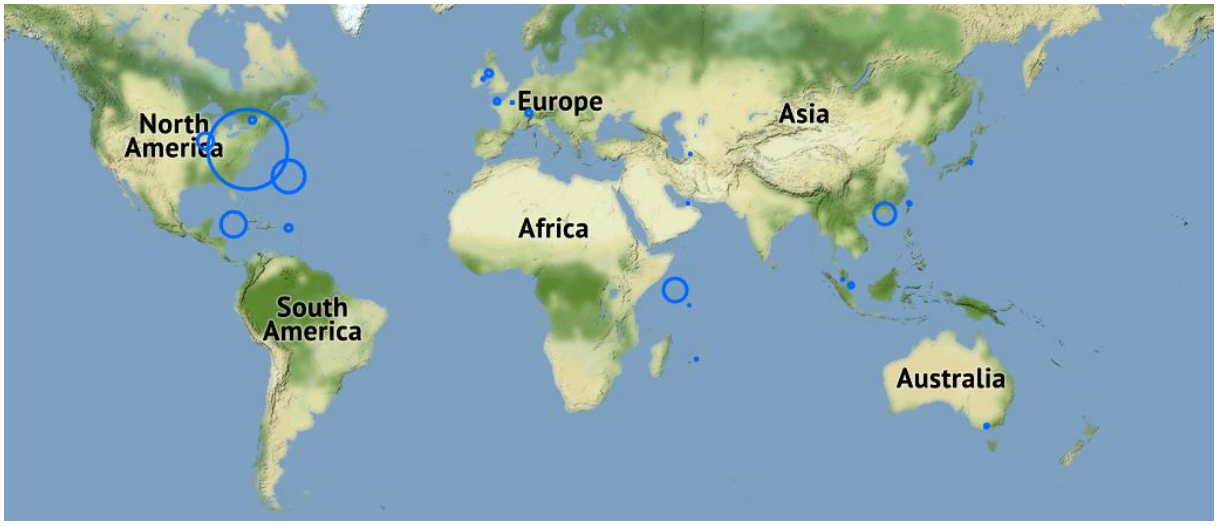


Figure 14 : localisation des principaux pays impliqués dans les Paradise Papers

Le fichier edges.csv, de dimension 364 456 × 7, fournit les liens existant entre les différents acteurs. Un algorithme peut alors permettre d'étudier certains réseaux et de les illustrer. Cet algorithme peut être élaboré avec les élèves de lycée avec plus ou moins d'autonomie selon le niveau de classe et de connaissances en programmation en langage Python.

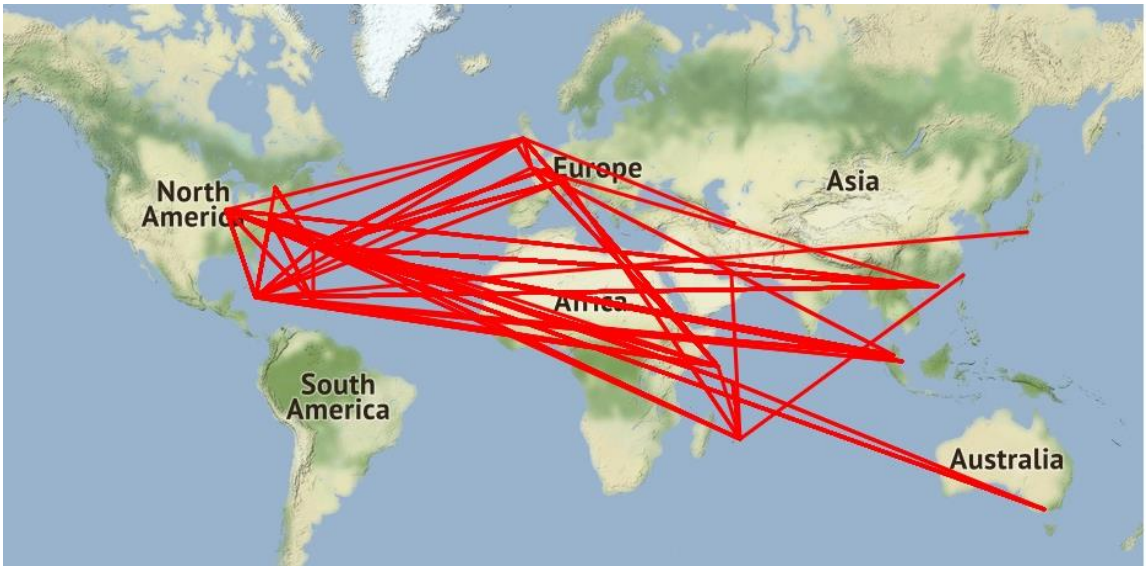


Figure 15 : représentation des principaux liens dans les Paradise papers

[8] Au risque d'une confusion : c'est plutôt l'aire du disque qui est perçue.

CONCLUSION

Les programmes de mathématiques font une part croissante à l'éducation du citoyen, notamment dans le cadre du socle commun. Les exemples développés montrent la place de premier ordre qu'occupe dans ce cadre l'enseignement de la statistique, des probabilités, de l'algorithmique et de la programmation, enseignement qui, lui même, a pris de l'importance dans les programmes de mathématiques. Le développement des « big data » (« mégadonnées »), du rôle de l'algorithmique et de l'intelligence artificielle dans notre quotidien offre de nouvelles perspectives d'apprentissage pour « armer » le futur citoyen des connaissances nécessaires à son jugement critique et pour qu'il puisse jouer un rôle actif et éclairé dans la société.

REFERENCES BIBLIOGRAPHIQUES

BERHOUET, J., DUTARTE, P., & GLEBA, F. (2017). Statistique, probabilités et jugement critique. In *Actes du séminaire académique Sciences et jugement critique*, Académie de Créteil 2017, maths.ac-creteil.fr.

BERNARD, A., & EHRHARDT, C. (2017). *Les lois du hasard : enjeux mathématiques, historiques, citoyens*. In T. Barrier & C. Chambris (Eds.), *Actes du séminaire national de l'ARDM de l'année 2017*. Paris : IREM de Paris.

DELAHAYE, J.-P., & GAUVRIT, N. (2012). *Comme par hasard !* Book-e-book.com.

DUTARTE, P., DELZONGLE, F., MAATI, H., CARDINAL, J.-P., COUPRY, A., & DHERISSARD, S. (2007). *Statistique et citoyenneté, le citoyen face au chiffre*. Brochure 135 de l'IREM de Paris Nord.

DUTARTE, P. (2011). Évolution de la pratique statistique dans l'enseignement du second degré en France. *Statistique et Enseignement*, *2*(1), 31-42.

ZEISEL, H., & KAYE, D. (2006). *Prove It with Figures Empirical Methods in Law and Litigation*. Springer.