Gestion des Données Ouvertes en Agriculture et Nutrition

Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environnemental Research (WUR), AgroKnow, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).

GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]

MODULE 2: UTILISATION DES DONNÉES OUVERTES

LEÇON 2.3 : Analyse et visualisation des données

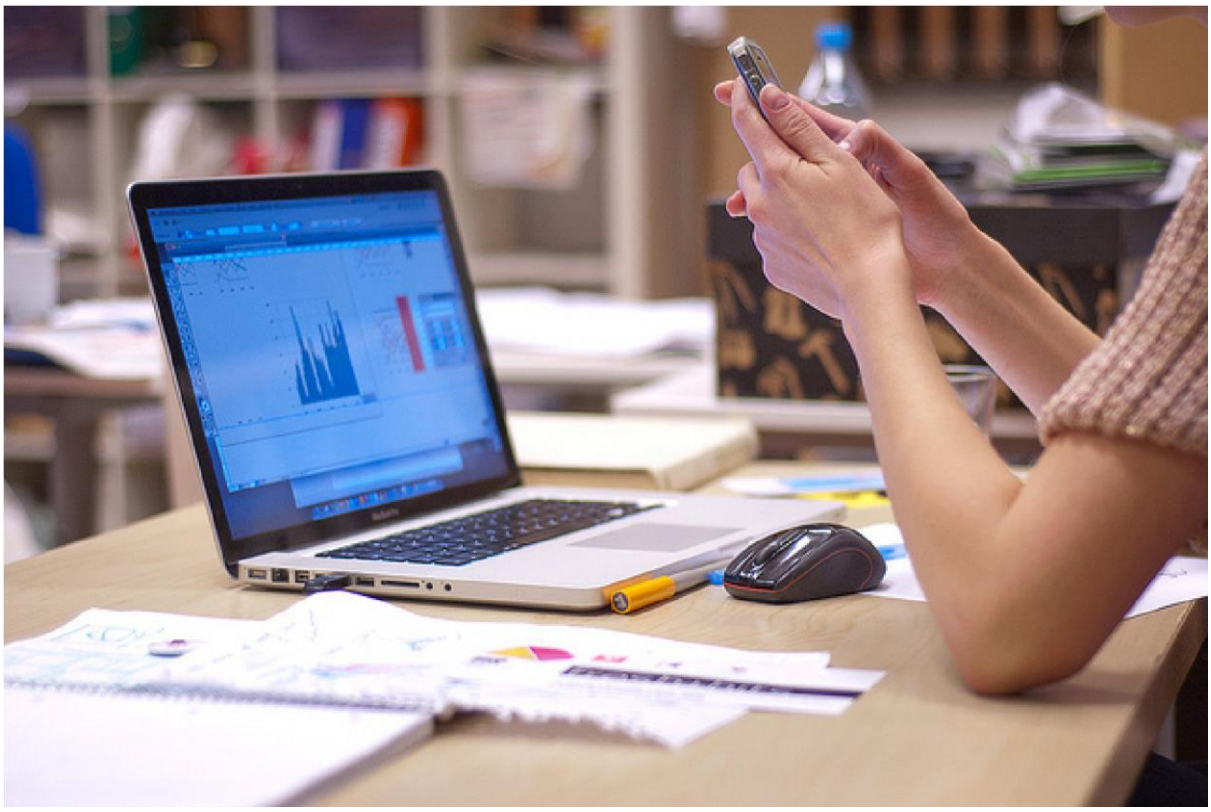


Photo par Sebastian Sikora Sous licence CC BY 2.0

Objectifs et résultats d'apprentissage

Cette leçon a pour objectif de fournir une base pour la préparation, l'analyse et la présentation des résultats des données.

A la fin de cette leçon, vous devrez être en mesure:

- D'expliquer pourquoi les données doivent être analysées
- De préparer des données pour l'analyse
- D'appliquer un certain nombre de techniques d'analyse des données
- D'analyser les risques reliés aux différents types de données
- D'expliquer l'objectif de la visualisation des données
- De choisir une visualisation appropriée pour les données
- D'évaluer l'efficacité de visualisations de données différentes

Sommaire

Liste des illustrations

Liste des tableaux

1. Introduction

La leçon 2.1 a présenté les méthodes de découverte des données " sur " et " dans " le Web ; l'objectif de cette leçon était de présenter des techniques pour trouver les données de base. La leçon 2.2 a été suivie d'une analyse de la qualité et de la provenance des données et des étapes nécessaires pour nettoyer et préparer les données pour l'analyse.

Cette leçon porte sur la prochaine étape dans l'analyse et la visualisation des données. Comme les autres leçons, les connaissances requises varient énormément en fonction de l'objet exact de l'analyse et de la visualisation. Cette leçon vise à adopter une approche pragmatique des deux sujets et à proposer une large théorie de l'analyse qui conduit à des visualisations claires génératrices d'impact. Des exemples pratiques se concentrent sur des données quantitatives dans un tableur et une analyse basée sur l'extraction de caractéristiques de données qualitatives.

2. Objectif de l'analyse des données

Les données brutes non traitées sont souvent en désordre et ne sont pas toujours prêtes pour la visualisation. Cette section examine un certain nombre de techniques pouvant être utilisées pour transformer des données en informations, notamment :

- Dérivation et extraction de caractéristiques
- Combinaison des ensembles de données
- Enrichissement des ensembles de données

3. Dérivation et extraction de caractéristiques

La dérivation et l'extraction de caractéristiques sont similaires en ce sens qu'elles sont conçues pour s'ajouter aux données existantes sans nécessiter des ensembles de données externes. Cela entraîne l'ajout de colonnes (ou de fonctions) supplémentaires aux données basées uniquement sur les données existantes.

La dérivation

Un élément de données dérivé est un élément de données dérivé d'autres éléments de données en utilisant un type de transformation mathématique,

logique ou autre, par exemple une formule arithmétique, une composition, une agrégation.

Par exemple, les données sources peuvent contenir une série de colonnes de dépenses mensuelles. Une somme totale pourrait être faite sur tous les mois pour ajouter une autre colonne. Cette colonne est donc dérivée des autres.

L'extraction de caractéristiques

L'extraction de caractéristiques est très similaire aux données dérivées mais ne doit pas nécessairement impliquer une fonction.

Par exemple, la ville pourrait être extraite d'une liste de données d'adresses non structurées, faisant de la ville une caractéristique distincte de l'ensemble de données. De même, la couleur du pixel central d'une image pourrait être extraite sans l'utilisation d'une fonction mathématique.

Un court tutoriel sur l'extraction de caractéristiques est disponible pour Excel ici[1].

3.1. Combinaison d'ensembles de données

Ici, nous ne parlons pas d'ajouter des données supplémentaires à la fin d'un ensemble de données existant (un processus connu sous le nom de consolidation), mais plutôt le processus d'ajout aux données existantes.

Si vous utilisez un tableur, le résultat de la combinaison de deux ensembles de données de 10 colonnes chacun sera un ensemble de données comportant 19 colonnes. La combinaison de données peut uniquement être effectuée de cette manière s'il existe une colonne dans chaque ensemble de données avec une valeur partagée. Cette colonne devient la clé sur laquelle la combinaison peut être effectuée.

Par exemple : prenez un ensemble de données de routes et le nombre de voitures pendant chaque heure de la journée. Chaque période d'une heure est une colonne et chaque route est une ligne. S'il y avait plusieurs sources de données, par exemple des sources avec des heures différentes dans tous recueillis individuellement, alors vous pouvez saisir le nom de la route et combiner toutes les données dans un seul ensemble de données.

Découvrez comment combiner deux ensembles de données dans Excel[2] ou dans Open Refine[3].

[1] https://www.extendoffice.com/documents/excel/3639-excel-extract-part-of-string.html
[2] https://superuser.com/questions/366647/merge-two-excel-files-using-a-common-column
[3] https://blog.ouseful.info/2011/05/06/merging-datesets-with-common-columns-in-googlerefine/

3.2. Enrichissement des données géographiques

Tout comme la combinaison de données, où deux ensembles de données peuvent être combinés en fonction d'une caractéristique commune (par exemple le nom de la route), le même résultat peut être obtenue avec des données géographiques. Une grande différence est qu'il est possible de combiner des données géographiques en fonction de l'emplacement et de placer un point géographique à l'intérieur d'une limite. Ce processus s'appelle jonction spatiale.

Une jonction spatiale peut être utile lorsque vous examinez les caractéristiques de différents bâtiments ou services et que vous les cartographiez sur la région de compétence pour voir si des tendances émergent. Par exemple, cela pourrait être utilisé pour voir si les écoles dans différentes zones contrôlées par le conseil fonctionnent de manière sensiblement différente.

En savoir plus sur l'enrichissement et l'ajout de données cartographiques à l'aide des excellents guides fournis par CartoDB[3].

Une autre étape essentielle dans la préparation des données géographiques à analyser est celle du géocodage. Le géocodage est le processus qui consiste à prendre toute référence ou description d'un emplacement physique et à ajouter les coordonnées de l'emplacement physique réel aux données. Le géocodage inverse est donc l'inverse, en extrayant la description (par exemple Londres) des coordonnées. Le géocodage peut également se référer au processus de transformation d'un système de représentation de coordonnées (par exemple abscisse/ordonnée) à un autre (par exemple latitude-longitude). Le géocodage est essentiel pour essayer d'analyser des données géographiques et d'effectuer d'autres opérations telles que des jonctions spatiales.

4. Analyse qualitatives et quantitatives des données

Il y a deux principaux types d'analyse statistique dont il est question dans les données : qualitative et quantitative. En termes simples, la recherche qualitative porte sur des données ouvertes, souvent basées sur du texte, tandis que la recherche quantitative tente de se concentrer sur des données objectives et mesurables sous forme de chiffres ou d'autres données structurées. Le tableau 1 montre certaines des principales différences entre les deux types de recherche.

Tableau 1 Recherche qualitative versus quantitative (Source : Open Université)

[3] https://carto.com/learn/guides

	Qualitative research	Quantitative research
Type of knowledge	Subjective	Objective
Aim	Exploratory and observational	Generalisable and testing
Characteristics	Flexible	Fixed and controlled
	Contextual portrayal	Independent and dependent variables
	Dynamic, continuous view of change	Pre- and post-measurement of change
Sampling	Purposeful	Random
Data collection	Semi-structured or unstructured	Structured
Nature of data	Narratives, quotations, descriptions	Numbers, statistics
	Value uniqueness, particularity	Replication
Analysis	Thematic	Statistical

Bien que le tableau ci-dessus illustre la recherche qualitative et quantitative comme distinctes et opposées, dans la pratique, elles sont souvent combinées ou tirent des éléments les unes des autres. Par exemple, une question d'enquête pourrait recevoir une réponse sur une échelle de 1 à 10, mais la question pourrait être sujette à des préjugés personnels.

Même une chose aussi simple que de compter les moutons peut être dangereuse si votre champ contient aussi des agneaux ; sont-ils des moutons ? Quand deviennent-ils moutons ?

4.1. Analyse des données quantitatives

Les bons chercheurs quantitatifs chercheront à maintenir un niveau de contrôle des différentes variables et définiront soigneusement la portée et la taille de l'échantillon mesuré. Ils essaieront également d'éliminer ou d'accepter l'influence d'autres facteurs sur l'échantillon et le souligneront clairement dans la recherche.

L'un des aspects les plus importants de la recherche est l'obtention d'un résultat statistiquement significatif. La signification statistique est essentiellement une mesure qui indique qu'il y a moins de 5% de chance que le résultat de l'analyse soit dû à un hasard.

Cela s'explique le mieux par les tirages à pile ou face et par le fait de vérifier si une pièce de monnaie est faussée, ce qui peut être vérifié par un simple test d'hypothèse nulle. L'hypothèse nulle est une affirmation sur le monde qui peut expliquer de manière plausible les données que vous observez, par exemple : "la pièce est juste". Vous pourriez alors retourner la pièce 100 fois et si les têtes n'ont été soulevées qu'une seule fois, alors on peut dire que l'hypothèse nulle peut être rejetée et que la pièce est faussée.

Mais que se passe-t-il si 51 des 100 lancers sortent la tête ? Ou vous retournez la pièce 100 000 fois et 51 000 fois il sort des têtes. L'un ou l'autre de ces éléments peut-il être aléatoire ou la pièce est-elle faussée ?

Le calcul de la représentativité d'un résultat peut se faire de deux façons, soit par un calcul de signification statistique, soit par un calcul de score z. Quoi qu'il en soit, ce que nous essayons de calculer, c'est si le résultat se situe en dehors des 95% d'observations où l'hypothèse nulle est vraie, ce qui la réfute. Comme la majorité des données quantitatives peuvent être associées à une distribution normale, c'est ce qui fait qu'un résultat est statistiquement significatif. Le diagramme ci-dessous montre que 95 % des valeurs se situent toujours dans les limites de 1,98 écart-type de la moyenne (x).¯

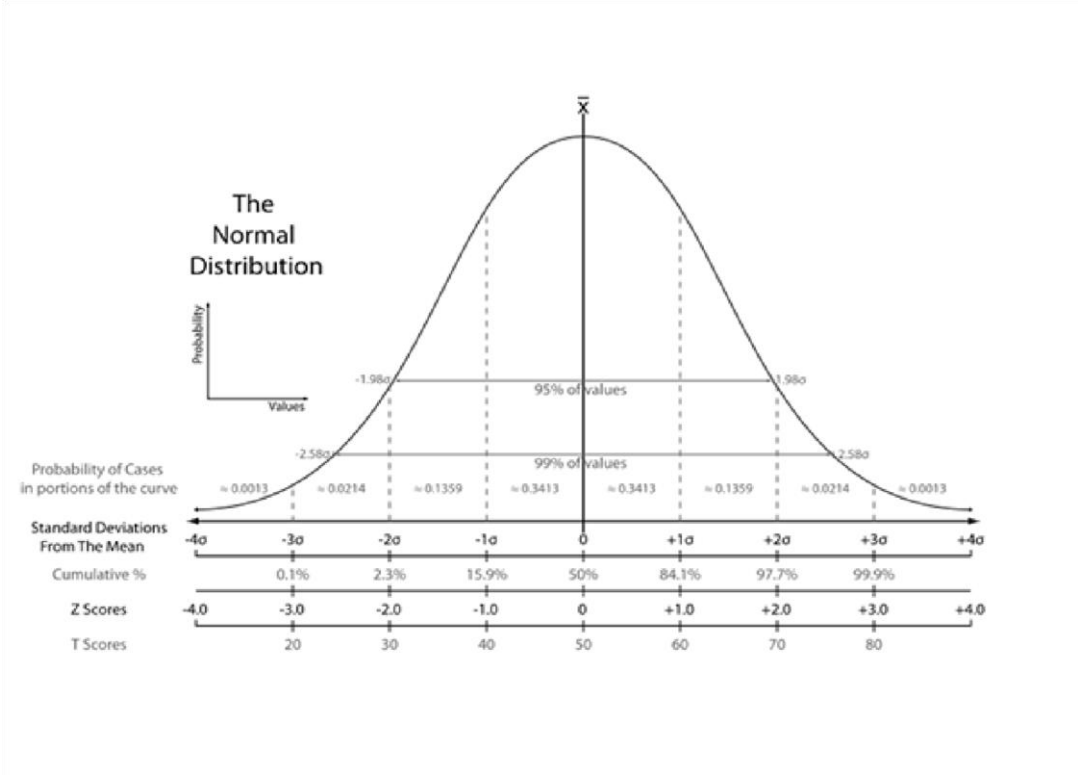


Illustration 1 Distribution normale et mesures de l'importance

Un niveau de confiance de 95 % signifie que nous rejetons l'hypothèse nulle si le calcul de la signification statistique se situe en dehors de 95 % de l'aire de la courbe normale. Le score z est un calcul du nombre d'écarts-types par rapport à la moyenne de l'échantillon.

Ainsi avec un tirage au sort d'une pièce de monnaie avec une hypothèse nulle que la pièce est juste, basée sur différentes quantités de flips et de résultats nous obtenons ce qui suit :

Pièce de monnaie	Flips	Pourcentage de têtes	score *Z*	Faussée?
#1	100	51%	0,2	Non

#2	100	60%	2,00	Oui
#3	50	60%	1,41	Insignifiant
#4	50	65%	2,12	Oui

Ici, on peut voir que lorsque la taille de l'échantillon augmente, la signification du résultat augmente même si pourcentage de lancers de pièces qui sortent des têtes demeure constant.

La taille de l'échantillon est un facteur clé et elle aide à éliminer le risque que d'autres facteurs aient un effet, par exemple que la pièce soit toujours lancée de la même manière, avec la même force et à partir de la même position. Cependant, certains tests peuvent nécessiter un contrôle absolu lorsque la taille de l'échantillon est faible, par exemple en comparant la résistance à l'écrasement de téléphones chers [4].

Comme la majorité des analyses quantitatives sont axées sur les moyennes, la taille de l'échantillon est importante pour éliminer l'effet des valeurs extrêmes sur les données. De plus, il est important de choisir la moyenne la plus représentative pour l'ensemble de données, car la moyenne n'est pas toujours un reflet exact des données s'il y a des valeurs extrêmes dans ces données. Par exemple, dans son discours sur l'état de l'Union de 2008, le président Bush s'est attaqué à son opposition, affirmant que ses plans fiscaux se traduiraient par une hausse d'impôt moyenne de 1 800 $ par personne. Cependant, c'était la valeur moyenne et les revenus des gens ne sont pas répartis également ; les propositions frappaient en fait durement les personnes à revenu élevé, tout en remettant de l'argent dans les poches de ceux qui en avaient besoin.

Il n'y a pas que la taille de l'échantillon qui influe sur la représentativité d'un résultat, cependant, et il y a d'autres facteurs qui doivent être éliminés avant qu'un résultat puisse être déclaré significatif :

Fluctuation

Il est donc important de garder à l'esprit la fluctuation des événements aléatoires. Tourner toute une série de têtes sur une pièce de monnaie ne garantit pas qu'il y en aura d'autres à suivre (à moins que la pièce ne soit faussée). Il en est de même pour les accidents de la route. L'installation de radars dans les points noirs des accidents n'est pas nécessairement corrélée à la réduction du nombre d'accidents qui s'ensuit.

Cibles

Bien que la plupart des données soient normalement distribuées, les cibles ont un effet profond sur le comportement des gens. Une fois qu'une cible est

[4] http://bgr.com/2017/09/25/iphone-8-plus-vs-galaxy-note-8-drop-test-glass/

introduite, les gens vont jouer sur leur résultat pour l'atteindre. Par exemple, un hôpital américain n'a opéré que des patients en bonne santé afin qu'ils puissent obtenir des fonds en apparaissant comme l'hôpital le plus performant.

Corrélation

Ce n'est pas parce qu'il semble y avoir une corrélation qu'il y en a une. Un exemple célèbre de corrélations fallacieuses est que le prix de la banane est fortement corrélé avec le nombre de personnes qui sont mortes en s'empêtrant dans leurs draps de lit. Ou la corrélation de 99% entre les dépenses spatiales, scientifiques et technologiques des États-Unis et les suicides par pendaison, strangulation et suffocation.

Pourcentages

Les pourcentages peuvent donner l'impression que les petits nombres sont grands et sont souvent utilisés de cette façon. Par exemple, l'année dernière, il y a eu une augmentation de 100% des cas de cancer liés aux téléphones portables : il est passé de 1 à 2, l'autre problème avec les pourcentages dans la mesure où beaucoup confondent points de pourcentage et pourcentages. Considérons par exemple le cas où la taxe sur la valeur ajoutée au Royaume-Uni a augmenté de 2,5 % (soit 2,5 points de pourcentage), passant de 17,5 % à 20 % ; une augmentation de 2,5 % aurait en fait porté son taux à 17,9375%.

Données non normalisées

Un autre piège courant est de ne pas ajuster les chiffres pour tenir compte des variables flottantes ; dire que L'incidence mondiale du cancer devrait augmenter de 75% d'ici 2030[5] est assez choquant. Jusqu'à ce que vous réalisiez que ce pourcentage a été calculé à partir du nombre brut de personnes et n'a pas été ajusté en fonction de la croissance démographique. Un autre type de gros titre courant est de dire que les dépenses consacrées à un secteur ou à un service ont augmenté chaque année, mais qu'elles ne sont pas normalisées en fonction de l'inflation, ce qui montre que la valeur réelle de ces dépenses est en fait en baisse.

Quelle que soit la technique utilisée pour analyser les données quantitatives, une vraie objectivité peut demander beaucoup d'efforts, même lorsque l'objet de l'analyse est une simple pièce de monnaie.

4.2. Analyse des données qualitatives

Toutes les données ne sont pas présentées sous forme de tableaux structurés ou de données géographiques localisées avec précision. Souvent, ce sont les données qualitatives qui peuvent être les plus difficiles à analyser et à utiliser. Les

[5] http://www.cancerresearchuk.org/about-us/cancer-news/news-report/2012-05-31-globalcancer-incidence-predicted-to-increase-by-75-per-cent-by-2030

données qualitatives sont des informations qui ne peuvent être mesurées et sont donc très subjectives (Même la couleur peut être subjective)[6].

L'objectif principal de l'analyse qualitative des données est de consolider et de donner un sens à de grandes quantités d'informations, souvent de sources différentes. Le résultat de cette analyse est d'offrir une explication, une interprétation ou un résumé thématique des données. Les apports à l'analyse qualitative peuvent prendre de nombreuses formes : transcriptions d'entrevues, documents, blogues, sondages, photos, vidéos, etc.

L'analyse qualitative des données est un processus plus naturel pour les humains qui cherchent naturellement à distiller des intrants autour de thèmes et des résultats clés, comme cela est particulièrement vrai pour les réunions ou les groupes de discussion. Les gens utilisent souvent des cartes mentales ou des cartes autocollantes « post-it » pour les aider à se regrouper et à catégoriser les éléments d'une vaste discussion sur des thèmes clés.

L'analyse qualitative des données doit tenir compte du contexte de la parole, de la cohérence et des contradictions des opinions, de la fréquence et de l'intensité des commentaires, de leur spécificité ainsi que des thèmes et tendances émergents.

Il y a deux façons principales d'analyser les données qualitatives : l'analyse du cadre et l'analyse des réseaux thématiques.

L'analyse du cadre implique la construction d'un ensemble prédéfini de critères qui reflètent clairement vos buts, objectifs et intérêts. En utilisant cet ensemble de critères, les informations pertinentes peuvent être extraites des données et comparées aux autres entrées du cadre. L'utilisation d'un cadre permet à de nombreux chercheurs de faire l'extraction tout en minimisant les risques de biais d'analyse qualitative. Un biais peut encore être introduit à l'étape de la conception du cadre, ce qui peut entraîner l'omission d'informations clés.

Une approche alternative consiste à ne pas construire un cadre, mais plutôt à appliquer une analyse de réseau thématique. *L'analyse de réseau thématique* est une approche plus exploratoire qui encourage l'analyse de toutes les données d'entrée qui peuvent façonner la sortie dans des directions inattendues. En réalité, la majorité des analyses de données qualitatives impliqueront une combinaison des deux approches.

Quelle que soit l'approche choisie, la première étape de toute analyse qualitative des données implique une familiarisation avec les données ; lire et relire les réponses. En même temps, c'est une bonne idée de commencer à codifier les données en écrivant des mots-clés et des sujets qui tentent de réduire et d'interpréter les données. Le résultat du processus de codage

[6] https://en.wikipedia.org/wiki/The_dress

pourrait être l'analyse de réseau thématique ou un cadre par lequel toutes les réponses doivent être codées. Quoi qu'il en soit, le codage peut être un processus long, lent et répétitif, mais il existe un certain nombre d'outils qui peuvent aider à l'analyse de réseau thématique.

4.2.1. Outils d'analyse de réseaux thématiques

Les outils de reconnaissance des entités fournissent une telle technique qui peut aider à analyser et à enrichir les données qualitatives. Essentiellement, la reconnaissance des entités vise à localiser et à classer les entités nommées dans le texte en catégories prédéfinies telles que les noms de personnes, d'organisations, de lieux, d'expressions de temps, de quantités, de valeurs monétaires, de pourcentages, etc.

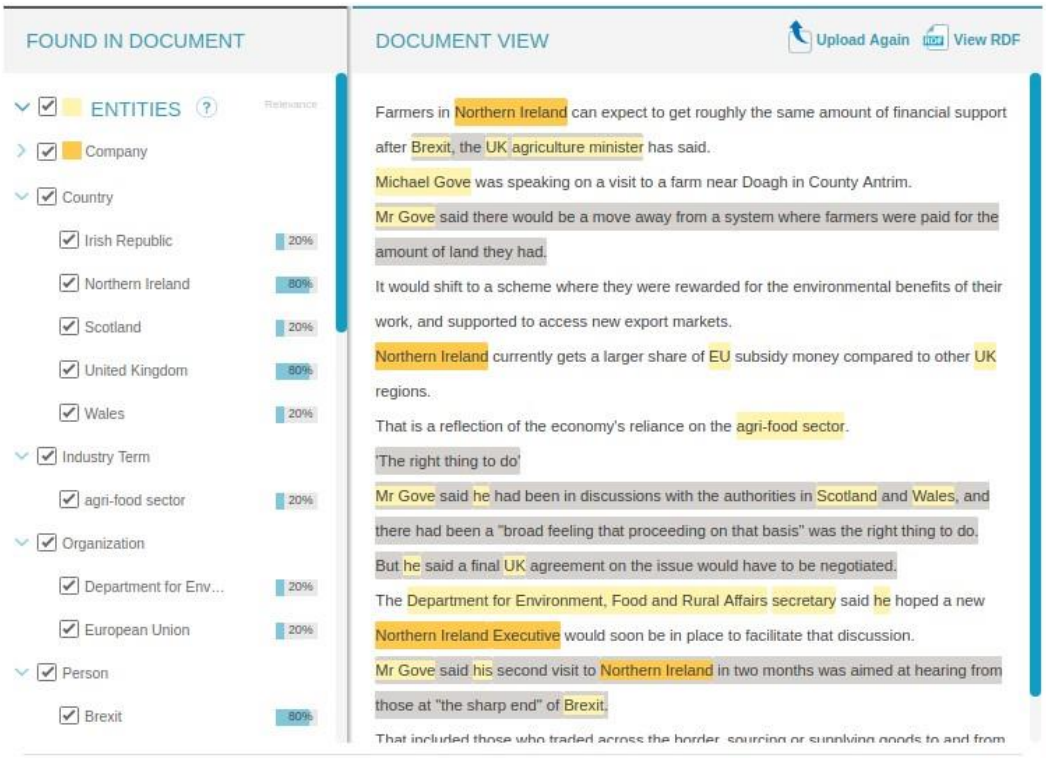


Illustration 2 Calais tagging de BBC News : Les fermes de l'IN peuvent s'attendre à recevoir le même soutien après Brexit - Gove

Calais[7] (montré ci-dessus) de Thomson Reuters est un exemple d'un tel moteur de reconnaissance d'entités. Plus qu'une simple reconnaissance d'entité, le service Calais relie ces entités aux enregistrements de données qui existent dans la base de données Thomson Reuters perm ID des entités et des enregistrements financiers.

[7] Open Calais – http://www.opencalais.com

Plus largement, une telle reconnaissance d'entité est utilisée par des services tels que TheyWorkForYou[8] afin de suivre les activités des politiciens et de fournir cela facilement au public.

Calais est un bon exemple d'outil en ligne capable de reconnaître des entités et de fournir des liens vers des données supplémentaires sur chaque sujet. D'autres techniques peuvent être beaucoup plus simples mais tout aussi efficaces telles que word cloud generators [9].

5. Visualisation des données

Une autre façon d'interpréter rapidement les données est de les visualiser. Le cerveau humain est beaucoup plus apte à consommer et à comprendre les données présentées dans une visualisation que sous forme de texte.

La plupart des graphiques utilisés dans la visualisation de données moderne dérivent des conceptions originales de William Playfair (1759-1823), un économiste politique. Playfair a inventé plusieurs types de diagrammes : en 1786 la ligne, la zone et le diagramme à barres des données économiques, et en 1801 le diagramme circulaire et camembert, utilisé pour montrer les relations partitives.

Le choix de la technique de visualisation à utiliser dépend à la fois de l'objectif de la visualisation et du type de données à visualiser. Cette section explore les différents objectifs et les types de visualisation adaptés à chacun avec des exemples.

5.1. Les objectifs de la visualisation des données

L'objectif de la visualisation des données est de communiquer aux utilisateurs des informations de manière claire et efficace.

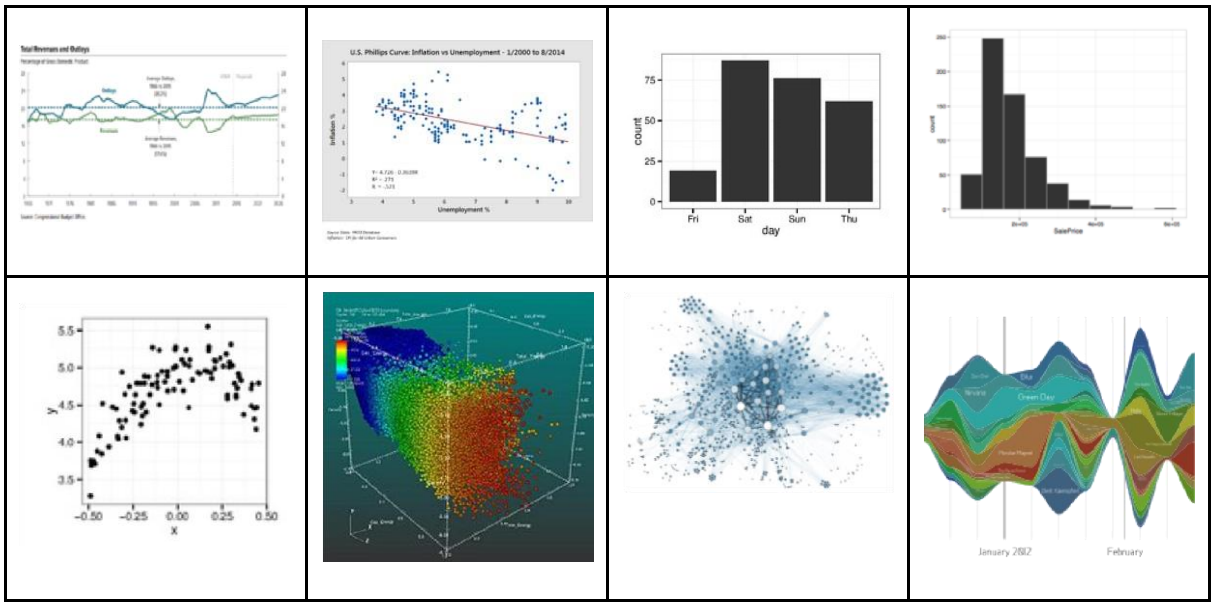
[8] https://www.theyworkforyou.com
[9] https://www.jasondavies.com/wordcloud/

Selon Friedman (2008), « l'objectif principal de la visualisation des données est de communiquer l'information clairement et efficacement par des moyens graphiques. Cela ne signifie pas que la visualisation des données doit être ennuyeuse pour être fonctionnelle ou extrêmement sophistiquée pour être belle. Pour transmettre efficacement les idées, la forme esthétique et la fonctionnalité doivent aller de pair, en donnant un aperçu d'un ensemble de données plutôt clairsemées et complexes, en communiquant ses aspects clés de manière plus intuitive. Pourtant, les concepteurs ne parviennent souvent pas à trouver un équilibre entre la forme et la fonction, créant de magnifiques visualisations de données qui ne remplissent pas leur objectif principal - communiquer l'information ».

L'un des principaux problèmes de la visualisation des données est qu'il s'agit d'un vaste sujet qui couvre un large éventail de visualisations différentes qui sont conçues à des fins différentes. La communication n'est qu'un des objectifs d'une visualisation de données. Les visualisations de données peuvent également être utilisées pendant la phase d'analyse des données afin d'aider à donner un sens à l'analyse des données et à la guider.

Si l'objectif principal d'une visualisation des données est de communiquer des informations, la visualisation doit pouvoir le faire sans que le lecteur n'ait besoin de texte explicatif ou de connaissances supplémentaires.

Prenez les exemples suivants de la page de visualisation des données de Wikipédia : On peut soutenir que seuls les diagrammes à barres (en haut à droite) et la carte arborescente (en bas à gauche) sont de bonnes visualisations pour la communication instantanée des données. C'est parce que ces deux graphiques utilisent une astuce de visualisation appelée 'pop out'.



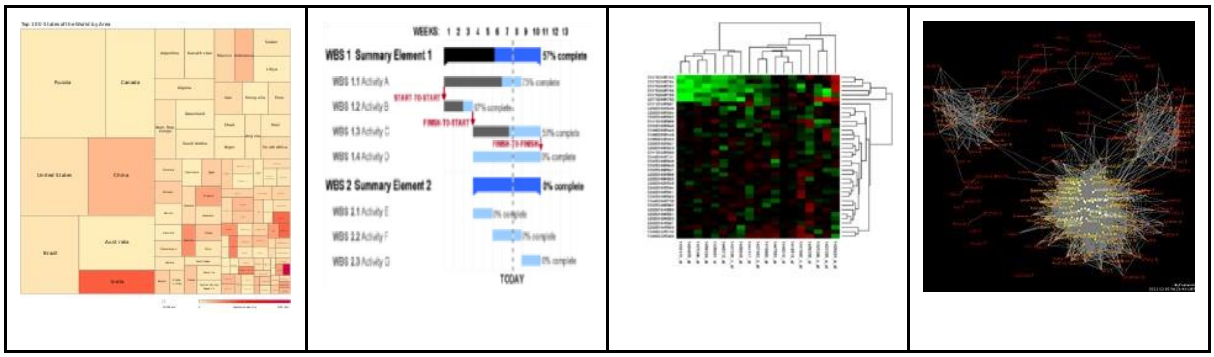


Illustration 3 Exemples de visualisation de données (à partir de Wikipedia)

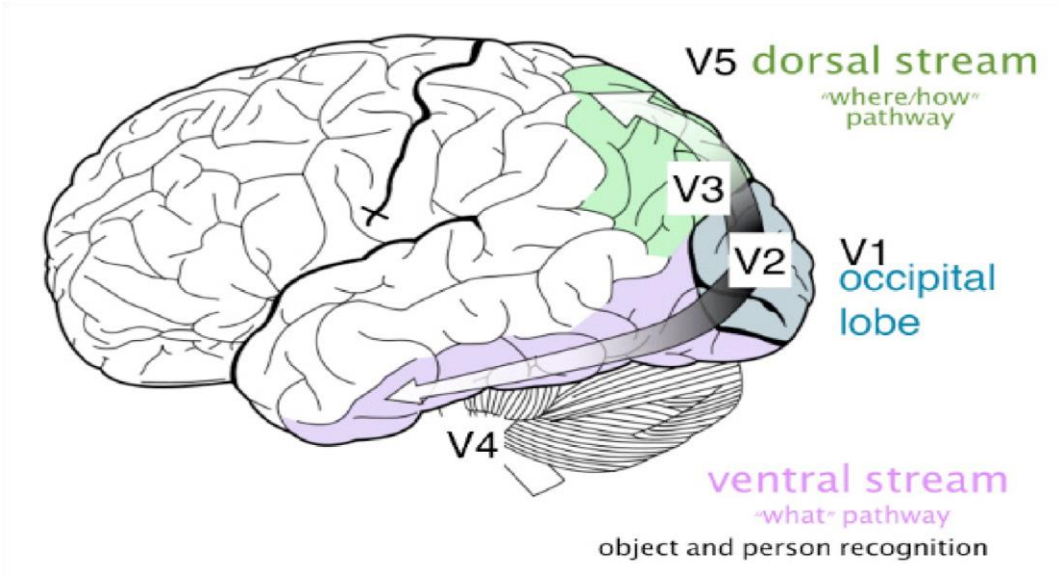


Illustration 4 Voies dans le cortex visuel

Le pop out aide à diriger l'œil vers le bon endroit, instantanément et l'œil humain est attiré vers les couleurs les plus vives, des objets plus gros et des choses qui font la différence. C'est une caractéristique qui est programmée dans notre cortex visuel. À l'intérieur du cortex visuel, il y a deux flux, le flux ventral (quoi) et le flux dorsal (où / comment). C'est le flux dorsal qui traite les informations provenant de l'œil sur notre environnement en temps réel afin que nous puissions réagir instantanément dans des situations à risque (par exemple, un objet qui vous est lancé).

C'est le courant dorsal qui regarde où sont les choses et comment elles se rapportent à d'autres choses qui font que les pop-ups fonctionnent si vite. Inversement, le flux ventral est responsable de l'élaboration de « ce que » la chose est. C'est un processus beaucoup plus lent et c'est la raison pour laquelle vous pourriez reconnaître le visage d'une personne, mais ne pas être en mesure de lui donner un nom.

Les meilleures visualisations de données pour la communication font appel au flux dorsal et mettent en évidence l'information. Si une visualisation nécessite l'utilisation du flux ventral pour aider à contextualiser l'information, alors il y a une

forte probabilité que les individus interprètent l'information différemment les uns des autres.

Essayez-le par vous-même avec le jeu de pop-out ci-dessous (Figure 5). Tout ce que vous avez à faire est de trouver l'intrus dans les deux images ci-dessous.

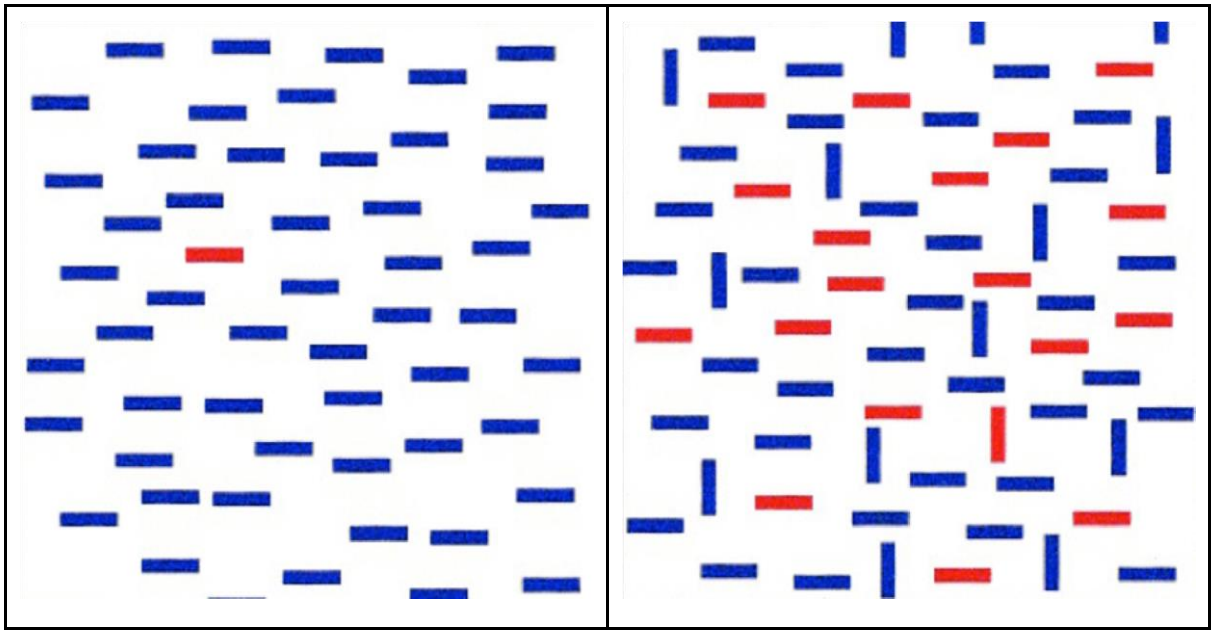


Illustration 5 Trouvez l'intrus.

Avec laquelle avez-vous été le plus rapide ?

Un autre aspect important de la visualisation des données est l'utilisation de la couleur. En plus de mettre l'information en évidence, la couleur peut aussi transmettre une signification. Par exemple : le rouge, l'ambre et le vert représentent le danger, la prudence et la sécurité, tels qu'ils sont utilisés dans les feux tricolores (Illustration 6).

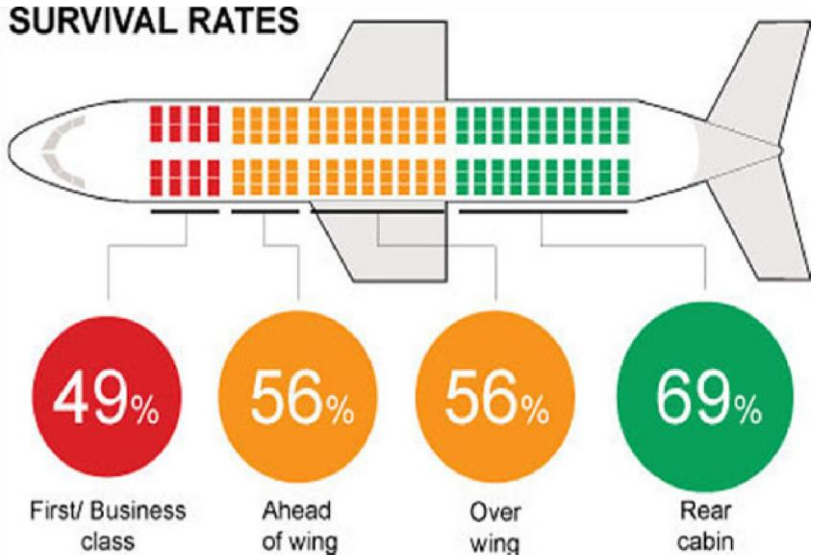


Illustration 6 Couleurs des feux tricolores indiquant la sécurité

Cependant, toutes les couleurs ne transmettent pas une signification facilement reconnaissable, et trop d'utilisation de la couleur peut être source de confusion. De cette façon, essayez d'utiliser le moins possible le nombre de couleurs.

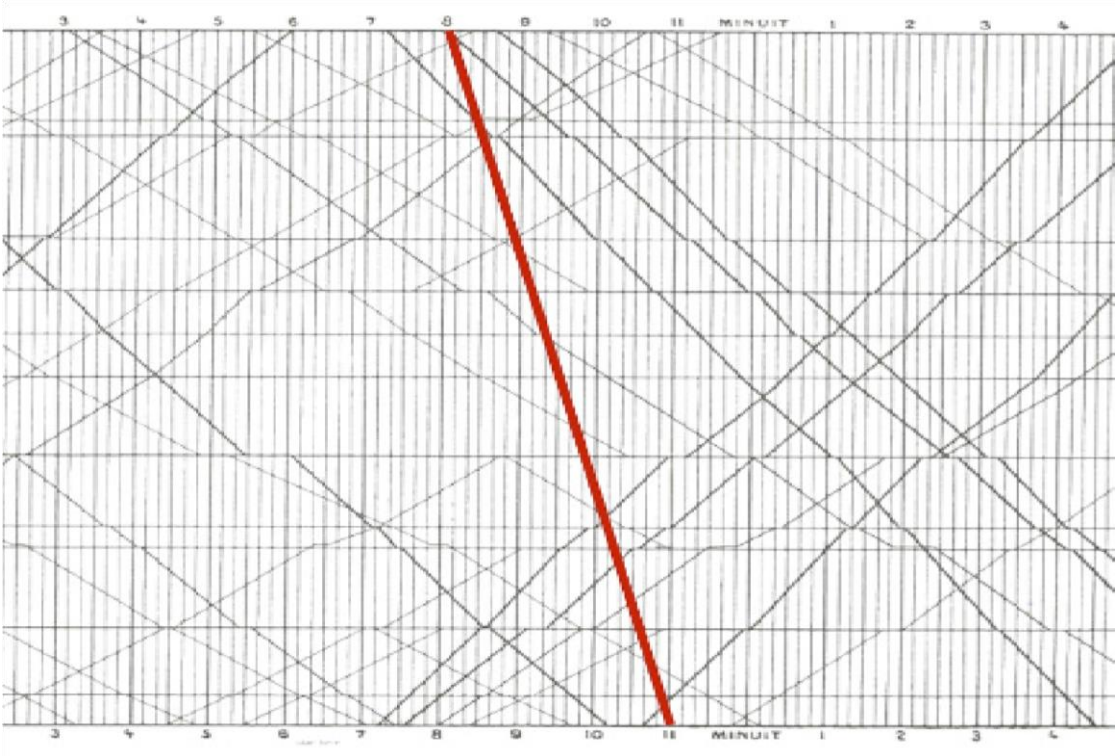


Illustration 7 'Pop-out' et autres repères visuels

L'emphase fonctionne mieux lorsqu'une combinaison de différents repères visuels est utilisée en combinaison, par exemple la couleur, la taille et l'épaisseur du trait.

Dans l'illustration 7, une combinaison de couleur et d'épaisseur guide l'œil vers la ligne la plus inclinée. Cette ligne représente le premier train à grande vitesse de Paris à Lyon. L'inclinaison de la ligne indique la vitesse.

Dans toute visualisation, il est essentiel de se débarrasser de tout encombrement extérieur pouvant nuire au message transmis.

5.2. Choisir la visualisation correcte des données

Le choix de la visualisation correcte des données dépend de deux aspects clés.
1. Le type de données
2. Le message à transmettre

Trois principaux types de données sont présentés dans le tableau 2 :

Tableau 2 Types de données

Type de données	**Description**	**Exemple de visualisation**
séries chronologiques	Observations des mêmes objets au fil du temps	Diagramme linéaire, diagramme de mouvement, diagramme de zone polaire, diagramme de Gantt, diagramme à barres
Population	Observation de différents objets à un moment donné	Diagramme à barres, treemap, diagramme circulaire
Multivariable	Observations de différents objets à différents moments dans le temps	Diagramme de mouvement multidimensionnel, diagramme à barres, Treemap

Prenons par exemple les données sur les rendements céréaliers de la Communauté de l'Afrique de l'Est : Tanzanie, Rwanda, Ouganda, Kenya, Burundi et Sud-Soudan, qui sont disponibles sous forme de données ouvertes auprès de la Banque mondiale[10]. Ces données peuvent être visualisées de nombreuses façons en raison de leur nature multivariée.

Les données disponibles auprès de la Banque mondiale sont disponibles pour chaque pays depuis 1961. Si les données ne concernaient qu'un seul pays, nous aurions alors un ensemble de données chronologiques. Si les données n'étaient que d'environ un an, il s'agirait d'un ensemble de données sur la population, mais les deux sont disponibles, donc l'ensemble de données est un ensemble multivarié.

[10] World Bank - Cereal yield (kg per hectare) - https://data.worldbank.org/indicator/AG.YLD.CREL.KG

Illustration 8 Visualisation des séries chronologiques du rendement céréalier dans la Communauté de l'Afrique de l'Est depuis 2000

Le graphique linéaire de l'illustration 8 est naturellement adapté pour illustrer une série chronologique pour ceux qui lisent de gauche à droite. Il permet également une comparaison relativement facile entre les pays à différents moments dans le temps. Notez comment la ligne rose plus foncée pour le Rwanda se détache, tandis que les tons plus clairs du bleu se fondent. Le bleu est une couleur très neutre, alors que les rouges et les jaunes sont beaucoup plus appropriés pour faire ressortir l'information. Ces paramètres permettent de constater que le Rwanda s'est amélioré par rapport aux autres pays de la CAE depuis 2000.

Inversement, le diagramme à barres (figure 9) compare les pays à un moment donné (2014 dans le cas présent). L'utilisation de la même couleur ici signifie que l'œil n'est pas attiré par un seul pays. Ce qui ressort ici, ce sont peut-être les trois regroupements de l'Ouganda avec le Rwanda, la Tanzanie avec le Kenya et le Burundi avec le Sud-Soudan. Par contre, le graphique chronologique nous indique que ces groupes peuvent ne pas être significatifs dans le temps

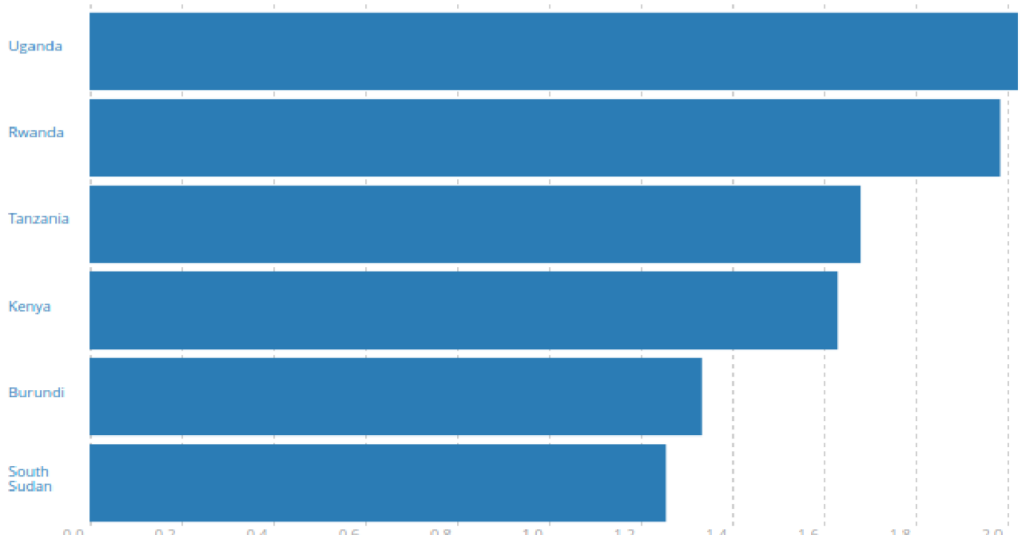


Illustration 9 Diagramme à barres de la population de la CAE en termes de rendement céréalier en 2014

Pour explorer davantage les regroupements, nous pourrions examiner une autre visualisation d'un ensemble de données sur la population, cette fois-ci une carte. Comme nous avons des pays, il est parfaitement logique de les montrer sur une carte (Illustration 10).

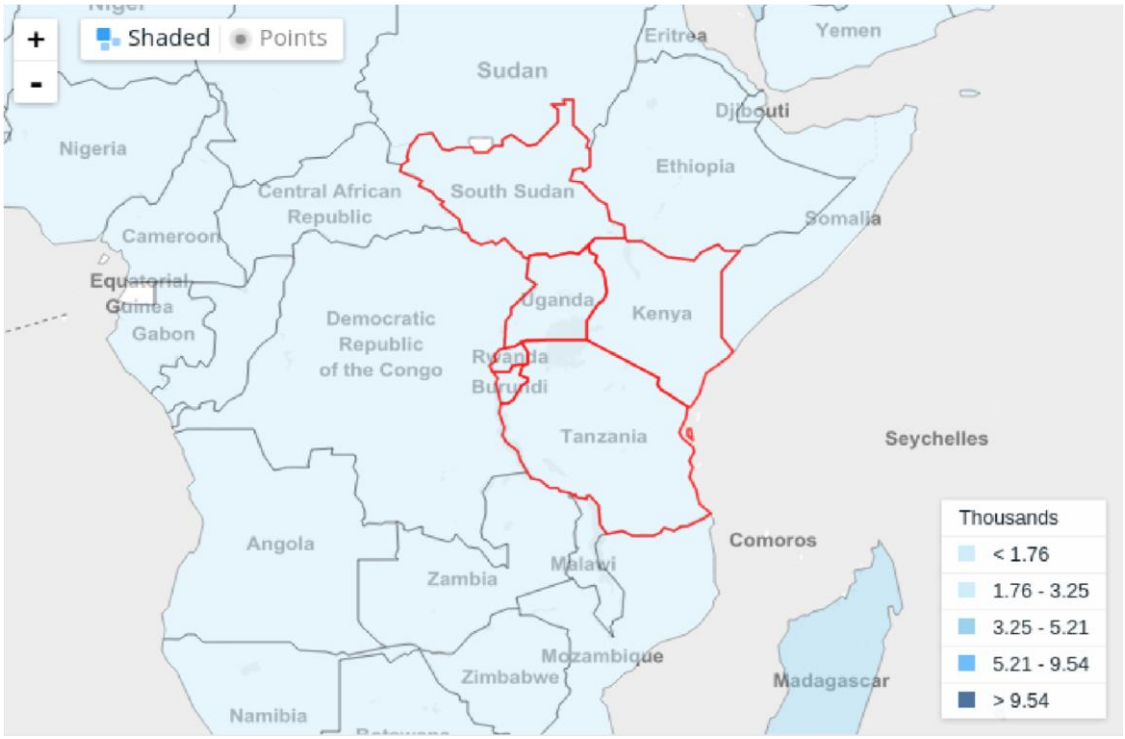


Illustration 10 Visualisation d'un ensemble de données géographiques démographiques sur le rendement céréalier dans la CAE en 2014

Malheureusement, la visualisation de la carte de la Banque mondiale ne filtre pas les pays pour simplement montrer ceux que vous avez sélectionnés. À l'échelle mondiale, les pays de la CAE sont très similaires en termes de production agricole, ce qui n'est pas reflété dans les graphiques à barres et les séries chronologiques.

Afin de révéler ces différences subtiles, les données de ces pays peuvent être utilisées avec un outil de cartographie différent (carto.com dans ce cas) afin de créer votre propre carte (comme dans l'illustration 11).

Illustration 11 Les pays de la CAE dans la cartographie alternative

Cette dernière série de visualisations montre comment la suppression de données peut modifier le message qui est transmis. Idéalement, la nouvelle visualisation devrait exclure les autres pays, car le fait qu'ils soient blancs est trompeur et il est inutile de montrer la carte du monde entier.

Résumé

L'analyse et la visualisation des données est un sujet difficile et complexe. C'est particulièrement le cas étant donné la nature multivariée des données et le fait que de nombreux ensembles de données sont maintenant " grands ", ce qui signifie que les problèmes de complexité sont multipliés.

Le choix de la bonne technique d'analyse est également fortement lié à la technique de collecte des données pour s'assurer que les erreurs systématiques sont limitées autant que possible. Il convient d'accorder une attention particulière à la collecte des données et à la manière dont l'analyse peut être influencée par des facteurs externes tels que les objectifs des politiques.

La visualisation des données doit également éviter d'introduire un biais ou tout autre aspect qui pourrait induire le spectateur en erreur. Étant donné que l'œil traite la visualisation en une fraction de seconde, la visualisation doit être conçue avec soin pour s'assurer que le message clé est transmis avec précision. Parce que 78 et 103 sont beaucoup plus proches que 71 et 78 ?

Références

- Friedman, V. 2008. 'Data Visualization and Infographics'. *Smashing Magazine*, 14 January, 2008. Available at: https://www.smashingmagazine.com/2008/01/monday-inspiration-datavisualization-and-infographics/#top

- Huberman, M. and Miles, M.B., 2002. *The Qualitative Researcher's Companion*. Sage, Thousand Oaks, CA, USA. 410 pp.