ROYAUME DU MAROC

MINISTÈRE DE L'ECONOMIE, DES FINANCES
ET DE LA RÉFORME DE L'ADMINISTRATION

الجامعة الدولية للرباط
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵜⴰⴳⵔⴰⵖⵍⴰⵏⵜ ⵏ ⵕⵕⴱⴰⵟ
Université Internationale de Rabat
L'UNIVERSITÉ INNOVANTE

DEPF Etudes

Apports des Big Data pour le suivi de l'activité économique et la prévision

depf.finances.gov.ma

Juillet 2020

Equipe de travail

DEPF
- Rafik NASHI
- Ahlam ERRAHMANI
- Ilyes BOUMAHDI

UIR
- Imane KHAOUJA
- Ibtissam MAKDOUN
- Ibrahim RAHHAL
- Ghita MEZZOUR

TABLE DES MATIERES

INTRODUCTION 4

PARTIE I : REVUE DE LITTERATURE SUR LES BIG DATA 5

1. Big Data : Quelques points de repères 6

1.1. Définition .....

1.2. Genèse des Big Data : quelques repères historiques 8

2. Types des sources de données et des Big Data 10

2.1. Caractéristiques des données d'enquêtes et des données administratives 10

2.2. Typologie des Big Data 11

3. Big Data : cadre réglementaire 13

3.1. Union Européenne : une réglementation couverte par le RGPD 14

3.2. Chine : Protection des données personnelles à travers la loi sur la cyber-sécurité 15

3.3. Inde : Projet de Personal Data Protection Bill 15

4. Revue des principaux travaux pour le suivi de l'activité économique, la prévision à court terme 15

4.1. Prévision et suivi du PIB et ses composantes 15

4.2. Emploi et statistiques du chômage 17

4.3. Indices des prix et inflation 18

4.4. Suivi de l'activité économique 19

4.5. Secteur financier 21

5. Utilisation des Big Data fiscales et budgétaires pour le suivi et la prévision macroéconomiques 22

5.1. Surveillance et gestion budgétaire 22

5.2. Surveillance et prévision de l'activité économique réelle 23

5.2.1. Utilisation des recettes fiscales pour la prévision immédiate 23

5.2.2. Autres applications pour des travaux analytiques macro-budgétaires 23

5.3. Cas du Maroc : Utilisation des Big Data pour la surveillance fiscale au sein de la DGI 24

6. Utilisation des Big Data dans le contexte de la pandémie du Coronavirus 25

6.1. Suivi et lutte contre la propagation du Coronavirus 25

6.2. L'analyse des effets de la pandémie sur l'activité économique 27

PARTIE II : ETUDES DE CAS POUR LE MAROC 30

Etude de cas N° 1 : Opportunités du Big Data pour un suivi avancé de l'activité touristique au Maroc 31

1. Caractéristiques de l'offre touristique sur le web de 10 destinations méditerranéennes 32

2. Appréciation de l'offre touristique sur le web de 10 destinations méditerranéennes 33

3. Appréciation de l'offre touristique sur le web de Marrakech et Agadir 34

4. Enseignements et recommandations 36

Etude de cas N° 2 : Usage des sciences des données pour améliorer l'employabilité des jeunes au Maroc 38

1. Étude des besoins du marché du travail 39

2. Analyse de l'inadéquation des compétences 43

3. Analyse des parties prenantes de l'emploi au Maroc 44

3.1. Analyse de la collaboration par questionnaire 44

3.2. Analyse de la collaboration entre les parties prenantes à travers leurs rapports officiels 44

4. Enseignements et recommandations 45

CONCLUSION GÉNÉRALE : DÉFIS ET IMPLICATIONS STATISTIQUES DES BIG DATA 46

Introduction

La disponibilité croissante et massive des données, souvent désignées sous le terme de « Big Data », ainsi que leur prolifération exceptionnelle suite aux avancées sans précédent qu'ont connues les technologies du numérique, sont considérées comme une opportunité pour améliorer et enrichir la production d'information et comme un changement de paradigme en termes d'analyse qui incite les chercheurs à ouvrir un nouveau champ d'investigation pour repenser notre façon de traiter la profusion des données offertes et de se projeter dans l'avenir.

En effet, les Big Data sont devenus un outil indispensable à la transformation de l'action publique et plus largement de l'économie. Ils s'imposent comme le socle incontournable de la transformation numérique et deviennent une priorité pour toutes les organisations. Ils sont le fruit des avancées technologiques regroupant plusieurs acteurs, notamment les gouvernements, les consommateurs ou les entreprises. Leur champ d'application relève de plusieurs domaines d'activité et peut servir à prendre des décisions, à faire des prévisions et/ou à lancer des initiatives.

Ainsi, des pays, à l'instar de la Corée du Sud, la Chine et le Taiwan, se sont appuyés sur les Big Data pour suivre l'évolution de l'épidémie du coronavirus et contenir sa propagation. L'utilisation et le croisement des données massives ont permis d'identifier les cas suspects et leurs mouvements avant de déterminer leurs contaminations et d'intervenir pour faire face à cette épidémie.

Si les Big Data représentent des opportunités, la disponibilité et l'utilisation croissantes des données pour créer de la valeur représentent, également, d'importants défis et enjeux. Il s'agit, essentiellement d'un enjeu d'accès à l'information, dans la mesure où ces nouvelles données appartiennent généralement à des entreprises privées, d'un enjeu juridique et de propriété de l'information et d'un enjeu de sécurité et de confidentialité des données.

La présente étude est structurée en deux parties. La première partie est consacrée à la présentation d'une revue de littérature sur les Big Data. Il s'agit de décliner le potentiel de contribution des Big Data en matière de suivi de l'activité et des prévisions économiques. Elle éclaire les différents contours des Big Data, leurs types, leurs caractéristiques et leur développement historique. Elle détaille les apports des Big Data pour le suivi de l'activité et la prévision économique. Elle présente les principaux défis qui doivent être relevés afin d'en optimiser l'utilisation dans le contexte de la statistique publique avant d'aborder un dernier point sur l'utilisation des Big Data dans le contexte de la pandémie du Coronavirus.

La deuxième partie, quant à elle, est consacrée à la présentation de deux études de cas pour le Maroc. La première est réalisée par la DEPF[1] afin de traiter les opportunités du Big Data pour un suivi avancé de l'activité touristique au Maroc. La deuxième est réalisée par le laboratoire TICLab de l'UIR[2] afin d'étudier l'usage des sciences des données pour améliorer l'employabilité des jeunes au Maroc.

---

[1] M. Ilyes BOUMAHDI, chef du service des activités tertiaires et de l'économie de savoir à la Direction des Etudes et des Prévisions Financières (DEPF).

[2] Ghita MEZZOUR, Professeur associé à la faculté d'informatique et logistique de l'Université International de Rabat (UIR) et Directeur-Adjoint du TICLab.

PARTIE I

REVUE DE LITTERATURE SUR LES BIG DATA

Rafik NASHI

Ahlam ERRAHMANI

Service de l'Analyse des Données, DEPF

1. BIG DATA : QUELQUES POINTS DE REPÈRES

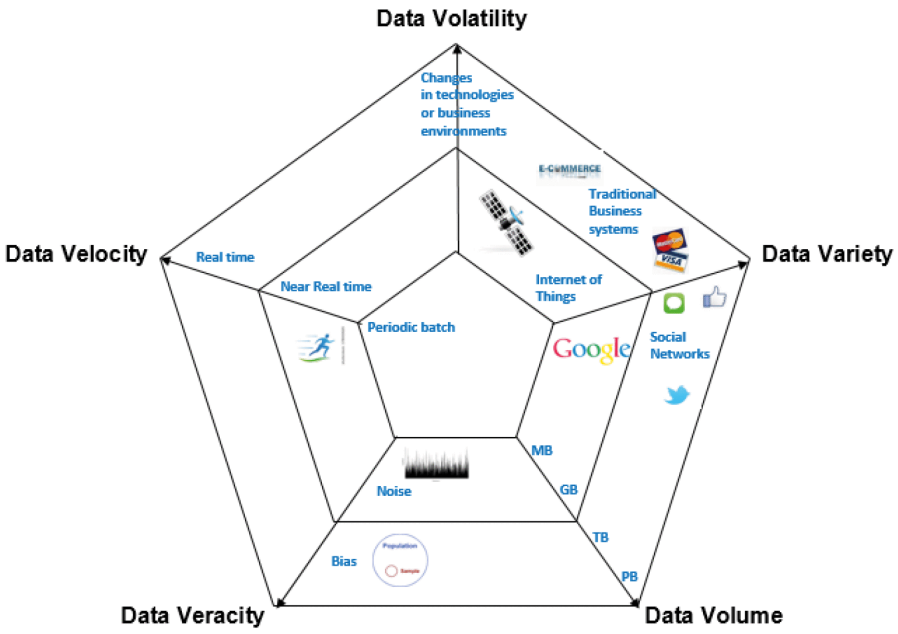
1.1. DÉFINITION

Le concept de Big Data comprend les aspects relatifs aux applications, à l'ingénierie et aux aspects scientifiques, mais il n'existe pas encore de définition unifiée du Big Data. Sa définition varie selon les communautés qui s'y intéressent en tant qu'usager ou fournisseur de services. Dans certaines communautés universitaires, le terme fait référence aux applications des technologies de l'information pour traiter des problèmes de données massives dans les entreprises et les composantes scientifiques ou les aspects de recherche du Big Data sont appelés science des données. Dans certaines communautés professionnelles, les termes « Business Intelligence (BI) » et analyse commerciale sont utilisés pour désigner l'analyse du Big Data ou l'exploration du Big Data (Chen[3] et al. 2012). La fondation « *National Science Foundation* » décrit le Big Data comme « *des ensembles de données volumineux, diversifiés, complexes, longitudinaux et / ou distribués générés à partir d'instruments, de capteurs, de transactions Internet, de courriers électroniques, de vidéos, de flux de clics et/ou de toutes les autres sources numériques disponibles aujourd'hui et dans le futur* » (NSF 2012).

Cependant, Shi Y[4] (2014) a présenté deux définitions du Big Data. Pour les universitaires, le Big Data est « *un ensemble de données complexes, difficiles à traiter et à analyser dans un délai raisonnable, complexes, hétérogènes et d'une grande valeur potentielle », tandis que pour les décideurs, le Big Data constitue « un nouveau type de ressource stratégique à l'ère numérique et le facteur clé de l'innovation, qui modifie la manière de produire et de vivre de l'homme* ».

Une autre possibilité réside au niveau des caractéristiques nommées les"5 Vs", créé par IBM[5], ils concernent : (i) le Volume (échelle de données), (ii) la Vélocité (analyse des données en continu), (iii) la Variété (différentes formes de données), (iv) Volatilité (changement de technologie et environnement d'affaires) et (v) la Véracité (incertitude des données).

Figure 1: Les 5 Vs des Big Data



Source: Hammer C.L., Kostroch D.C. et Quiros G., "Big Data: Potential, Challenges and Statistical Implications". FMI, septembre 2017.

[3] Chen H, Chiang RHL, "Business intelligence and analytics: From Big to Big Impact", MIS Quarterly, vol. 36, N° 4, pp. :1165 –1188, décembre 2019.
[4] Shi Yong, "Big Data: history, current status, and challenges going forward", Bridge vol. 44, n° 4, pp : 6–11, winter 2014.
[5] International Business Machines.

- **Volume :**

  Il représente l'une des caractéristiques les plus importantes du Big que l'on associe au Big Data. Cette association avec l'ampleur de l'ensemble de données se produit naturellement car tous les domaines tendent actuellement à collecter et stocker des quantités massives de données. Ce comportement est favorisé à la fois par les faibles coûts de stockage des données et des résultats plus précis, du point de vue de l'analyse des données.

- **Variété :**

  La collecte de données provenant de diverses sources conduit à une forte hétérogénéité. Le traitement du Big Data implique le plus souvent de manipuler des données sans structure relationnelle prédéfinie. Par conséquent, organiser les données avant de les stocker et de les traiter devient une tâche critique. La puissance du Big Data provient de la capacité à traiter et à extraire des informations de toutes sortes de données.

- **Vélocité :**

  C'est la vitesse à laquelle les données sont générées, collectées, stockées et traitées. La terminologie commune utilisée pour les données à déplacement rapide est la "transmission de données en continu". Initialement, les défis liés à la vélocité étaient limités à des segments spécifiques de l'industrie, mais il devient un problème d'un cadre beaucoup plus large avec l'Internet des objets.

- **Volatilité :**

  La volatilité est différente de la variété. Elle fait référence à des données dont la signification change continuellement. C'est particulièrement le cas lorsque la collecte de données repose sur le traitement du langage. En effet, les mots n'ont pas de définitions statiques, et leur signification peut varier énormément selon le contexte.

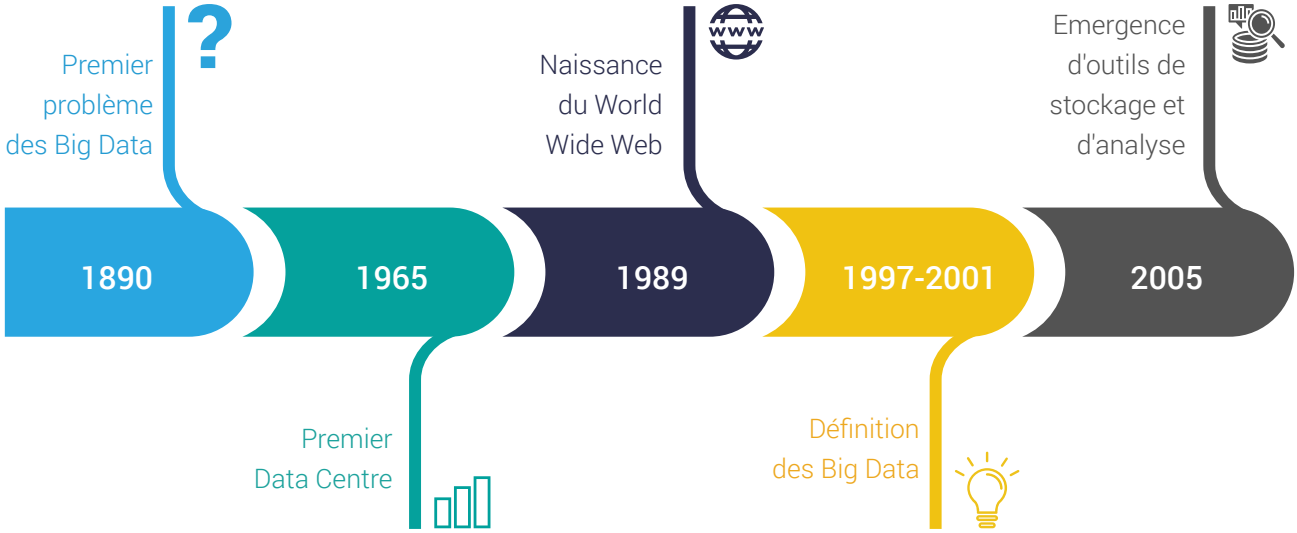
- **Véracité :**

  La diversité des sources et des formes que prennent les Big Data, offre moins de contrôle sur son exactitude. La fiabilité des données a une grande incidence sur leur valeur. Par conséquent, l'un des nouveaux défis identifiés en matière de Big Data est la véracité, qui implique un processus d'élimination des mauvaises données avant le traitement (épuration) et de production d'un ensemble de données exact.

1.2. GENÈSE DES BIG DATA : QUELQUES REPÈRES HISTORIQUES

Le terme Big Data n'est pas entièrement nouveau. Dans un numéro de la revue Harvard Business Review en 2006, Tom Davenport[6] note une méthode utilisée par des organisations telles qu'Amazon, Capital One et les Boston Red Sox pour dominer dans leurs secteurs : l'analyse en tant que différenciateur concurrentiel, « les entreprises étaient submergées de données et de data crunchers » (Davenport 2006). En 2010, Hal Varian a discuté des transactions informatisées, selon lesquelles les transactions économiques impliquent un ordinateur tel qu'un terminal de point de vente, une caisse enregistreuse et, plus récemment, le commerce électronique. Bien que les auteurs n'utilisent pas explicitement le terme «Big Data», le phénomène et les informations auxquels il se réfère entreront par la suite dans la discussion sur les Big Data.

Figure 2 : Chronologie du Big Data



Source : Shi Yong (2014), "Big Data: history, current status, and challenges going forward". Bridge 44(4) : 6–11

1890 : le premier problème des Big Data

En 1890, un recensement du gouvernement américain a été effectué. Les 60 millions d'habitants recensés devaient être comptés manuellement. Herman Hollerith a résolu ce problème avec sa machine de tabulation de pantographes qui s'appuyait sur des cartes perforées pour indiquer des données qui étaient ensuite lues, enregistrées et analysées par sa machine. Cette procédure a permis de réduire les temps de recensements de 10 ans à moins de 24 mois.

1965 : Le premier Data Centre

Encore une fois, le gouvernement américain avait besoin d'un endroit pour stocker 742 millions de déclarations de revenus et 175 millions de jeux d'empreintes digitales. Tous ces enregistrements ont été transférés sur un ordinateur magnétique et stockés sur un grand ordinateur. Bien que ce plan ait été abandonné pour des raisons de sécurité, c'était désormais la naissance du concept de Data Center ou centre de données.

6 Davenport, T. (2006). "Competing on Analytics", Harvard Business Review. https://hbr.org/2006/01/competing-on-analytics.

1989 : la naissance du World Wide Web

L'avènement du web, inventé au centre de recherche CERN par Tim Berners-Lee, offre de nouvelles potentialités de production et d'échange instantané d'informations à l'échelle planétaire. La masse d'information accessible croît à une vitesse impressionnante sous différents formats (texte, image, vidéo, ...) et en différentes langues.

1997-2001 : apparition de l'expression Big Data

Selon les archives de la bibliothèque numérique de l'Association for Computing Machinery[7], l'expression « Big Data » serait apparue en octobre 1997. En 2001, Douglas Laney de Gartner, a décrit les Big Data comme un « défi de données en trois dimensions présentant une grande variété, arrivant en volumes croissants, à grande vitesse ». Cette définition est, depuis, la plus préférée pour décrire ces types de données.

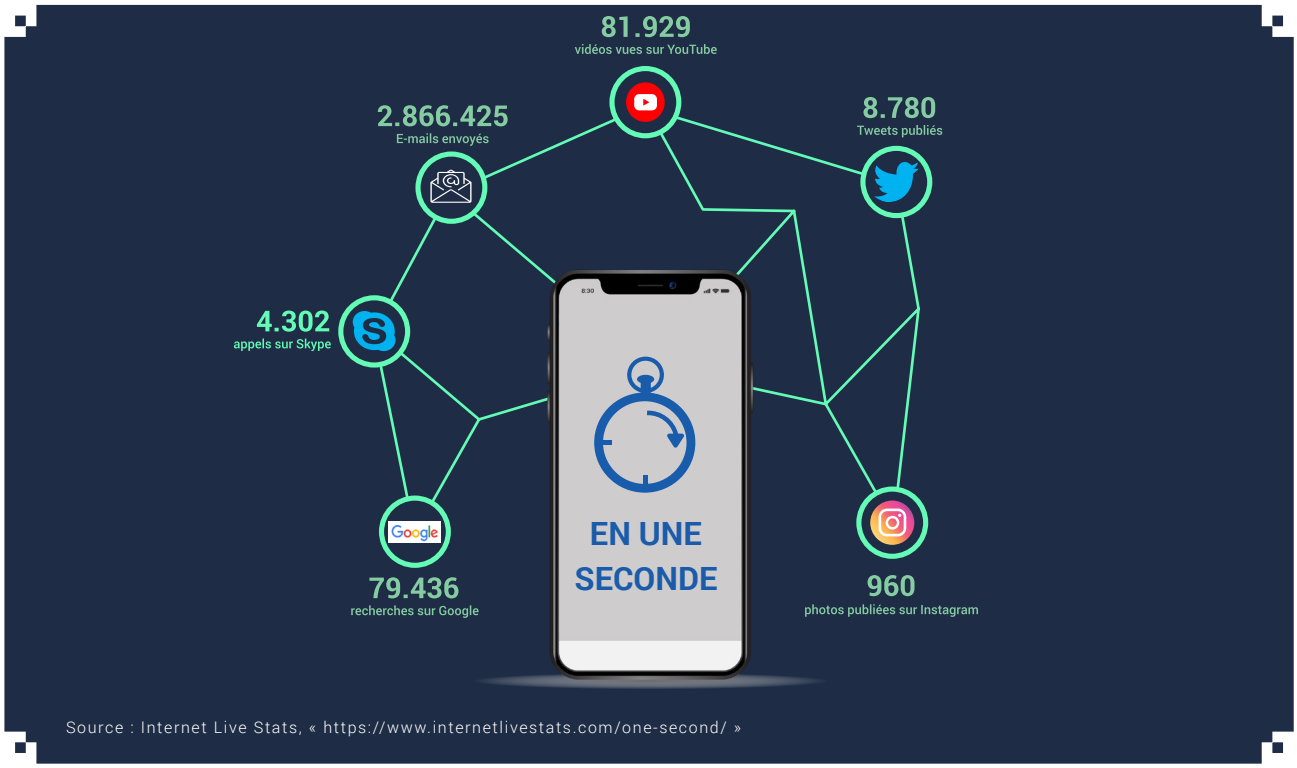
2005 : développement d'infrastructures de stockage et d'analyse dédiées aux Big Data

En 2005, on assiste à une prise de conscience de la masse importante d'information générée par les différents services en ligne, notamment les données des utilisateurs des différentes plateformes de streaming et réseaux sociaux. Cette année, l'outil Hadoop a été créé spécifiquement pour stocker et analyser ces jeux de Big Data.

Période actuelle : dominance des données massives

A l'ère de l'Open Data, d'Internet, des géants du web, du cloud et des réseaux sociaux, le Big Data est devenu essentiel dans le monde d'aujourd'hui. La figure ci-dessous expose quelques «Big chiffres» qui montrent l'incroyable masse de données qui nous entourent quotidiennement ("Internet Live Stats, 2019").

7 Gil Press, « A Very Short History Of Big Data », Forbes, 9 mai 2013.



2. TYPES DES SOURCES DE DONNÉES ET DES BIG DATA

Les données utilisées à des fins statistiques peuvent être tirées de diverses sources, qu'il s'agisse de sources traditionnelles telles que les recensements et les enquêtes statistiques, les sources administratives et d'autres sources alternatives à l'instar des Big Data. La section en cours décrit les principales caractéristiques de ces trois catégories de sources de données.

2.1. CARACTÉRISTIQUES DES DONNÉES D'ENQUÊTES ET DES DONNÉES ADMINISTRATIVES

Les enquêtes, exhaustives telles que le recensement de population, ou par sondage comme l'enquête sur l'emploi, constituent un outil important de collecte des données. Elles sont habituellement réalisées dans un but précis auprès d'un échantillon d'unités, pour atteindre un objectif économique, scientifique ou public donné. De même, les organismes gouvernementaux recueillent et conservent des données à des fins administratives, comme les données fiscales des personnes physiques et des entreprises, ou le registre de l'état civil. Suite à la demande croissante des données statistiques, ainsi que du fardeau de réponse et des coûts associés à la collecte, les sources administratives sont devenues des sources de données de plus en plus recherchées à des fins statistiques. Généralement, ces deux sources administratives et d'enquêtes peuvent se compléter afin de fournir des données plus complètes.

- **Données administratives**

L'American Statistical Association[8] (1977) définit les données administratives comme « les données collectées et conservées dans le but de prendre des mesures ou de contrôler les actions d'une personne ou d'une autre entité». Aux pays développés, les données administratives ont une longue histoire d'utilisation dans la production des statistiques gouvernementales. Avec le progrès technologique, le traitement des grands ensembles de données par les organismes de statistique est devenu facile, encourageant une utilisation encore plus grande de ces données à des fins de recherche.

[8] Rob Kitchin, The opportunities, challenges and risks of big data for official statistics, National University of Ireland, Maynooth, (2015).

En tant qu'une source d'information potentielle, les données administratives ont de nombreuses utilisations, notamment la tabulation directe et l'estimation indirecte de modèles ou d'autres statistiques, ainsi que la construction de bases de sondage et l'évaluation des résultats de l'enquête (Brackstone, 1987). Ces données administratives peuvent présenter plusieurs avantages par rapport aux données d'enquêtes traditionnelles, notamment une couverture plus complète d'une population, de faibles coûts de collecte de données, une réduction des charges des répondants et une meilleure qualité des données. Les problèmes potentiels liés à l'utilisation des données administratives à des fins statistiques comprennent la stabilité d'un programme dans le temps, les problèmes de confidentialité concernant l'utilisation non administrative des données, les problèmes conceptuels relatifs à la population et aux éléments collectés, et les coûts de transformation des données sous une forme utile à des fins de recherche.

- **Les données d'enquêtes**

Les enquêtes diffèrent des données administratives par leurs objectifs, et ces différences ont souvent des implications sur leur structure statistique, leur cadre conceptuel et leur contenu. Presque toutes les enquêtes sont menées pour répondre à des catégories spécifiques de recherche ou à des questions de politique publique par rapport à l'exercice d'une fonction administrative. Cette différence d'objectif se reflète dans la population étudiée, l'unité d'observation, la taille de l'échantillon et la portée de données. Certains avantages des données d'enquête par rapport aux données administratives incluent le ciblage d'une population spécifique et des variables d'intérêt, l'interaction avec le répondant et la possibilité de garantir que les données seront utilisées uniquement à des fins statistiques (c'est-à-dire non administratives). Les problèmes potentiels avec les données d'enquêtes comprennent les difficultés à construire une base de données appropriée, le manque de participation légalement obligatoire, les coûts élevés de l'augmentation de la taille de l'échantillon, la non-réponse des unités et des articles et l'erreur de mesure.

2.2. TYPOLOGIE DES BIG DATA

Les classifications des Big Data sont multiples. On se focalisera ici sur deux d'entre elles, à savoir la classification selon le volume des données (celle de Doornik et Hendry) et la classification selon la structure des données (celle de la Commission Economique pour l'Europe des Nations Unies (UNECE)).

- **La classification des Big selon le volume des données (Doornik et Hendry-2015) :**

La classification de Doornik et Hendry[9] (2015) identifie trois principaux types de données volumineuses : Grand, Gros et Enorme.

› **Grand** : Un nombre important d'observations (T) pour un nombre limité de variables (N), soit T>N. C'est le cas des données relevées sur des transactions financières ou des requêtes de recherche sélectionnées. Dans ce cas, le nombre d'observations est effectivement très grand dans l'échelle de temps initiale, disons secondes, mais il convient de déterminer s'il est également assez grand dans l'échelle de temps par rapport à la variable macroéconomique cible de l'exercice de prévision immédiate, par exemple les trimestres.

› **Gros** : Un nombre important de variables pour un nombre limité d'observations, N>T. les grandes bases de données transversales entrent dans cette catégorie. Ces ensembles de données pourraient être utiles du point de vue de la prévision immédiate si le nombre d'observations est suffisamment grand ou si les variables permettent une estimation correcte du modèle (par exemple, au moyen de méthodes de panel).

[9] Doornik, J. A. and D.F. Hendry (2015), 'Statistical Model Selection with Big Data', Cogent Economics & Finance , 3(1), 2015.

› **Énorme** : de nombreuses variables et de nombreuses observations, c'est-à-dire très grand N et T. Ce type de données est idéal dans le contexte de prévision immédiate. Le principal inconvénient est que la collecte de données massives a débuté relativement récemment (au cours de la dernière décennie) et ne permet pas une validation croisée longue et immédiate. Google Trends, résumés accessibles au public d'un grand nombre de requêtes de recherche spécifiques dans Google, constituent peut-être le meilleur exemple de cette catégorie et non par hasard les indicateurs les plus couramment utilisés dans les exercices de prévision économique immédiate.

Tableau 1 : La classification de Doornik and Hendry (2015)

Type	Grand	Gros	Enorme
Marché financiers			✕
Paiements électroniques			✕
Téléphones portables	✕		✕
Données des capteurs	✕		✕
Images des satellites		✕	
Prix du scanner	✕		✕
Prix en ligne	✕		✕
Recherches en ligne		✕	

Source: Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). "Big data types for macroeconomic nowcasting". Eurostat Review on National Accounts and Macroeconomic Indicators, 1(2017), 93-145

- **La classification selon la structure des données (Commission Economique pour l'Europe des Nations Unies-UNECE) :**

Une deuxième possibilité pour classer les Big Data consiste à identifier le contenu des données. Une taxonomie particulièrement utile est fournie par la Division des Statistiques de la Commission Economique pour l'Europe des Nations Unies (UNECE), qui identifie trois principaux types de données volumineuses :

» **Réseaux sociaux** (informations d'origine humaine) : ces informations sont le récit d'expériences humaines, stockées presque entièrement sous forme numérique dans des ordinateurs personnels ou des réseaux sociaux. Les données, généralement structurées de manière souple et souvent non régies, comprennent :

- Réseaux sociaux : Facebook, Twitter, Tumblr, etc.
- Blogs et commentaires
- Documents personnels
- Images : Instagram, Flickr, Picasa, etc.
- Vidéos : YouTube etc.
- Recherches sur internet
- Contenu de données mobiles : messages texte
- Cartes générées par l'utilisateur
- E-mail

» **Systèmes de gestion traditionnels** (données médiées par processus) : ces processus enregistrent et surveillent les événements d'affaires présentant un intérêt, tels que l'enregistrement d'un client, la fabrication d'un produit, la prise de commande, etc. hautement structuré et inclut les transactions, les tables de référence et les relations, ainsi que les métadonnées qui définissent son contexte. Les données d'entreprise traditionnelles constituent la grande majorité de ce que les systèmes d'information gèrent et traitent, à la fois dans les systèmes opérationnels et dans les systèmes de l'informatique décisionnelle. Habituellement structuré et stocké dans des systèmes de bases de données relationnelles, y compris aussi des "données administratives", il peut être regroupé dans :

  - Données produites par les agences publiques : dossiers médicaux, assurances sociales…
  - Données produites par les entreprises : transactions commerciales, archives bancaires / stocks, commerce électronique, cartes de crédit, etc.

» **Internet des objets** (données générées par machine) : dérivé des capteurs et des machines utilisés pour mesurer et enregistrer les événements et les situations dans le monde physique. Il devient une composante de plus en plus importante des informations stockées et traitées par de nombreuses entreprises. Sa nature bien structurée convient au traitement informatique, mais sa taille et sa rapidité vont au-delà des approches traditionnelles.

  - Données provenant de capteurs : capteurs fixes (domotique, capteurs météo / de pollution, capteurs de trafic / webcam, etc.) ou capteurs mobiles (suivi : localisation du téléphone portable, voitures, images satellite, etc.)
  - Données provenant de systèmes informatiques : journaux, journaux Web, etc.

Du point de vue économique, les trois types des Big Data sont potentiellement pertinents. Par exemple, des recherches Internet sélectionnées et/ou des tweets (réseaux sociaux), des transactions par carte de crédit (systèmes commerciaux traditionnels) ou le nombre de navires de commerce naviguant dans un certain domaine (Internet des objets) pourraient tous fournir des indicateurs avancés utiles pour l'estimation de l'activité d'un pays.

3. BIG DATA : CADRE RÉGLEMENTAIRE

Le développement des Big Data et l'émergence de leur champ d'application se sont accompagnés par un effort important de la part des Etats, des institutions internationales et des organes publics et privés pour asseoir un cadre réglementaire et juridique à toutes les phases de la chaîne numérique. En outre, ces types de données soulèvent un ensemble de questions et d'enjeux réglementaires au regard de la nature des informations utilisées qui revêtent dans la majorité des cas un caractère personnel. Ces enjeux sont essentiellement liés à la propriété des données, à leur confidentialité et sécurité, à leur collecte et stockage, à l'échange et au transfert à un pays tiers au lieu de leur production.

La problématique des données personnelles se trouve ainsi au centre des préoccupations et au carrefour des possibilités du cadre juridique de l'utilisation des données numériques. L'usage des données à caractère personnel peut, en effet, porter préjudice à la vie privée et à l'intimité des individus et est, ainsi, régit, dans la plupart des pays, par des lois de protection des données personnelles. A cet égard, si lesdites lois interdisent

toute utilisation des données à caractère personnel, elles autorisent toutefois, le profilage de données[10], concept mentionné dans l'article 4 du Règlement Général sur la Protection des Données (RGPD) de l'Union Européenne.

Le cadre juridique des données numériques est essentiellement constitué d'obligations : obligations déclaratives liées au traitement des données personnelles, obligations d'acheminement des données, obligations de conservation et de communication des données pour le compte des autorités publiques et obligations de mise à disposition des données publiques.

3.1. UNION EUROPÉENNE : UNE RÉGLEMENTATION COUVERTE PAR LE RGPD

Le Règlement Général sur la Protection des Données (RGPD), adopté par l'Union Européenne en mai 2018, a été institué pour fédérer les législations et réglementations en vigueur dans les pays membres de l'UE au sujet de la gestion des données personnelles. Il présente un cadre réglementaire pour la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. Son objectif est de protéger les personnes, de contrôler leurs propres données personnelles et d'engager les responsables du traitement des données privées des citoyens. Il adopte une approche basée sur les risques ainsi qu'une analyse d'impact qui vise la protection des données. En outre, le RGPD permet d'accompagner et réglementer les étapes de la chaîne de valeur du Big Data.

Le RGPD s'applique aux entreprises, aux organismes publics et aux associations, de toutes tailles, quelles que soient leurs activités, du moment qu'ils traitent des données à caractère personnel de citoyens européens. Il couvre aussi bien les entreprises établies sur le territoire de l'UE que les responsables de traitement et sous-traitants hors UE ainsi que les grands agrégateurs de données, les sociétés de télécommunications et les entreprises de services informatiques.

Le RGPD couvre toute la chaine de valeur des méga données constituée de 4 étapes à savoir : (i) la collecte de données, (ii) le stockage des données, (iii) l'agrégation de données et (iv) l'analyse et l'utilisation des résultats. En effet, chaque activité est soumise à des règles spécifiques[11] : la licéité, la loyauté et la transparence, la limitation de la finalité, la minimisation des données, l'exactitude et l'actualisation, la limitation de la conservation et l'intégrité et la confidentialité.

Il cible toutes les opérations de traitement des données à caractère personnel effectuées à l'aide de procédés automatisés[12], constituées essentiellement des opérations de collecte, d'enregistrement, d'organisation, de structuration, de conservation, d'adaptation ou de modification, d'extraction, de consultation, d'utilisation, de communication par transmission, de diffusion ou toute autre forme de mise à disposition, de rapprochement ou d'interconnexion, de limitation, d'effacement ou de la destruction.

---

10 Le profilage est « toute forme de traitement automatisé de données à caractère personnel consistant à utiliser ces données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment pour analyser ou prédire des éléments concernant le rendement au travail, la situation économique, la santé, les préférences personnelles, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique ».

11 Article 5 du Règlement Général sur la Protection Données.

12 Article 4, paragraphe 2, du Règlement Général sur la Protection Données. Les types de traitements concernent les opérations de collecte, d'enregistrement, d'organisation, de structuration, de conservation, d'adaptation ou modification, d'extraction, de consultation, d'utilisation, de communication par transmission, de diffusion ou toute autre forme de mise à disposition, de rapprochement ou d'interconnexion, de limitation, d'effacement ou la destruction.

3.2. CHINE : PROTECTION DES DONNÉES PERSONNELLES À TRAVERS LA LOI SUR LA CYBER-SÉCURITÉ

La Chine a développé son arsenal législatif sur la cyber sécurité et la protection des données personnelles depuis l'apparition de l'économie du Big Data. Les lois les plus significatives ont été réalisées depuis 2012 avec une loi sur la cyber sécurité entrée en application le 1er juin 2017. Comme aux États-Unis, et à la différence de l'Europe et de nombreux pays, la Chine a une approche sectorielle de la protection des données. Autrement, au lieu d'avoir une grande loi couvrant tous les aspects de la protection des données personnelles comme le RGPD, les dispositions sur le sujet sont disséminées dans plusieurs textes. Une des dimensions importantes de la CSL est la contrainte de stocker les données importantes et les données à caractère personnel sur des serveurs localisés sur le territoire chinois.

3.3. INDE : PROJET DE PERSONAL DATA PROTECTION BILL

Le Projet de Personal Data Protection Bill (PDPB), Consacre de nouveaux droits sensiblement similaires à ceux confirmés ou consacrés par le RGPD, notamment le droit d'accès, le droit de rectification, le droit à la portabilité des données personnelles et le droit à l'oubli. Le PDPB[13] est soumis au contrôle de l'autorité indienne de protection de données où il est applicable aux acteurs publics et privés. Le traitement des données concerne : l'utilisation, le partage, la divulgation, la collecte à l'intérieur du territoire indien, le traitement dans le cadre d'une activité exercée en Inde indépendamment du lieu de leur établissement. En outre, le traitement renvoi à toute donnée personnelle collectée, utilisée, partagée, divulguée ou encore traitée par une personne morale indienne ou par un citoyen indien, que le traitement ait lieu ou non en Inde.

Un ensemble de pays a adopté une réglementation similaire et qui s'inspire du Règlementation européenne afin de se mettre en conformité avec le RGPD et d'assurer une globalisation des échanges avec l'Europe et d'augmenter ainsi l'attractivité économique des dites. D'autre pays, tel que les Etats-Unis, ont signés des protocoles d'échanges avec l'union européenne pour cadrer les opérations de transfert ou d'échange des données à caractère personnelles entre ces pays.

4. REVUE DES PRINCIPAUX TRAVAUX POUR LE SUIVI DE L'ACTIVITÉ ÉCONOMIQUE, LA PRÉVISION À COURT TERME

Dans cette section, il serait question de présenter les applications existantes des techniques statistiques et d'exploration de données pour la prévision avec les Big Data. Elle résume ces éléments en fonction du domaine et du sujet (le cas échéant) afin de fournir une expérience plus enrichissante.

4.1. PRÉVISION ET SUIVI DU PIB ET SES COMPOSANTES

On commencera tout d'abord par la famille des travaux s'intéressant à la mesure et la prévision de l'agrégat le plus important d'une économie, à savoir le Produit Intérieur Brut (PIB).

[13] Informations juridiques, Inde, le personal data protection bill, 2018 : « émergence d'une politique de protection des données personnelles ».

Prévoir la croissance du PIB à partir des articles de journaux, l'INSEE de France

Le premier travail entrant dans cette catégorie est celui de Bortoli[14] et al. (2018) qui consiste à prévoir la croissance du PIB en lisant le journal. Les statistiques du PIB en France sont publiées trimestriellement, 30 jours après la fin du trimestre. Dans cet article, les auteurs considèrent le contenu des médias comme une source de données complémentaire aux outils conjoncturels classiques pour améliorer les prévisions du PIB français.

Les données de plus d'un million d'articles publiés dans le journal « Le Monde » entre 1990 et 2017 ont été utilisées pour créer un nouvel indicateur synthétique de « sentiment médiatique » sur l'état de l'économie. En mettant l'accent sur la prévision du PIB à court terme, un «modèle médiatique» (modèle autorégressif augmenté de l'indicateur de sentiment des médias) a été comparé avec un modèle autorégressif simple et un modèle autorégressif augmenté de l'indicateur INSEE de climat des affaires fondé sur des enquêtes de conjoncture menées auprès des chefs d'entreprise.

L'ajout d'un indicateur médiatique améliore les prévisions du PIB français par rapport à ces deux modèles de référence. Par la suite, une approche automatisée par régression pénalisée a été testée, où les auteurs utilisent les fréquences d'apparition des mots ou d'expressions dans les articles plutôt qu'une information agrégée. Cette approche, étant plus aisée à mettre en œuvre, elle apporte cependant des résultats inférieurs.

La prévision du PIB chinois : contenu informatif des données économiques et financières, Hong Kong Institute for Monetary Research

Matthew S. et Kenneth K. (2011)[15] appliquent le modèle factoriel proposé par Giannone, Reichlin et Small (2005) à un ensemble de données volumineuses permettant de prédire en temps réel (c'est-à-dire les prévisions du trimestre en cours) le taux de croissance du PIB trimestriel de la Chine. L'information utilisée est constituée de 189 séries d'indicateurs de plusieurs catégories, telles que les prix, la production industrielle, les investissements en immobilisations, le secteur extérieur, le marché monétaire et le marché financier. Les auteurs appliquent également les critères de Bai et Ng. (2002) pour déterminer le nombre de facteurs communs dans le modèle factoriel.

Le modèle identifié génère des prévisions en temps réel hors échantillon pour le PIB chinois avec des moyennes des erreurs de prévision au carré moins grandes que celles du modèle de référence, à savoir la marche aléatoire (Random Walk). De plus, en utilisant le modèle factoriel, les auteurs constatent que les données de taux d'intérêt constituent le bloc le plus important pour l'estimation du PIB du trimestre en cours en Chine. Les données sur les prix à la consommation et au détail et les indicateurs d'investissement en immobilisations constituent d'autres blocs importants.

La prévision macroéconomique en temps réel en utilisant les probabilités de Google, BCE

Koop et Onorante[16] (2013) suggèrent de procéder à une prévision immédiate en utilisant des méthodes de sélection de modèle dynamique (DMS) qui permettent la commutation de modèle entre des modèles de régression à paramètres variables dans le temps. Ceci est potentiellement utile dans un environnement d'instabilité des coefficients et de paramétrisation excessive pouvant survenir lors de la prévision avec des

---

14 Bortoli, C., Combes, S. & Renault, T. (2018). "Nowcasting GDP Growth by Reading Newspapers". INSEE, Economie et Statistique / Economics and Statistics, 505-506, 17–33.

15 Matthew S. et Kenneth K. (2011), "Nowcasting Chinese GDP: Information Content of Economic and Financial Data.", Hong Kong Institute for Monetary Research, HKIMR Working Paper No.04/2011.

16 Koop, G. and L. Onorante (2013), "Macroeconomic Nowcasting Using Google Probabilities", European Central Bank Presentation.

indicateurs de recherche sur Google. Ces variables permettent que le changement de modèle soit contrôlé par les indicateurs de recherche sur Google à travers des probabilités de recherche déduites de ces derniers. En d'autres termes, au lieu d'utiliser les indicateurs de recherche sur Google comme variables explicatives, ces probabilités permettent de déterminer le modèle de prévision à utiliser à chaque instant.

Dans un exercice empirique impliquant neuf variables[17] macroéconomiques mensuelles aux États-Unis, les auteurs ont constaté que les méthodes DMS apportaient de grandes améliorations à la prévision en temps réel. L'utilisation des probabilités tirées de Google dans le modèle DMS est souvent plus performante que le DMS conventionnel.

Prévision en temps réel du PIB avec les données de paiements électroniques, BCE

Galbraith et Tkacz[18] (2015) évaluent l'utilité d'un vaste ensemble de données de paiements électroniques comprenant des transactions par carte de débit et de crédit, ainsi que des chèques émis dans le système bancaire, en tant qu'indicateurs potentiels de la croissance actuelle du PIB au Canada. Ces variables[19] capturent un large éventail d'activités de dépenses et sont disponibles très rapidement, ce qui en fait des indicateurs actuels appropriés. Bien que chaque transaction effectuée avec ces mécanismes de paiement soit en principe observable, les données sont agrégées pour les prévisions macroéconomiques.

En contrôlant les dates de diffusion de chacun des indicateurs, les auteurs génèrent des prévisions immédiates de la croissance du PIB pour un trimestre donné sur une période de cinq mois, soit la période au cours de laquelle un intérêt pour les prévisions immédiates existerait. Il est constaté que les erreurs de prévision immédiate chutent d'environ 65% entre la première et la dernière prévision immédiate. Parmi les variables de paiement considérées, les transactions par carte de débit semblent apporter les améliorations les plus importantes en termes de précision des prévisions.

4.2. EMPLOI ET STATISTIQUES DU CHÔMAGE

La seconde famille de travaux que nous exposerons dans cette note est liée à la mesure et la prévision du niveau de chômage.

Le pouvoir prédictif de l'indice de recherches sur Google dans la prévision du chômage, Banque d'Italie

D'Amuri et Marcucci[20] (2012) suggèrent l'utilisation d'un indice de l'intensité de la recherche d'emploi sur Internet (Google Index, GI) comme meilleur indicateur avancé pour prédire le taux de chômage mensuel aux États-Unis. Ils effectuent une comparaison approfondie des prévisions hors échantillon en analysant de nombreux modèles qui adoptent leur indicateur principal, les revendications initiales les plus standards ou une combinaison des deux. Les auteurs constatent que les modèles enrichis avec l'indice de Google dépassent les modèles traditionnels en prédisant le taux de chômage pour différents intervalles hors échantillon qui commencent avant, pendant et après la grande récession. Les modèles basés sur l'indice de Google dépassent également les modèles standards dans la plupart des prévisions au niveau des États et par rapport à l'Enquête sur les prévisionnistes professionnels. Ces résultats survivent à un test de falsification et sont également confirmés lorsque différents mots clés sont utilisés.

---

17 Inflation, Inflation salariale, Chômage, Spread Terme, Inflation du prix des produits de base, Production industrielle, Indice des conditions financières (FCI), Inflation du prix du pétrole, masse monétaire,…

18 Galbraith, J.W and G. Tkacz (2015), 'Nowcasting GDP with electronic payments data', European Central Bank, Working Paper No 10 / August 2015.

19 Transactions par carte de débit et par carte de crédit, chèques, PIB décalé, indice du logement, emplois dans les ventes aux particuliers et aux entreprises, indice boursier, FCI, masse monétaire, semaine moyenne de travail (heures), nouvelles commandes, biens durables, stocks, commerce de détail.

20 D'Amuri, F. and J. Marcucci (2012), 'The Predictive Power of Google Searches in Predicting Unemployment', Banca d'Italia Working Paper, 891.

L'utilisation des données sur les activités Web pour améliorer les statistiques du chômage, Eurostat

Reis[21] et al. (2015) analysent l'activité Web comme une source de données volumineuse. Les traces électroniques laissées par les utilisateurs pendant qu'ils utilisent des services Web pourraient être utilisées comme données en temps réel ou avec un très petit décalage. Du fait que la majorité des activités humaines mesurées par les statistiques officielles sont étroitement liées au comportement des internautes en ligne, ces données sur l'activité des internautes offrent la possibilité de produire des prévisions d'indicateurs socioéconomiques afin d'accroître l'actualité des statistiques. Des articles dans la littérature ont démontré que ces prédictions peuvent être faites. Cependant, ce type de données devrait être davantage contrôlé quant à sa transparence, sa continuité, sa qualité et son potentiel d'intégration avec les méthodes traditionnelles de la statistique officielle. L'application empirique que les auteurs ont mise en œuvre est une meilleure prévision immédiate du chômage français et italien.

Amélioration de la prévision des statistiques du chômage avec Google Trends, Eurostat

Ferreira[22] (2015) utilise un modèle à facteurs dynamiques pour extraire une variable latente des données de Google Trends, ce qui est un bon indicateur de la dynamique du chômage. Les modèles de prévision du chômage qui utilisent la variable latente estimée ont donné de meilleurs résultats que les approches proposées dans les travaux précédents, en particulier pendant une période où la tendance a été brutalement modifiée.

4.3. INDICES DES PRIX ET INFLATION

Dans cette sous-section on s'intéressera aux travaux permettant la mesure et la prévision des indices des prix et, notamment, l'inflation à travers les Big Data.

Utilisation de données du Web Scraping pour construire des indices des prix à la consommation, Office for National Statistics (ONS), Royaume Uni

Breton[23] et al. (Office for National Statistics (ONS), UK) fournissent un aperçu des recherches de l'ONS sur le potentiel d'utilisation des données extraites sur le Web pour les statistiques des prix à la consommation. La recherche couvre la collecte, la manipulation et l'analyse de données extraites sur le Web. Les principaux avantages des données Web récupérées sont la réduction des coûts de collecte, la couverture accrue (plus d'éléments du panier), fréquence accrue, production de nouveaux indices et une meilleure capacité à répondre aux nouveaux défis.

L'ONS utilise des données extraites sur le Web pour calculer 3 types d'indices de prix, ceux qui augmentent le nombre d'articles utilisés, le nombre de jours considérés et à la fois le nombre d'articles et les jours considérés. La construction de ce type d'indices peut être utile aux économistes et aux décideurs.

Collecte de données de l'habillement sur Internet pour le suivi de l'IPC, Netherlands Technical Report

Griffioen[24] et al. (Statistiques Pays-Bas) se sont intéressés à l'utilisation des prix en ligne des vêtements pour l'analyse de l'IPC. Cette étude entre dans la catégorie du web scraping et présente les résultats et les difficultés

---

21 Reis, F., P. Ferreira and V. Perduca (2015), 'The Use of Web Activity Evidence to Increase the Timeliness of Official Statistics Indicators', Eurostat Working Paper.

22 Ferreira, P. (2015), 'Improving Prediction of Unemployment Statistics with Google Trends: Part 2', Eurostat Working Paper.

23 Breton, R., N. Swiel and R. O'Neil (2015), 'Using Web Scraped Data to Construct Consumer Price Indices', New Techniques and Technologies for Statistics, Eurostat Conference, 9–13 March 2015.

24 Griffioen, R., J. de Haan and L. Willenborg (2014), 'Collecting Clothing Data from the Internet', Statistics Netherlands Technical Report.

de la collecte de prix en ligne sur une période de deux ans. Les prix des vêtements raclés sur le Web permettent de réduire le coût de la collecte des prix dans les magasins physiques, d'augmenter la taille de l'échantillon et d'avoir des informations de bonne qualité. Toutefois ce type de collecte de données est sensible aux modifications apportées au site Web et au choix de la stratégie du web scraping.

Utilisation des données de prix du Web Scraping pour la construction de l'indice des prix du commerce électronique, les Nations Unis

Dans un autre travail, Jiang Shu[25] a utilisé des données du Web-scraping pour construire l'indice de prix du commerce électronique. Il s'agit d'analyser les données de prix des téléphones portables spécifiques par programme Crawler et établir l'indice de prix quotidien comme référence pour les données de prix mensuelles.

4.4. SUIVI DE L'ACTIVITÉ ÉCONOMIQUE

Concernant le suivi de l'activité économique dans son ensemble, nous présenterons dans cette sous-section les principaux travaux officiels qui ont utilisé les Big Data comme source d'information.

Utilisation des données de la télédétection par satellite et par antenne pour compléter les statistiques agricoles, les Nations Unis

Dans un projet pilote destiné à passer en production pour remplacer les données existantes, Jiang Shu[26] (Bureau national de statistique, Chine), a mené une enquête sur les cultures par terre agricole en utilisant les données de la télédétection par satellite et aérienne pour aider à estimer les statistiques agricoles. Premièrement, l'auteur a construit le cadre d'échantillonnage spatial en utilisant les données des enquêtes sur l'utilisation des terres et du recensement de l'agriculture. Ensuite, le cadre d'échantillonnage a été mis à jour à travers la télédétection par satellite et aérienne. En utilisant les échantillons sélectionnés par la méthode d'échantillonnage spatial, la zone de plantation des cultures et la production chaque saison a été estimée.

Utilisation des données de Google Trends dans les enquêtes mensuelles de la Banque de France sur le commerce de détail, INSEE

Dans le cadre du partenariat la liant à la Banque de France, la Fédération du e-commerce et de la vente à distance (FEVAD) fournit mensuellement le chiffre d'affaires réalisé en e commerce auprès des particuliers, depuis 2012. Dans l'attente des livraisons, la Banque de France procède à des estimations, dont l'enjeu est renforcé par la croissance du e-commerce. Le modèle autorégressif (SARIMA (12)) utilisé peut désormais être complété par d'autres modèles statistiques s'appuyant sur des données exogènes grâce à un historique plus long de données. Le travail de François Robin[27] détaille les différents choix opérés conduisant à la prévision finale : transformation des données, modèles à sélection de variables et stratégie pour la prévision. Les requêtes sur Google notamment, mesurées par Google Trends, permettent d'améliorer la capacité prédictive du modèle final, obtenu en combinant les modèles simples.

---

25 Jiang S. (2014), "Using web scraping price data for price index of e-commerce", United Nations, Big Data Project Inventory.

26 Jiang S. (2014), "Crop survey by farmland: using satellite and aerial remote sensing to help estimate agricultural statistics", United Nations, Big Data Project Inventory.

27 Robin, F. (2018). "Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys.", INSEE Economie et Statistique / Economics and Statistics, 505-506, 35–63.

Utilisation des données de téléphonie mobile pour les statistiques du tourisme et du transport, Statistics Belgium

Marc Debusschere[28] (Statistiques de Belgique, Belgique) évalue la possibilité d'utiliser les données de la téléphonie mobile pour compléter, voire remplacer, les sources de données de produits statistiques, principalement dans les domaines du tourisme ou des statistiques du transport. Il s'agit d'explorer la possibilité d'enregistrer des phénomènes non encore accessibles par les méthodes traditionnelles. Plusieurs volets ont été traités dont l'exploration des données, projet pilote destiné à passer en production pour améliorer la rapidité d'exécution, projet pilote destiné à passer en production pour compléter les données existantes et projet pilote destiné à passer en production pour remplacer les données existantes.

Estimation de la population résidente à partir des données des téléphones portables, INSEE

De nombreux travaux s'intéressent à l'utilisation des données issues de la téléphonie mobile pour construire des indicateurs statistiques. Ces données ont l'intérêt de fournir des informations à la fois à une résolution spatiale élevée et à une haute fréquence. Plusieurs applications proposent par exemple de mesurer la population présente à des niveaux spatiaux ou temporels fins. L'exploitation de ces données pour construire des indicateurs statistiques soulève néanmoins des difficultés : les données d'un seul opérateur ne sont pas représentatives de la population totale, et ces données anonymes sont souvent pauvres en caractéristiques sociodémographiques ce qui limite la qualité des redressements.

Le travail de Sakarovitch[29] et al. (2018) s'appuie sur un fichier issu des enregistrements des activités d'abonnés d'un grand opérateur français pour donner un premier aperçu du potentiel mais aussi des problèmes posés par de telles données, illustré par l'estimation d'indicateurs de populations résidentes inférés à partir du simple enregistrement des activités des personnes.

Mesurer l'activité économique en Chine en utilisant les données de téléphonie mobile, Chine

Dong[30] L. et al. (2017) explorent le potentiel d'utilisation des données de téléphonie mobile pour mesurer l'activité économique en Chine dans une perspective ascendante. Les nouvelles tendances en matière d'utilisation des smartphones, d'applications de cartographie en ligne et de médias sociaux, ainsi que des données géolocalisées qu'ils génèrent, offrent la possibilité de suivre les activités socio-économiques des utilisateurs d'une manière sans précédent et granulaire, et ont déclenché une révolution dans la recherche empirique. Ces vastes données mobiles offrent de nouvelles perspectives et approches pour mesurer la dynamique économique et élargissent les domaines des sciences sociales et de l'économie. Premièrement, les auteurs ont construit des indices permettant de jauger les tendances de l'emploi et de la consommation à partir de milliards de données de géo-positionnement.

Par la suite, l'estimation du trafic piétonnier dans les magasins hors ligne a été effectuée à l'aide de données de recherche d'emplacement dérivées de Baidu et Maps, qui sont ensuite appliquées pour prévoir les revenus d'Apple en Chine et pour détecter avec précision les fraudes au guichet. Troisièmement, les auteurs ont construit des indicateurs de consommation pour suivre les tendances dans diverses industries du secteur des services

---

28 Marc D., "Feasibility study on the use of mobile telephone data for tourism & transportation statistics", Belgium - Statistics Belgium, United Nations, Big Data Project Inventory.

29 Sakarovitch, B., Bellefon, M. (de), Givord, P. & Vanhoof, M. (2018). "Estimating the Residential Population from Mobile Phone Data, an Initial Exploration". INSEE, Economie et Statistique / Economics and Statistics, 505-506, 109–132.

30 V Dong L. et al. (2017), "Measuring economic activity in China with mobile Big Data", Big Data Lab, Baidu Research, Baidu, Beijing, China.

et les vérifier avec plusieurs indicateurs existants. Il s'agit de la première étude à mesurer la deuxième plus grande économie du monde en exploitant des données temporelles spatiales d'une taille et d'une granularité sans précédent. De cette manière, cette recherche fournit de nouvelles approches et de nouvelles perspectives pour mesurer l'activité économique.

4.5. SECTEUR FINANCIER

S'intéressant au secteur financier, de nombreux travaux ont analysé et établi des prévisions de ce secteur vu la masse de données disponible en temps réel. Les paragraphes ci-dessous exposent les principaux travaux qui en découlent.

Utilisation des approches de traitement textuel pour prédire les rendements des actions

Heston et Sinha[31] (2014) utilisent un ensemble de données de plus de 900 000 articles pour tester si les informations peuvent prédire les rendements des actions. Les auteurs constatent que les entreprises sans nouvelles ont des rendements futurs moyens nettement différents des entreprises avec des nouvelles. Confirmant les résultats de la littérature, les nouvelles quotidiennes prédisent des rendements boursiers pour seulement 1-2 jours. Mais les nouveaux hebdomadaires prédisent des rendements boursiers pour un quart d'année. Les reportages positifs augmentent rapidement les rendements boursiers, mais les reportages négatifs ont une réaction longtemps retardée.

Exploitation des Big Data pour l'évaluation des risques systémiques, BCE

Nyman[32] et al. (2014) ont étudié les moyens d'utiliser les Big Data dans la gestion du risque systémique. Les nouvelles et les récits sont les principaux moteurs de l'activité économique et financière. Leurs données de presse comprennent les commentaires quotidiens sur les événements du marché, les rapports hebdomadaires de recherche économique et les nouvelles de Reuters. L'apprentissage automatique et les composantes principales sont inclus dans la méthodologie afin de calculer les indices de consensus basés sur les sources ci-dessus. Cette étude conclut que les rapports de recherche économique hebdomadaires pourraient potentiellement prévoir l'indice du consommateur du Michigan et que des commentaires quotidiens sur les événements de marché pourraient potentiellement prévoir la volatilité du marché.

Apport de la diffusion de l'information dans la perspective de la bulle boursière, Université de Miami

Andrade[33] et al. (2009) analysent le rôle des analystes et de la diffusion de l'information dans la perspective de la bulle boursière chinoise de 2007. Ils ont utilisé la corrélation entre les différentes mesures d'intensité de la bulle et sa couverture par les analystes en tant que mesure de la diffusion de l'information et l'indice de recherche Google pour vérifier leur timing et leur intensité. Ils ont trouvé une relation négative significative entre l'intensité de la bulle et la couverture par les analystes et une forte corrélation positive entre l'indice de recherche Google et le volume de nouveaux comptes. Cette étude est essentiellement liée aux problèmes de prévision.

---

31 Heston, S.L. and N.R. Sinha (2014), 'News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns', Working Paper.

32 Nyman, R., D. Gregory, S. Kapadia, R. Smith and D. Tuckett (2014a), 'Exploiting Big Data for Systemic Risk Assessment: News and Narratives in Financial Systems', Working Paper, ECB Workshop on using Big Data for forecasting and statistics, 07-08/04/2014, Frankfurt.

33 Andrade, S. C., Bian, J. & Burch, T. R. (2009). "Does information dissemination mitigate bubbles? The role of analyst coverage in China". University of Miami Working Paper.

5. UTILISATION DES BIG DATA FISCALES ET BUDGÉTAIRES POUR LE SUIVI ET LA PRÉVISION MACROÉCONOMIQUES

Traditionnellement, les données fiscales et budgétaires pour l'analyse des politiques sont dérivées de rapports officiels qui, selon les pays, sont publiés tous les mois, tous les trimestres ou tous les ans, souvent avec un décalage important. Cependant, les innovations en matière de numérisation des systèmes de paiement et de comptabilité des administrations publiques impliquent que des données fiscales et budgétaires quotidiennes en temps réel existent dans de nombreux pays. Ces données contiennent des informations précieuses, mais sous-utilisées et sous-exploitées.

Les utilisations possibles incluent la surveillance budgétaire en temps réel, qui permet de réagir beaucoup plus rapidement aux signes de stress budgétaire émergents, et la prévision de l'activité économique, particulièrement utile dans les pays où les statistiques du PIB à haute fréquence ne sont pas disponibles.

5.1. SURVEILLANCE ET GESTION BUDGÉTAIRE

Les données budgétaires quotidiennes peuvent améliorer la surveillance et la gestion budgétaires. Les avantages de l'utilisation des données budgétaires quotidiennes par opposition aux données à fréquence plus basse pour le suivi des principaux agrégats de revenus et de dépenses dans le cadre de la surveillance budgétaire comprennent, entre autres, la rapidité des données :

- Surveillance et prévision des recettes fiscales : la disponibilité de données budgétaires quotidiennes améliore considérablement la pertinence et l'immédiateté de l'évolution des recettes fiscales et de l'analyse des prévisions de fin d'année.
- Surveillance des dépenses publiques : les données budgétaires quotidiennes apportent également une valeur ajoutée pour la surveillance des changements intervenus dans les habitudes de dépenses des administrations publiques lors de la consolidation budgétaire. Elles permettent de détecter plus rapidement les modifications de la taille et de la composition des dépenses publiques que les statistiques budgétaires conventionnelles.
- Surveillance des soldes de trésorerie du gouvernement : Le suivi des soldes de trésorerie quotidiens agrégés des administrations publiques peut donner des indications sur le niveau de liquidité disponible des administrations publiques, leur capacité à s'acquitter de leurs obligations de manière permanente et, de manière plus générale, la vulnérabilité fiscale potentielle de chaque économie.
- Amélioration de la planification de la trésorerie : L'accès à des données à haute fréquence sur les soldes de trésorerie historiques aide de deux manières. Premièrement, les données quotidiennes peuvent être utilisées pour évaluer la volatilité réelle des soldes de trésorerie, ce qui permet de mesurer le degré de sophistication de la gestion de la trésorerie, alors que les données mensuelles peuvent potentiellement masquer une grande partie de la volatilité. Deuxièmement, les données journalières historiques sur les soldes de trésorerie constituent un apport important aux efforts visant à élaborer des prévisions de trésorerie précises. Plus les prévisions sont précises, moins les fonds nécessaires pour garantir que le gouvernement puisse respecter ses engagements en cours sont réduits.

5.2. SURVEILLANCE ET PRÉVISION DE L'ACTIVITÉ ÉCONOMIQUE RÉELLE

Les données quotidiennes sur les revenus fiscaux peuvent renforcer les efforts déployés pour anticiper l'activité économique. Selon Banbura[34] et al. (2013), la prévision immédiate (ou mise à jour en temps réel) peut être définie comme la prédiction de la production (PIB) dans le présent, le proche avenir ou le passé récent.

5.2.1. Utilisation des recettes fiscales pour la prévision immédiate

L'hypothèse principale est que les modifications de l'assiette fiscale des principaux impôts et taxes reflètent globalement l'activité économique et entraînent des modifications des recettes fiscales, qui peuvent à leur tour être observées. De toute évidence, les modifications des recettes fiscales peuvent également refléter des modifications de la politique fiscale, qu'il faudra peut-être corriger si elles ont des incidences importantes sur les recettes dans le contexte de la prévision immédiate.

La TVA et les impôts sur le revenu et les plus-values sont particulièrement bien adaptés à de tels exercices car elles sont souvent déclarées à une fréquence plus élevée, ce qui implique un léger décalage entre les modifications des tendances en matière de recettes fiscales et celles de l'assiette fiscale. On peut également s'attendre à ce qu'ils reflètent la consommation privée (pour la TVA) et l'activité économique au sens large (impôts sur le revenu et les plus-values). Cela est donc particulièrement utile dans les pays où les données budgétaires quotidiennes sont disponibles mais où les statistiques des comptes nationaux, trimestrielles sur le PIB, sont soit indisponibles, peu fiables, ou considérablement retardées, et d'autres indicateurs mensuels de l'activité économique (tels que la production industrielle) ne sont également pas fournis par les autorités.

5.2.2. Autres applications pour des travaux analytiques macro-budgétaires

La littérature économique a essayé d'examiner différentes questions de recherche liées à la politique budgétaire en utilisant des données budgétaires quotidiennes.

Etude des effets des chocs budgétaires dans un monde globalisé à travers les données massives, Bureau National de l'Economie et de la Recherche, Cambridge

Auerbach et Gorodnichenko[35] (2015) ont construit deux séries quotidiennes de dépenses publiques afin d'analyser leurs effets sur les taux de change. Une des séries budgétaires quotidiennes fait référence aux paiements aux entrepreneurs de la défense rapportés dans les données budgétaires quotidiennes des États-Unis. L'autre série rassemble le volume annoncé des contrats passés quotidiennement par le département américain de la Défense. Les auteurs ont montré que les annonces concernant les dépenses futures entraînent une appréciation significative et en temps réel du dollar américain. Ce résultat contrasté par rapport à la littérature précédente est dû à l'utilisation de données journalières, ce qui permet une précision beaucoup plus fine du calendrier des chocs budgétaires et des réponses des autres variables économiques.

Etude de l'effet des dépenses publiques sur l'investissement privé, FMI

Hebous et Zimmermann[36] (2016) étudient les effets des achats fédéraux américains sur les investissements des entreprises et utilisent des données budgétaires quotidiennes issues des contrats d'achat fédéraux américains,

---

[34] Banbura, M., Domenico G., Michele M. & Lucrezia R. (2013), "Now-Casting and the Real-Time Data Flow," In Handbook of Economic Forecasting, edited by Graham Elliott, Clive Granger, and Allan Timmermann (Amste rdam: Elsevier).

[35] Auerbach, Alan J., and Yuriy Gorodnichenko, 2015, "Effects of Fiscal Shocks in a Globalized World." NBER Working Paper 21100, National Bureau of Economic Research, (Cambridge, MA).

[36] Hebous, Shafik, and Tom Zimmermann (2016), "Can Government Demand Stimulate Private Investment? Evidence from U.S. Federal Procurement," IMF Working Paper 16/60, (Washington: International Monetary Fund).

associées à des informations clés relatives aux entreprises financières. Plusieurs restrictions ont été incluses pour empêcher les entreprises de ne pas anticiper le choc de la demande publique. Les auteurs ont ainsi constaté que si les dépenses fédérales américaines augmentent, l'investissement en capital des entreprises augmente également.

Les effets sont plus importants pour les entreprises confrontées à des contraintes de financement, alors qu'ils sont proches de zéro pour les entreprises non contraintes. Conformément au modèle de l'accélérateur financier, leurs conclusions indiquent que l'effet des achats du gouvernement agit en facilitant l'accès des entreprises à des emprunts extérieurs. En outre, une analyse au niveau du secteur suggère que l'augmentation de l'investissement au niveau de l'entreprise se traduit par un effet à l'échelle du secteur sans éviction des investissements en capital des autres entreprises du même secteur.

Analyse des ventes des actions des investisseurs durant la crise financière mondiale, Bureau National de l'Economie et de la Recherche, Cambridge

Hoopes[37] et al. (2016) étudient l'hétérogénéité de la propension des investisseurs à vendre des actions pendant la crise financière mondiale en utilisant un ensemble unique de données quotidiennes sur les ventes d'actions et de parts de fonds communs de placement dans la population d'investisseurs individuels imposables aux États-Unis. Les données sont extraites de l'univers des déclarations fiscales (anonymisées) déposées auprès de l'Internal Revenue Service, ce qui leur permet de faire correspondre les ventes d'actifs déclarées à des fins d'imposition des gains en capital à des informations démographiques sur chaque contribuable. Bien que les auteurs n'observent pas l'achat d'actifs dans ces registres fiscaux, ils présentent des preuves indirectes à partir des encaissements de dividendes et d'un ensemble de données de compte de courtage supplémentaire, ce qui suggère que les individus avec des ventes brutes élevées sont aussi, dans une large mesure, également des vendeurs nets d'actions.

5.3. CAS DU MAROC : UTILISATION DES BIG DATA POUR LA SURVEILLANCE FISCALE AU SEIN DE LA DGI

En tant que créateur et gestionnaire des données fiscales au sein du Ministère de l'Economie, des Finances et de la Réforme de l'Administration, la Direction Générale des Impôt (DGI) a réalisé de grandes avancées en matière de digitalisation ces dernières années (lancement et généralisation des télédéclarations et télépaiements, échange de données avec les partenaires…).

Malgré que les données fiscales soient dématérialisées, leur exploitation par l'administration fiscale demeure laborieuse : il faut accéder à plusieurs applications et effectuer de nombreuses recherches. Pour remédier à ce problème, La DGI a décidé de se doter, dorénavant, d'un nouveau système basé sur le Big Data et l'Intelligence Artificielle.

Pour se faire, la Direction a lancé un appel d'offres pour la réalisation d'une étude relative à la définition de l'architecture de l'environnement analytique Big Data, et la mise en œuvre d'un système de recoupement et d'analyse des données (SRAD).

Ce système est destiné à jouer un rôle d'intégration et de valorisation des données issues des différentes applications de la DGI et également des données fournies par les partenaires, en vue d'améliorer l'efficacité et la pertinence des contrôles.

[37] Hoopes, Jeffrey, Patrick Langetieg, Stefan Nagel, Daniel Reck, Joel Slemrod, and Bryan Stuart, 2016, "Who Sold During the Crash of 2008–9? Evidence from Tax-Return Data on Daily Sales of Stock." NBER Working Paper 22209, National Bureau of Economic Research, (Cambridge, MA).

Améliorer l'efficacité et la pertinence des contrôles

L'architecture cible devra couvrir l'ingestion et le traitement massif des données, la gestion de la qualité des données, et leur rattachement à un référentiel consolidé des contribuables. Elle devra également disposer des capacités de détection de fraude à travers des règles configurables, et aussi à travers des algorithmes de Machine Learning. Elle devra également permettre la visualisation des résultats au travers de plusieurs vues configurables.

Un mode vision 360° des contribuables

Ce système va générer automatiquement des fiches profil permettant d'avoir une vue unique actualisée de l'ensemble des données agrégées et détaillées d'un même contribuable. Par ailleurs, le système appliquera des règles et des algorithmes du « Machine Learning », de « scoring » et de sélection des contribuables pour les différents programmes de contrôle fiscal, de vérification générale, de vérification ponctuelle, de contrôle sur pièce, de droit de constatation ou d'examen de l'ensemble de la situation fiscale.

6. UTILISATION DES BIG DATA DANS LE CONTEXTE DE LA PANDÉMIE DU CORONAVIRUS

La crise de la pandémie du Coronavirus que traverse le monde aujourd'hui est exceptionnelle, de par sa portée, la vitesse de sa propagation, ses impacts et répercussions dévastateurs sur tous les plans, mais de par aussi, les mesures et moyens mobilisés pour sauvegarder voire atténuer ses effets sur les hommes et les entreprises.

La crise COVID-19, étant donné la multidimensionnalité et l'intersectorialité de l'impact qu'elle a engendrés, a permis d'émerger l'utilisation des Big data comme outil avancé, efficace et réactif pour alimenter et éclairer la prise de décision dans un contexte de grande incertitude. L'Intelligence Artificielle (IA) et le Big Data sont testés au niveau des différents pays du monde comme jamais auparavant. Cette crise a également souligné la nécessité de poursuivre la recherche, l'exploration et le développement des innovations adaptées pour soutenir le rôle des organismes de collecte, de traitement et des rassembleurs et catalyseurs pour un écosystème de données solide.

Les utilisations des Big Data sont ainsi multiples. Différents pays utilisent la collecte massive de données et les smartphones pour des « systèmes de suivi et de traçage » des personnes pour un meilleur suivi de la propagation du virus. Par ailleurs, des modèles quantitatifs de prévision de la propagation du virus ainsi que des analyses d'impact basées sur les données massives pour évaluer les effets de cette pandémie sur l'activité économique se multiplient.

6.1. SUIVI ET LUTTE CONTRE LA PROPAGATION DU CORONAVIRUS

L'un des avantages que nous avons aujourd'hui dans la lutte contre le coronavirus, qui n'était pas aussi sophistiqué lors des épidémies antérieures, dont la plus récente celle du SRAS de 2003, est cette capacité à collecter, stocker, manipuler et traiter une quantité importante d'information en lien avec le développement des Big Data et la prolifération des nouvelles technologies de l'information de la communication.

En outre, les pays asiatiques, notamment la Chine, le Taïwan et la Corée du Sud ont exploité les Big Data, l'apprentissage automatique et d'autres outils numériques afin de suivre et contenir la propagation de l'épidémie.

Les expérimentations effectuées au niveau de ces pays ont constitué des modèles à suivre qui se sont répandus dans les quatre coins du globe.

D'autres pays dans la lutte contre la propagation du virus, dont l'Italie qui a été sévèrement impactée, utilisent la technologie numérique pour élaborer des prévisions en temps réel et fournir aux professionnels de la santé et aux décideurs gouvernementaux des informations qu'ils peuvent utiliser pour adapter les mesures prises à l'évolution de la situation sanitaire et également afin de prédire l'impact du coronavirus sur l'économie dans sa globalité et pour certains secteurs névralgiques.

L'infrastructure de surveillance de la Chine est utilisée pour suivre les personnes exposées

Les systèmes de surveillance développés par la Chine se sont avérés très utiles dans la réponse du pays à COVID-19. Des scanners thermiques ont été installés dans les gares pour détecter des températures corporelles élevées, signe potentiel d'infection. En effet, si une température élevée était détectée, la personne étant alors mise en quarantaine par les responsables de la santé pour subir un test de coronavirus. Si le test du coronavirus est déclaré positif, les autorités alerteraient tous les passagers contacts susceptibles d'avoir été exposés au virus afin qu'ils puissent se mettre en quarantaine. Cette notification a été activée en raison des règles de transport du pays qui exigent l'utilisation de la carte d'identité pour chaque passager voyageant dans les transports publics.

La Chine possède des millions de caméras de surveillance et de sécurité qui sont utilisées pour suivre les mouvements des citoyens en plus de repérer les crimes. Ce dispositif a permis aux autorités d'inspecter le respect des personnes des règles sanitaires et de mise en quarantaine. En effet, si une personne mise en quarantaine, a été repérée par les caméras de surveillance à l'extérieur de son domicile les autorités sont alertées. Les données des téléphones portables sont également utilisées pour suivre la mobilité des citoyens.

Le gouvernement chinois a également déployé une application « Close Contact Detector » qui a pour objectif d'alerter les utilisateurs en cas de contact avec une personne contaminée au cours des 14 derniers jours. L'information collectée fournie par les rapports de vérification des voyages produits par les fournisseurs de télécommunications pourraient ainsi répertorier tous les sites de déplacement d'un utilisateur au cours des 14 derniers jours pour identifier le risque d'exposition au virus et déterminer si une mise en quarantaine est recommandée. L'intégration de l'ensemble des données collectées à travers ses systèmes de surveillance chinois a permis au pays de développer des solutions adéquates pour lutter contre la propagation du coronavirus.

Plateformes pour suivre et analyser la propagation du virus et de l'épidémie

Dans le cadre du suivi de la progression de la pandémie plusieurs organismes nationaux et internationaux et instituts de recherches ont développé des plateformes de collecte et de diffusion de l'information relative aux différents indicateurs d'évolution de la pandémie ainsi que les différentes mesures prises par les gouvernements pour atténuer les effets socioéconomiques de la crise sur le plan sanitaire, social, fiscal, monétaires, … Ces tableaux de bord informationnels, se sont avérés très utiles pour les décideurs gouvernementaux, les professionnels de la santé et le grand public pour mesurer le degré de propagation de la contagion et pour la prise de décision. Ils sont également importants pour servir les modèles de prévision et de mesure d'impact de la crise.

Ainsi les plateformes de l'Organisation Mondiale de la Santé, celle de Google ou celle de de l'université Johns-Hopkins, et bien d'autres, fournissent des statistiques en temps réel et rassemblent des données du monde entier relatives au nombre de cas de contamination au coronavirus confirmés, de décès, de rémissions, de tests effectués, … tout en précisant l'emplacement des personnes contaminées. Cet ensemble de données complet

peut ensuite être utilisé pour créer des modèles de prédiction et d'identification des points de propagation de la maladie afin d'aider les systèmes de santé à se préparer à une flambée de cas.

L'analyse des épidémies prend, ainsi, toutes les données disponibles (y compris le nombre de cas confirmés, les décès, le suivi des contacts des personnes infectées, les densités de population, les cartes, le flux des voyageurs, …), puis les traite par apprentissage automatique pour créer des modèles spécifiques de la maladie. Ces modèles représentent les meilleures prévisions concernant les taux d'infection de pointe et les résultats.

Analyse des données volumineuses : Cas de Taiwan

Eu égard de sa proximité de la Chine et des fortes relations économiques avec ce pays, notamment en matière de flux commerciaux de déplacement de personnes et de main d'œuvre, l'île de Taiwan a été exposée à un grand risque de propagation du virus Covid-19. Toutefois, grâce à sa réactivité pour la gestion du risque, de l'utilisation des nouvelles technologies et la création d'un solide plan de lutte contre la pandémie après celui de SRAS de 2003, le pays a pu minimiser l'impact du coronavirus au niveau national.

Une partie de la stratégie du pays s'est basée sur les données nationales d'assurance maladie, de la base de données sur l'immigration et les douanes. Avec le croisement de ces différents systèmes, les autorités ont développé un système d'alertes en temps réel concernant les personnes susceptibles d'être infectées en fonction des symptômes et des antécédents de voyage.

L'analyse et la notification en ligne des symptômes de voyage et de santé, couplée à une ligne d'assistance téléphonique gratuite pour que les citoyens signalent les symptômes suspects, ont permis de classer les risques d'infection des voyageurs. Les autorités ont pris des mesures immédiates dès que l'OMS a alerté sur une pneumonie de cause inconnue en Chine le 31 décembre 2019. Il s'agissait du premier cas de coronavirus signalé, et la réponse rapide et l'utilisation de la technologie de Taiwan sont les raisons probables du faible taux d'infection relativement à d'autres pays.

Mesure de la population présente en temps de confinement et statistiques expérimentales en utilisant les données de téléphonie mobile, INSEE.

Savoir comment se répartit effectivement la population sur le territoire est essentiel pour organiser la réponse sanitaire et sociale face à l'épidémie du coronavirus. L'INSEE a ainsi diffusé le 8 avril les premiers résultats de population présente en métropole. Il a utilisé des informations issues des données de téléphonie mobile. Ces données sont des statistiques de comptage, agrégées territorialement, collectées au niveau des antennes relais. Ce ne sont ni des données GPS de localisation des téléphones portables, ni des données issues d'application téléchargées. Elles ne permettent pas de suivre le déplacement des personnes, mais plutôt d'effectuer des comptages par zones à différentes dates, ce qui aide à comprendre et à analyser la propagation de la pandémie.

6.2. L'ANALYSE DES EFFETS DE LA PANDÉMIE SUR L'ACTIVITÉ ÉCONOMIQUE

Utilisation des données à « haute fréquence » dans la prévision et l'analyse économique en période de crise, INSEE

L'ampleur et la rapidité de transmission du choc engendré par la pandémie du Covid-19 ont limité la pertinence et le pouvoir prédictif des indicateurs conjoncturels mobilisés usuellement pour mesurer et prévoir l'activité économique. Le suivi conjoncturel en cette période a donc privilégié l'utilisation de nouvelles sources de données,

à plus haute fréquence que mensuelle ou trimestrielle. En temps normal, ces indicateurs sont produits au moins un mois de décalage à l'exception des nouvelles données mobilisées pour le suivi de l'activité française depuis la crise de la Covid-19.

Pour les quatre principales économies de la zone euro, les États-Unis et le Royaume-Uni, ces nouvelles données expliquent une part non négligeable de la variation d'indicateurs traditionnels de production et de consommation. Ainsi, dans l'attente des résultats mensuels des enquêtes, les données à haute fréquence s'avèrent donc utiles à l'analyse et à l'estimation de l'activité. Ils apportent une information complémentaire aux enquêtes de conjoncture permettant de mieux appréhender la perte d'activité à très court terme.

La plupart des indicateurs conjoncturels d'activité économique utilisés habituellement dans les notes de conjoncture sont mensuels ou trimestriels, et disponibles seulement à la fin du mois ou du trimestre écoulé. Néanmoins, connaître l'évolution de l'activité de manière encore plus précoce peut être crucial, en particulier pendant la crise sanitaire du Covid-19 qui a occasionné des mouvements économiques soudains et de grande ampleur. L'information qualitative contenue dans les articles de la presse économique française peut être mobilisée dans ce sens. Elle permet, notamment, de calculer un indicateur mesurant la tonalité de l'opinion médiatique qui peut être utilisé pour estimer l'évolution de l'activité économique. Cet indicateur donne des indications en temps réel sur l'économie française souvent concordantes avec l'évolution du PIB mesurée ex post.

- **Utilisation des données de transactions par carte bancaire pour l'analyse du comportement des consommateurs, INSEE.**

Parmi les indicateurs à « haute fréquence » renseignant sur les évolutions de la consommation des ménages pendant la période de confinement, les données de transactions par carte bancaire constituent une source privilégiée. En effet, l'INSEE a procédé à l'exploitation de données de cartes bancaires pour analyser les comportements d'achat courant des ménages avant puis durant la période de confinement qui a débuté le 17 mars 2020. Si ces données comportent certaines particularités et limites, elles offrent une vision rapide et riche des évolutions de la consommation d'ensemble comme par secteur, le soutien que constitue la vente à distance ainsi que sur l'évolution du nombre d'achats quotidiens et du panier moyen.

Ces données illustrent de façon journalière les comportements d'achat des ménages, tant au niveau de tous les produits qu'à un niveau plus fin, et permettent dès lors d'identifier les catégories de produits présentant un profil de consommation moins atone que les autres. Si les dépenses de carburants ou d'hôtellerie-restauration n'évoluent pas, en revanche certaines dépenses de biens manufacturés (équipement du foyer, habillement-chaussure) se redressent très progressivement. La vente en ligne continue, logiquement, de progresser.

- **Consommation au temps du Covid-19 : analyse à partir des données transactionnelles au Royaume-Uni, Centre for Economic Policy Research.**

En utilisant les données de transaction d'une grande société « Fintech », les auteurs de cette étude ont utilisé les données détaillées au niveau des transactions de l'un des plus grands gestionnaires financiers personnels du Royaume-Uni, Money Dashboard (MDB). L'application est un agrégateur de comptes en temps réel qui rassemble les données de transactions financières des comptes courants, de crédit et d'épargne des utilisateurs, quel que soit le fournisseur, sur une seule plateforme. L'échantillon s'étend sur la période du 1er janvier au 26 avril 2020 et concentre plus de 34 000 utilisateurs qui ont régulièrement utilisé l'application au cours de cette période. Cela

donne un échantillon de près de 8,5 millions de transactions. La granularité des données permet de construire des mesures de dépenses hebdomadaires non seulement pour les biens non durables, durables et les services mais aussi pour un certain nombre de catégories plus détaillées telles que la vente au détail, la restauration, les voyages et les transports. Le système couvre également les résultats mensuels des revenus, les frais bancaires et les versements hypothécaires.

Les auteurs ont constaté une baisse de 40% à 50% des dépenses des ménages britanniques pendant la crise de Covid-19. La baisse est concentrée dans les services tels que la vente au détail, les restaurants et les transports. La hausse initiale des achats en ligne et des achats d'épicerie a par la suite été inversée. Les réductions de revenus sont devenues beaucoup plus fréquentes, avec une baisse médiane d'environ 30%. La part des emprunteurs confrontés à des problèmes de financement a considérablement augmenté pour les prêts garantis et non garantis. Les inégalités de consommation et de revenu ont augmenté. Les groupes les plus vulnérables économiquement connaissent la plus forte baisse en pourcentage. Les débiteurs hypothécaires et les hauts salariés de Londres enregistrent le changement de livre le plus important.

Coronavirus : la chute de la production mondiale vue de l'espace, « Les Echos ».

Pour saisir l'ampleur et la dynamique de ce qui s'avère déjà la plus grande crise industrielle du siècle, « Les Echos » ont lancé « l'indice de la reprise Kayrros-EY Consulting », qui permet le suivi hebdomadaire de l'activité de l'industrie des principales zones économiques de la planète - la Chine, les Etats-Unis, l'Union européenne et l'Inde grâce aux images des satellites.

Analyse de l'activité économique en période de confinement à travers les données de production et de consommation d'électricité, INSEE.

Les données de production et de consommation d'électricité quotidiennes françaises sont une source utile pour suivre en temps réel les évolutions de l'activité des entreprises et des ménages. La consommation d'électricité, en particulier, reflète la modification des comportements induite par la crise du coronavirus, qu'il s'agisse de la baisse de production dans des secteurs intensifs en électricité, comme les transports, ou du mode de vie transformé des ménages confinés. Le climat et la saisonnalité affectant fortement la production et la consommation d'électricité, les comparaisons temporelles sont effectuées après correction des effets des variations de température, des jours ouvrés et des mois de l'année.

Selon les données de RTE (réseau de transport d'électricité), la consommation totale pendant la période du 23 mars au 26 avril 2020 est inférieure de 14% à celle d'une période normale. De manière relativement cohérente avec l'ordre de grandeur des baisses d'activité estimées par ailleurs, la consommation des entreprises directement raccordées à RTE (pour la plupart de gros industriels) est inférieure de 24%. Par ailleurs, selon les données issues d'Enedis, pendant la période du 23 mars au 3 avril, la consommation hors résidentiel (entreprises – hors celles directement raccordées à RTE – et secteur public) est inférieure d'environ 27 % tandis que celle des ménages est supérieure d'environ 4% à la normale.

PARTIE II

ETUDES DE CAS POUR LE MAROC

Etude de cas N° 1

Opportunités du Big Data pour un suivi avancé de l'activité touristique au Maroc

Dans le cadre de sa mission d'analyse économique, de veille et de prévision, la DEPF tente d'explorer le gisement des données massives du web relatives à l'activité touristique. Cette analyse, visant à mieux orienter les actions publiques, se propose de faire un diagnostic de l'offre touristique sur le web de 10 destinations méditerranéennes (Marrakech, Agadir, Cordoue, Séville, Tunis, Le Caire, Prague, Rome, Istanbul, Athènes) et d'en relever l'appréciation multidimensionnelle faite par la clientèle. A cette fin, une quarantaine de variables (numérique, alphanumérique, géodésique…) ont été collectées[38] à partir du web couvrant les 10 destinations méditerranéennes.

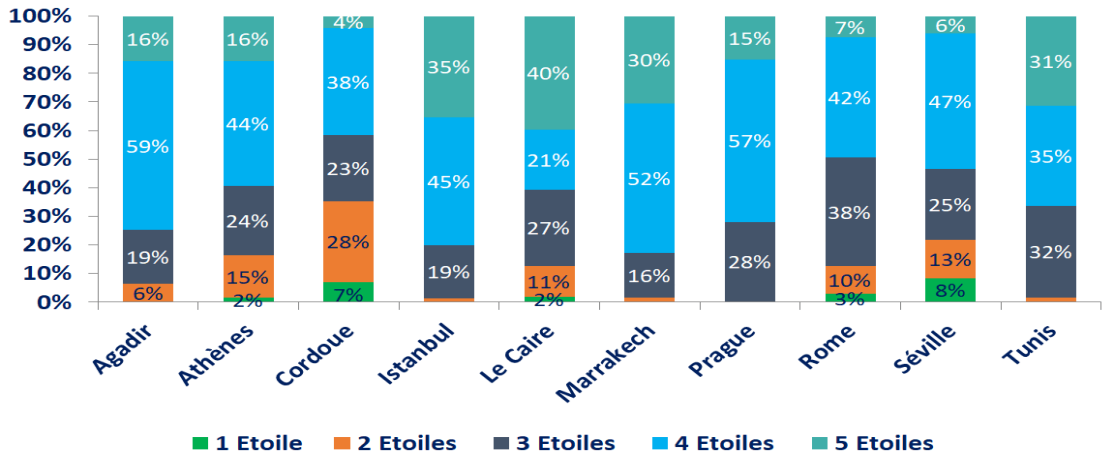
Ilyes BOUMAHDI
Chef du service des Activités Tertiaires et de l'Economie du Savoir, DEPF

[38] En utilisant l'outil Data Miner. Pour plus de détails voir : https://data-miner.io/

1. CARACTÉRISTIQUES DE L'OFFRE TOURISTIQUE SUR LE WEB DE 10 DESTINATIONS MÉDITERRANÉENNES

L'analyse de l'offre touristique web est basée sur 8000 hébergements au niveau de dix destinations méditerranéennes. L'offre de Marrakech se caractérise par la prépondérance des chambres d'hôtel (74%) au même titre qu'Istanbul (84%) et Rome (74%) au moment où celle d'Agadir est relativement équilibrée entre chambre d'hôtel (36%) et appartement (57%).

Graphe 1 : Structure de l'Offre hôtelière web de 10 destinations méditerranéennes en 2019



Source : Calcul DEPF à partir de données extraites du Web

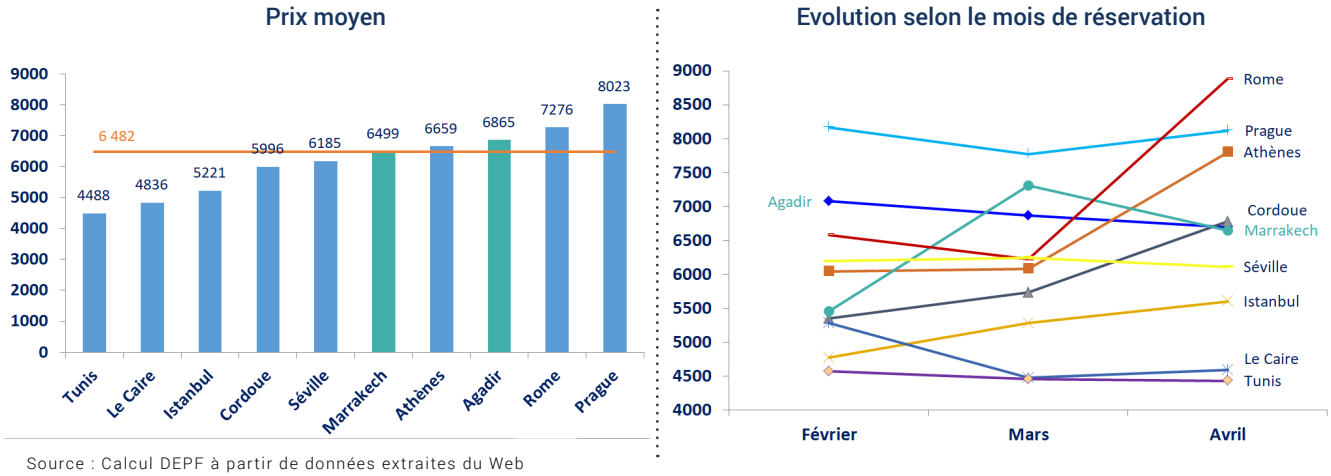
L'offre hôtelière web de Marrakech est concentrée sur le moyen-haut de gamme à hauteur de 82%, soit une structure légèrement supérieure à celle de sa capacité réelle (73% avec, respectivement, 43% en 5 étoiles et 33% en 4 étoiles). La même configuration est perceptible à Istanbul alors que Agadir donne une place, également, importante aux 3 étoiles (19%). L'offre web d'Agadir diffère de celle de sa capacité (25% en 5 étoiles, 33% en 4 étoiles et 27% pour les 3 étoiles) en faveur des 5 étoiles. Par contre, Cordoue offre une palette plus diversifiée avec un ancrage sur la gamme moyenne-inférieure (respectivement 38% en 4 étoiles, 23% en 3 étoiles et 28% en 2 étoiles).

En termes de commodité, l'offre touristique du Caire demeure la plus dispersée avec une distance moyenne au centre de 8,7 km contre respectivement 624 m et 595 m pour Séville et Cordoue. L'offre de Marrakech présente une dispersion plus équilibrée autour du centre avec une distance moyenne de 2,3 km mettant moins de pression sur les modes de mobilité. D'ailleurs, cette proximité du centre a été relevée parmi les atouts de Marrakech et d'Agadir dans les commentaires des visiteurs.

Par ailleurs, le prix moyen des hébergements dans toutes les destinations est de 6482 dirhams[39]. Les prix les plus faibles sont enregistrés au niveau de Tunis et du Caire alors que Prague se distingue par un prix élevé. Rome a pratiqué, pour la période de l'analyse, la baisse promotionnelle des prix la plus importante avec une réduction moyenne de 23% pour certains hébergements contre 15% seulement pour Prague. Quant à Marrakech (6499 dirhams), elle a pratiqué des prix analogues à la moyenne de l'échantillon alors que Agadir est sur un prix relativement plus élevé (6865 dirhams).

[39] Prix moyen pour une réservation de deux adultes effectuée en février, mars et avril pour six nuitées du 1 au 8 août 2019.

Graphe 2 : Prix de l'offre touristique web de 10 destinations méditerranéennes en 2019



Source : Calcul DEPF à partir de données extraites du Web

Cependant, des différences de prix pourraient être dues à la période de séjour qui coïncide avec le mois d'août qui s'avère un mois de haute saison pour certaines régions comparativement à d'autres. Ceci dit, l'indice[40] des prix web serait un indicateur potentiel pour le suivi de l'activité conjoncturelle des destinations nationales. En effet, la variabilité des prix selon la date de réservation suggérerait un lien, entre autres[41], avec l'attractivité de la destination. Ainsi, Rome a réalisé le renchérissement le plus important, pour une réservation établie en avril, alors que les prix ont fléchi en mars. Marrakech a connu une évolution inverse pour converger vers les prix appliqués par Cordoue.

Compte tenu de la structure de l'offre hôtelière par ville, la charge fiscale est plus élevée à Rome (591 dirhams[42]) et à Marrakech (494 dirhams) au moment où elle est plus faible à Istanbul (15 dirhams). La différence en termes de charge fiscale et donc de compétitivité prix des destinations est également liée au système fiscal national et local de chaque pays[43]. Ainsi, la Turquie ne pratique pas de taxes de séjour[44] permettant de maintenir ses prix à un niveau très compétitif avec une attractivité bénéfique pour le secteur touristique malgré le débat récurrent sur l'opportunité des taxes de séjour pour réguler la pression touristique sur les territoires[45].

2. APPRÉCIATION DE L'OFFRE TOURISTIQUE SUR LE WEB DE 10 DESTINATIONS MÉDITERRANÉENNES

Sur la base de l'appréciation de près de 22 millions de touristes ayant vécus des expériences réelles en termes d'hébergement dans les 10 destinations méditerranéennes, il s'avère que le schisme nord-sud est consacré avec des notes supérieures à la moyenne de l'échantillon (8,7/10) pour les destinations de l'Union Européenne. Marrakech est la destination la mieux cotée du versant méridional (8,6) au moment où Agadir enregistre l'appréciation la plus faible (7,9). Cette appréciation générale discriminante est cependant plus homogène pour toutes les destinations du point de vue de la situation géographique des établissements (9,6).

---

40 L'extraction des prix selon la date de réservation suivant trois mois (février, mars et avril) n'est pas suffisante pour la construction d'un tel indice. Mais, un tel travail pourrait être envisagé à moyen-long terme, le temps de constituer une série exploitable.

41 Cette évolution pourrait, également, être due à la vente précoce des offres les moins chères durant les deux premiers mois ou à la programmation tardive d'un grand événement.

42 Relativement à un séjour de six nuitées pour deux adultes.

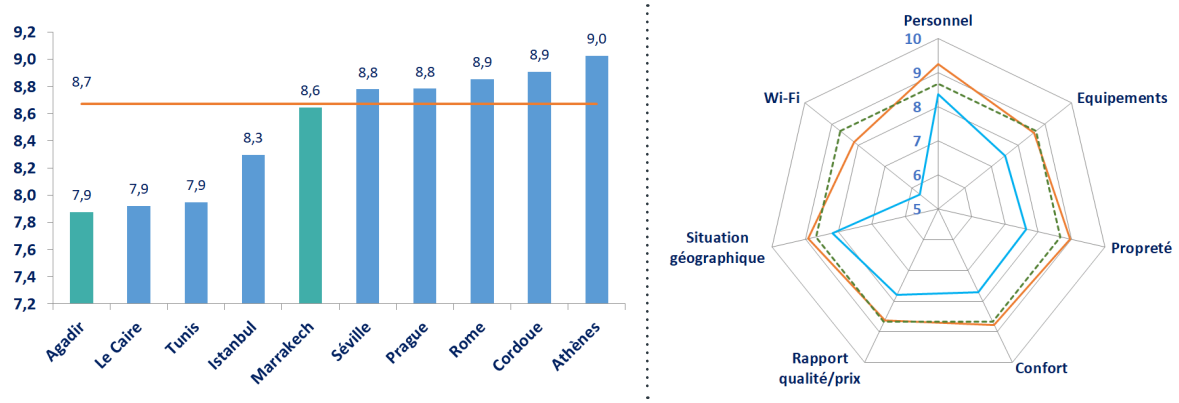
43 Cette taxe varie dans l'UE entre 0,1€ (Bulgarie) et 7,5€ (Belgique)/personne/nuit avec une fourchette moyenne entre 0,4 et 2,5 €. https://ec.europa.eu/growth/sectors/tourism/business-portal/financing-your-business/tourism-related-taxes-across-eu_en.

44 Les données ont été recueillies avant août 2019, soit avant l'entrée en vigueur d'une taxe annoncée en juillet 2019, contraignant les hôtels et les agences de voyage à verser jusqu'à 0,75% de leurs recettes à la nouvelle agence turque de promotion et de développement du tourisme.

45 Voir une analyse propre au cas d'Istanbul sur Cetin, Gurel. (2014), « Sustaining tourism development through city tax: The case of Istanbul. » e-Review of Tourism Research. 11. 26-41.

Les offres non classifiées sont mieux appréciées (8,8 pour Marrakech et 8,1 pour Agadir) relativement à celles classifiées. Ainsi, dans la catégorie des cinq étoiles, Agadir a une appréciation (7,5) au-dessous de sa moyenne (7,9) et celle de la même catégorie pour toutes les destinations (8,3). Par contre, si Marrakech arrive à avoir la même appréciation que la moyenne méditerranéenne pour les cinq étoiles (8,3), les autres catégories greffent l'appréciation générale de la destination.

Graphe 3 : Appréciation de l'offre web de 10 destinations méditerranéennes en 2019



Source : Calcul DEPF à partir de données extraites du Web

Le recul d'Agadir est en particulier lié à sa défaillance au niveau des équipements (7,5/10), de la propreté (7,6), du confort (7,7) et du rapport qualité/prix (7,8). Deux alternatives se présenteraient pour la destination, dont la première serait de réduire le prix du séjour pour créer moins d'attente de la part des touristes et donc une satisfaction plus améliorée compte tenu du prix déboursé. Le prix qui paraît adapté au profil d'Agadir[46] serait de -16% de celui relevé durant la période de réservation. L'estimation de ce prix alternatif s'est basée sur les caractéristiques des hébergements que nous avons pu recueillir du web[47] alors que d'autres critères, dont nous ne disposons pas, sont aussi importants dans la formation des prix (taux d'occupation, appartenance à une enseigne internationale, segment, …). La deuxième alternative serait de lancer une vaste mise à niveau d'équipement, notamment, en connectivité et en literie pour être à la hauteur des attentes des touristes. La destination est, par contre, mieux appréciée pour l'attention de son personnel (8,4).

3. APPRÉCIATION DE L'OFFRE TOURISTIQUE SUR LE WEB DE MARRAKECH ET AGADIR

Afin de faire une analyse plus fine de l'appréciation des destinations marocaines, une extraction a été faite des commentaires textuels des visiteurs ayant séjourné effectivement dans des hébergements à Marrakech et à Agadir. L'extraction a concerné aléatoirement 5% des hébergements proposés en ligne de chaque destination, soit 70 hébergements (10 pour Agadir et 60 pour Marrakech). Pour chaque hébergement, l'extraction a porté sur les dix commentaires les plus récents pour avoir l'appréciation la plus actualisée des destinations marocaines, soit près de 600 commentaires textuels de visiteurs de 58 nationalités en une trentaine de langues. Cette structure par pays est légèrement différente de celle relevée par les données conventionnelles[48], compte tenu de la différence d'utilisation des canaux de réservation par pays.

---

[46] Cette estimation est issue de la modélisation du prix pratiqué dans 8000 établissements en fonction de leurs notes d'appréciation, leurs types (appartement, chambre, studio, maison de vacance, suite), leurs distances au centre et leurs classifications hôtelières (R^2=0,627).

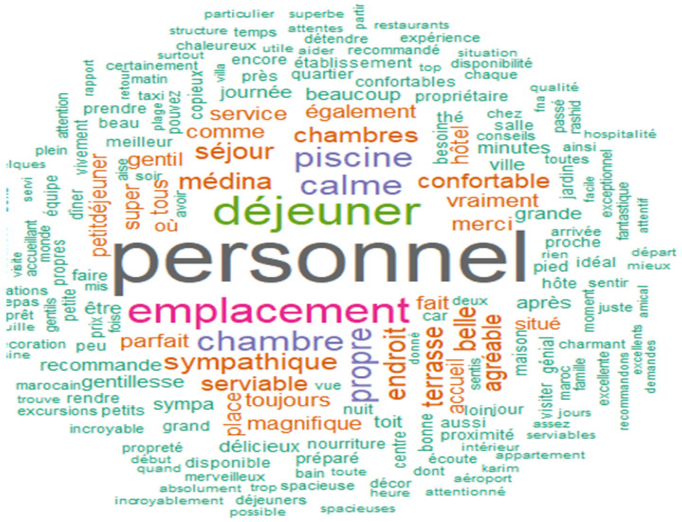
[47] Appréciation, types de l'hébergement (appartement, chambre, studio, maison de vacance, suite), distances au centre, classifications hôtelières.

[48] La France, le Royaume Uni, l'Allemagne et l'Espagne sont les marchés émetteurs les plus importants pour le Maroc concentrant 46% des arrivées (respectivement 38,3% et 40,1% pour Agadir et Marrakech) et 56% des nuitées (respectivement 52,3% et 49,5%).

L'analyse[49] des commentaires a été faite après correction des erreurs de rédaction, souvent réalisée en langage parlé non formel, et traduction en français[50]. Cette particularité de rédaction des commentaires sur le web a été favorable à l'analyse textuelle des sentiments en se basant sur les mots et non pas sur les phrases. Ainsi, il s'avère que les mots qui reviennent le plus souvent sont « personnel », « petit déjeuner », « emplacement » et « chambre ».

Ainsi, le mot « personnel », le plus fréquent avec 189 occurrences, est souvent associé à des mots à connotation positive (serviable et sympathique avec des coefficients de corrélation de 29%[51]). Les mots « petit déjeuner » (95) et « emplacement » (80) sont associés, également, à des sentiments positifs (respectivement copieux avec un coefficient de corrélation de 38%, et place et idéal avec un coefficient de corrélation de 21%). Le sentiment négatif le plus perceptible est lié à la petite taille de la piscine (33% de corrélation).

Graphe 4 : Nuage des mots des commentaires de l'offre web de Marrakech et Agadir en 2019



Source : Calcul DEPF à partir de données extraites du Web

L'analyse des sentiments des commentaires dépend des librairies du logiciel utilisées dans la préparation des données textuelles[52] et des routines[53] utilisées dans leur analyse qui, souvent, ont été développées pour les textes en anglais. Ainsi, l'analyse des sentiments des commentaires révèle la prépondérance des mots à connotation positive à hauteur de 90% avec une différenciation entre Marrakech (90%) et Agadir (85%) et ce, conformément à la notation différenciée des deux destinations (respectivement 8,6 et 7,9).

Aussi, la promotion des deux villes devrait-elle être différenciée pour faire valoir les avantages révélés dans les commentaires pour chaque ville, à savoir pour Marrakech, l'emplacement dans ou près de la Médina et de la Place Jamaa El Fna, et pour Agadir, la proximité, l'équipement et la disponibilité des espaces réservés dans la plage et ce, en plus des qualités communes aux deux destinations qu'il conviendrait d'entretenir à savoir le sens de l'accueil du personnel, l'authenticité du séjour, la qualité gastronomique, les dimensions des chambres, …

---

49 L'analyse a été faite par des packages R dont tidytext, dplyr et tm.

50 Le traitement manuel des commentaires a rendu difficile l'automatisation de la traduction. Cependant, il existe des solutions intéressantes à explorer dans l'avenir tel que l'API translate Cloud fourni par Google dans sa version beta. https://cloud.google.com/translate/docs/intro-to-v3.

51 Corrélation entre les deux vecteurs d'occurrences des deux mots selon leurs apparitions dans les 600 commentaires. Pour plus de détails voir «Introduction to the tm Package : Text Mining in R » , Ingo Feinerer, December 21, 2018.

52 Elimination des mots courants tels que « a, abord, absolument, afin, … ». La librairie utilisée est la version française de stopwords-iso augmentée par d'autres mots pour les besoins spécifiques à cette analyse (petit qui revenait souvent avec déjeuner). Pour plus de détails voir https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.json.

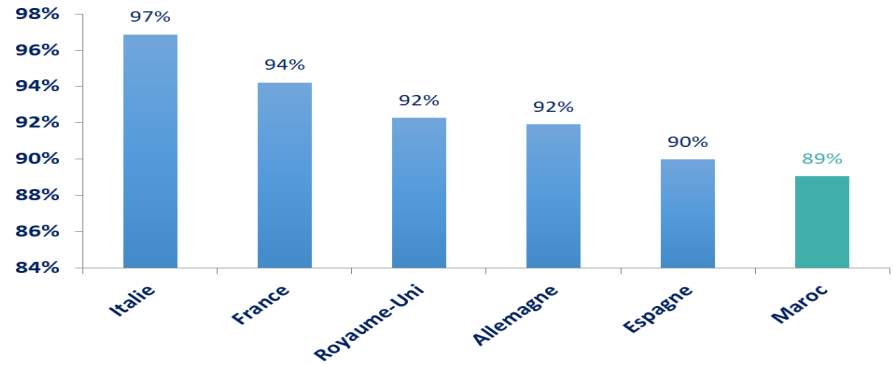
53 Pour l'analyse des sentiments le lexique francophone FEEL (French Expanded Emotion Lexicon), englobant 14000 mots exprimant des émotions et des sentiments, a été utilisé. Pour plus de détails, voir Amine Abdaoui, Jérôme Azé, Sandra Bringay et Pascal Poncelet. « FEEL: French Expanded Emotion Lexicon ». Language Resources and Evaluation, LRE 2016, pp 1-23. http://www.lirmm.fr/~abdaoui/FEEL.

Graphe 5 : Nuage des mots des commentaires de l'offre hôtelière visible sur le web de Marrakech et Agadir selon leurs polarités en 2019

Source : Calcul DEPF à partir de données extraites du Web

Cette différenciation est, également, perceptible par pays. Ainsi, il semble que les italiens et les français ont une appréciation plus positive que les autres nationalités. Le taux de satisfaction le plus bas étant celui des nationaux tirant ainsi la moyenne de notation de la destination nationale vers le bas relativement à celles concurrentes. En effet, il s'avère que le taux de corrélation entre le ratio de mots avec des sentiments positifs et la notation des établissements est fortement corrélé[54].

Graphe 6 : Ratio des mots positifs des commentaires de l'offre touristique web de Marrakech et Agadir relativement au total des mots positifs et négatifs en 2019



Source : Calcul DEPF à partir de données extraites du Web

4. ENSEIGNEMENTS ET RECOMMANDATIONS

Ainsi, il s'avère que l'offre web des destinations méditerranéennes est très diversifiée sortant de la catégorisation ordinaire des hôtels pour couvrir d'autres types d'hébergement (appartements, gîtes, maisons d'hôtes, …). Ces derniers contribuent à améliorer l'offre touristique authentique des destinations[55]. En effet, les offres non classifiées sont mieux appréciées à Marrakech et à Agadir relativement à celles classifiées, qui greffent l'appréciation générale de la destination, relevant ainsi le rôle important que joue les petites structures d'accueil

[54] 0,63 pour les 17 marchés émetteurs les plus importants.

[55] La relation entre la taille de l'entreprise et l'expérience client est relativement compliquée. Une augmentation de la taille de l'entreprise améliorerait l'expérience fonctionnelle et émotionnelle du client mais détériorerait l'expérience authentique. Ye, Shun & Xiao, Honggen & Zhou, Lingqiang. (2018). « Small accommodation business growth in rural areas: Effects on guest experience and financial performance ». International Journal of Hospitality Management.

dans l'attractivité des destinations et qu'il conviendrait de soutenir en termes de financement et de formation. Aussi, toute stratégie de soutien ou de promotion devrait tenir compte de ces acteurs touristiques qui ne sont pas organisés en corporation ce qui les met au ban des négociations et des outils mis en place par les instances publiques et les collectivités locales pour soutenir le secteur touristique.

Ces petites structures manquent de capacité concurrentielle et organisationnelle pour faire face aux grands opérateurs, notamment, en période de pénurie (annulation, taux d'occupation, saisonnalité, …). Il serait, donc, judicieux de soutenir ces établissements pour améliorer leurs offres en mettant à contribution les attentes révélées de la clientèle. Il s'agit, notamment, de la mise à niveau d'équipement en connectivité et en literie. Des fonds de soutien nationaux ou régionaux pourraient être mis à contribution avec une offre adaptée aux petites structures et un accompagnement dans le montage de leur dossier de candidature[56]. Il s'agit, également, de l'authenticité du séjour, de l'amabilité du personnel et de la recherche de la durabilité de l'activité.

Ainsi, le Maroc pourrait faire valoir, face à la concurrence qui a opté pour le tourisme de masse, son offre de qualité de moyenne à haute gamme basée sur le culturel, l'authenticité, la diversité et la durabilité. En effet, relativement à ce dernier point, certains touristes sont moins exigeants[57] par rapport à certains aspects d'hébergement en faveur de plus de responsabilité écologique des entreprises hôtelières. Aussi, la labélisation verte des petites structures pourrait relativement contrer certaines défaillances, particulièrement, en termes de confort, de propreté et d'équipements.

L'analyse a concerné la partie visible des données des plateformes web, or un flux volumineux d'information circule en arrière-plan qu'il conviendrait d'exploiter en partenariat avec ces dernières (recherche, annulation, motifs de réservation, moments de navigation, …) pour s'adapter aux attentes fluctuantes de la nouvelle génération opportuniste de la clientèle. Aussi, y a-t-il lieu de développer des partenariats avec ces plateformes afin d'exploiter leurs potentiels d'influence via leurs outils de plus en plus avancés de suggestion[58]. La capacité de ces plateformes pourrait, également, être déployée pour enrichir l'offre marchande dans des territoires peu dotées[59] en utilisant les nouvelles approches prédictives et de marketing (intelligence artificielle[60], machine learning, …) développées à la Big Data[61]. Ces plateformes pourraient, en plus, jouer le rôle important de tierce personne pour le recouvrement des taxes touristiques au bénéfice de l'Etat ou des collectivités locales. Le rôle de ces plateformes serait également très pertinent dans la labélisation verte des hébergements touristiques et leurs conseils quant aux attentes des touristes écoresponsables.

Ainsi, ce premier travail, dont la DEPF a été pionnière, lève, en partie, le voile sur les opportunités offertes par les données « non conventionnelles » dans l'accomplissement de sa mission d'analyse économique, de veille et de prévision. D'autres possibilités d'analyse sont envisageables ultérieurement qui appelle à créer une masse critique pluridisciplinaire d'experts nationaux à travers une coopération publique-privée-université à la hauteur des défis et des enjeux qui se dressent, notamment, en termes d'intelligence économique et de fracture de connaissance qui risquerait de pénaliser la vitesse de l'émergence socio-économique du pays.

---

56 Les petites entreprises d'hébergement ont du mal à obtenir le soutien d'institutions gouvernementales pour se développer et grandir. Siyabonga Mxunyelwa and Unathi Sonwabile Henama « Small to Medium Tourism Enterprises (SMTEs) promoting Local Economic Development in Hogsback, Eastern Cape, South Africa », African Journal of Hospitality, Tourism and Leisure, Volume 8 (3).

57 Les touristes seraient prêts à payer les prix des hôtels conventionnels pour un hôtel vert et pourraient tolérés des inconvénients mineurs (réutilisation des serviettes, utilisation de produits recyclés, …). Kim, Yunhi. (2010). « Intention to pay conventional-hotel prices at a green hotel – a modification of the theory of planned behaviour ». Journal of Sustainable Tourism.

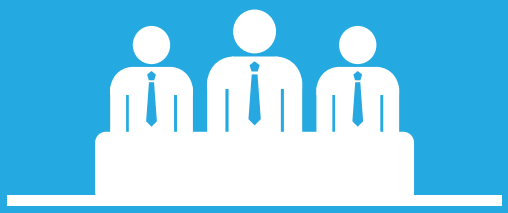
58 Google détient 92,8% du marché mondial de la recherche à fin octobre 2019 (97,5% du marché français et 98,9% du marché espagnol de recherche mobile qui sont les principaux marchés émetteurs du Maroc) avec des outils d'orientations des résultats de plus en plus sophistiqués. Cette hégémonie est mondialement acquise mis à part pour les marchés émetteurs émergeants pour le Maroc, à savoir, russe (Yandex (44,3%) et Google (52%)) et chinois (Baidu (61,1%), Sogou (24,1%) et Shenma (7,4%)). Source : Statcounter.

59 Le Comité Régional de Tourisme de la Nouvelle-Aquitaine a signé en mars 2019 une convention avec Airbnb pour le développement de l'activité touristique dans les départements ayant un fort potentiel non exploité.

60 Les régions de Rabat-Salé-Kénitra et Casablanca-Settat devraient accueillir ces filières dans les nouvelles cités des métiers et des compétences tel que présenté en avril 2019 dans la feuille de route relative au développement du secteur de la formation professionnelle.

61 Booking a mis en place une cellule spéciale d'intelligence artificielle au service des hébergeurs https://booking.ai/..

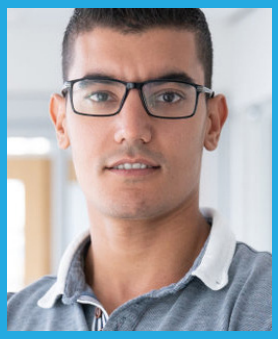
Etude de cas N° 2



Usage des sciences des données pour améliorer l'employabilité des jeunes au Maroc

Imane KHAOUJA

Ibtissam MAKDOUN


Ibrahim RAHHAL

Ghita MEZZOUR

Ecole Supérieure d'Informatique et du Numérique et TICLab (Université Internationale de Rabat)

Chaque année, un grand nombre de jeunes diplômés au Maroc n'arrivent pas à trouver un emploi. Le taux de chômage des jeunes au Maroc est 27,4%[62]. Paradoxalement, beaucoup d'entreprises peinent à trouver des profils adéquats. Ceci reflète un décalage entre les formations universitaires et les besoins du marché du travail et pose des problèmes sociaux, économiques et politiques.

Ce projet a développé et utilisé des techniques de sciences des données pour analyser la situation. Plus précisément, l'équipe collecte de manière hebdomadaire toutes les offres d'emploi publiées sur plus de 10 sites d'emploi marocains. Ces offres sont ensuite traitées et analysées pour ressortir les besoins du marché de l'emploi marocain. Le projet s'est attelé, aussi, à comparer les besoins du marché du travail à des programmes universitaires pour ressortir les inadéquations des compétences. Finalement, le projet a analysé le degré de collaboration entre les parties prenantes[63] de l'emploi car une telle collaboration est importante pour réduire le décalage entre le monde universitaire et le monde du travail.

---

62 https://www.hcp.ma/Taux-de-chomage-au-niveau-national-selon-les-tranches-d-age_a262.html

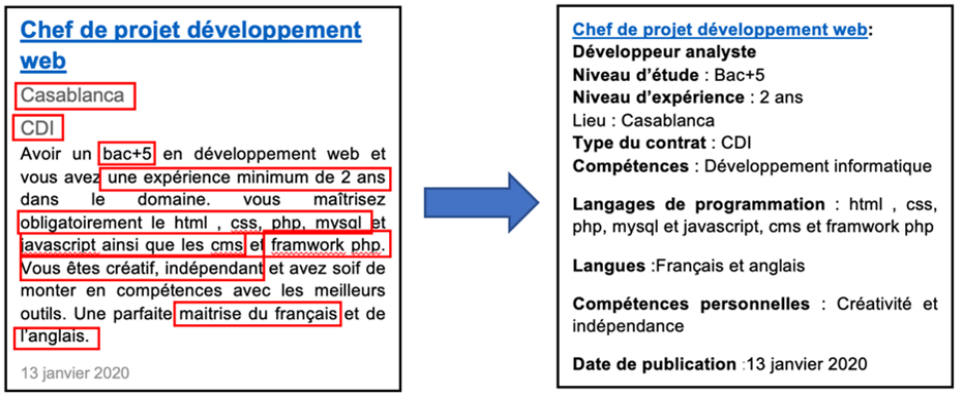
63 I. Khaouja, I. Makdoun, G. Mezzour, I. Rahhal, H. Benchekroun, Y. El Hatib, and I. Kassou. Social Network Analysis of Job Market Stakeholders in Morocco. Short paper, International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), July 2017, Washington, DC, USA

1. ÉTUDE DES BESOINS DU MARCHÉ DU TRAVAIL

Le projet s'est intéressé aux offres d'emploi mises en ligne pour 3 secteurs au Maroc à savoir : l'offshoring[64] (IT offshoring et centres d'appel), l'automobile[65] et la cyber sécurité[66]. L'offshoring et l'automobile ont une grande importance dans le plan émergence. Quant à la cyber-sécurité, c'est un domaine pointu qui est important pour donner confiance aux investisseurs. La méthodologie développée peut être étendue à d'autres secteurs du secteur productif national.

Dans un premier temps et pour des raison d'harmonisation de l'information, il a été question de procéder à suppression des offres d'emploi dupliquées en utilisant l'algorithme Simhash développé par Google. Cet algorithme permet de repérer et de supprimer les pages en double. Ensuite, les informations pertinentes comme les compétences techniques, les compétences personnelles[67], les langues, la ville, le salaire, le type de contrat, le niveau d'études, et les années d'expérience sont extraites et normalisées. Le graphe 7 illustre ce processus d'extraction d'information d'une offre.

Graphe 7 : Exemple d'extraction des compétences à partir d'une offre d'emploi



Le tableau ci-dessous décrit le nombre des offres d'emploi et la durée prise en considération dans l'étude.

Tableau 2 : Description des offres collectées dans les différents secteurs

Secteur	Nombre d'offres d'emploi	Durée	Région
Automobile	8000	1 an (02/2017 – 06/2018)	Tout le Maroc
IT offshore	1514	6 mois (02/2017 – 08/2017)	Casablanca
Call centers	26574	6 mois (02/2017 – 08/2017)	Casablanca
Cyber-Sécurité	76	18 mois (02/2017 –08/2018)	Tout le Maroc

Le graphe 8 met en avant les occupations les plus demandées dans l'IT offshore et les centre d'appel. L'IT offshore demande une panoplie de profils notamment les développeurs analystes, les administrateurs réseau et les designers web. Par contre, les centres d'appel demandent principalement des agents de call center et des agents de ressources humaines. En outre, le secteur de l'automobile demande des profils très variés allant des ingénieurs jusqu'au opérateurs. Le secteur de la cybersécurité recrute principalement des administrateurs réseau et des responsables de sécurité.

---

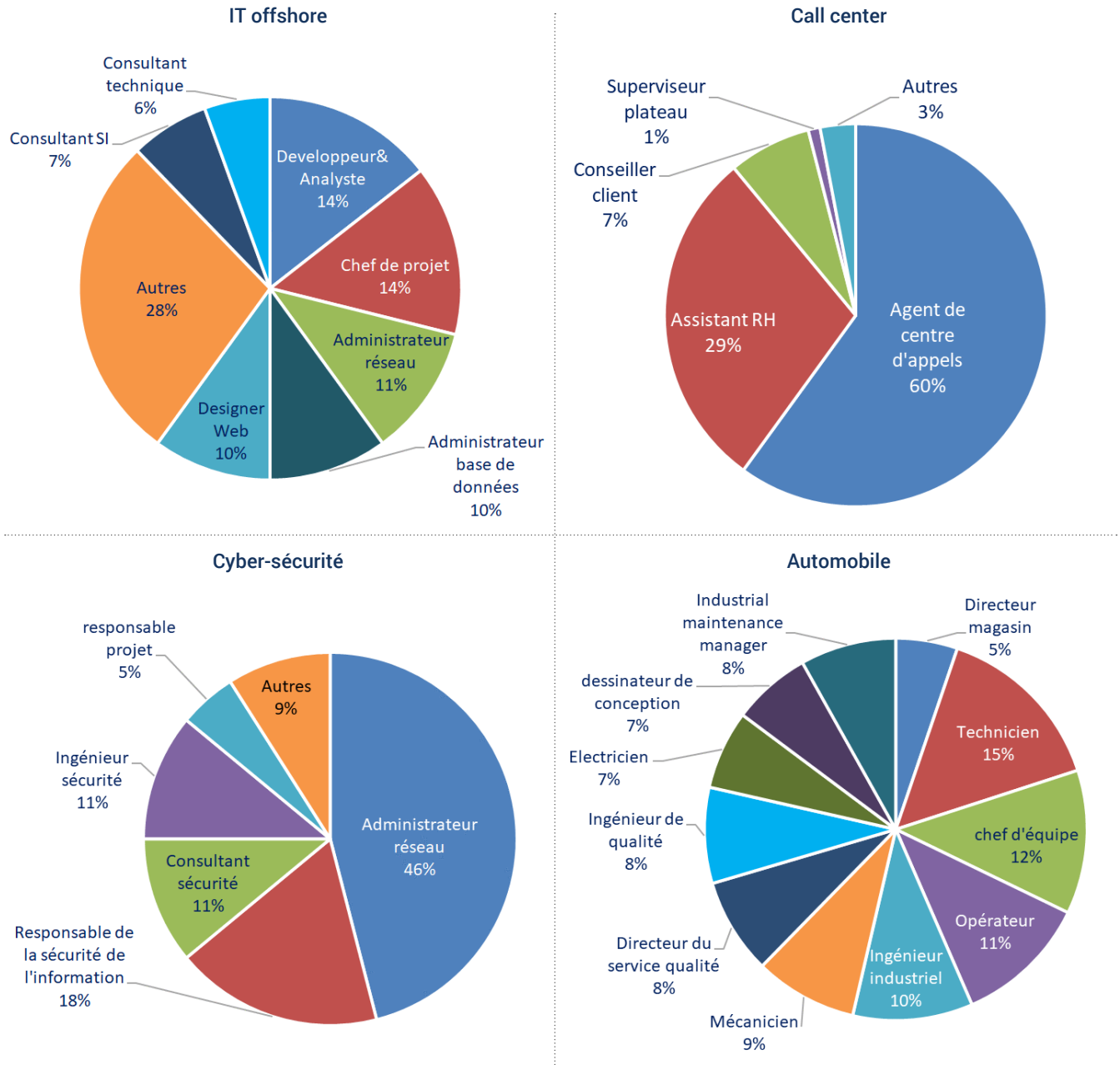
[64] Imane Khaouja, Ibrahim Rahhal, Mehdi El Ouali, Ghita Mezzour, Kathleen M. Carley, and Ismail Kassou. Analyzing the Needs of the Offshore Sector in Morocco. IEEE Global Engineering Education Conference (EDUCON), April 2018, Santa Cruz de Tenerife, Canary Islands, Spain.

[65] I. Makdoun, G. Mezzour, K. Carley, I. Kassou. Analyzing the needs of the Automotive Job Market in Morocco. Proceedings of the 13th International Conference on Computer Science and Education (ICCSE 2018): Colombo, Sri Lanka, 2018.

[66] Ibrahim Rahhal, Ibtissam Makdoun, Ghita Mezzour, Imane Khaouja, Kathleen M. Carley, and Ismail Kassou. Analyzing Cybersecurity Job Market Needs in Morocco by Mining Job Ads. IEEE Global Engineering Education Conference (EDUCON), April 2019, Dubai, Émirats arabes unis.

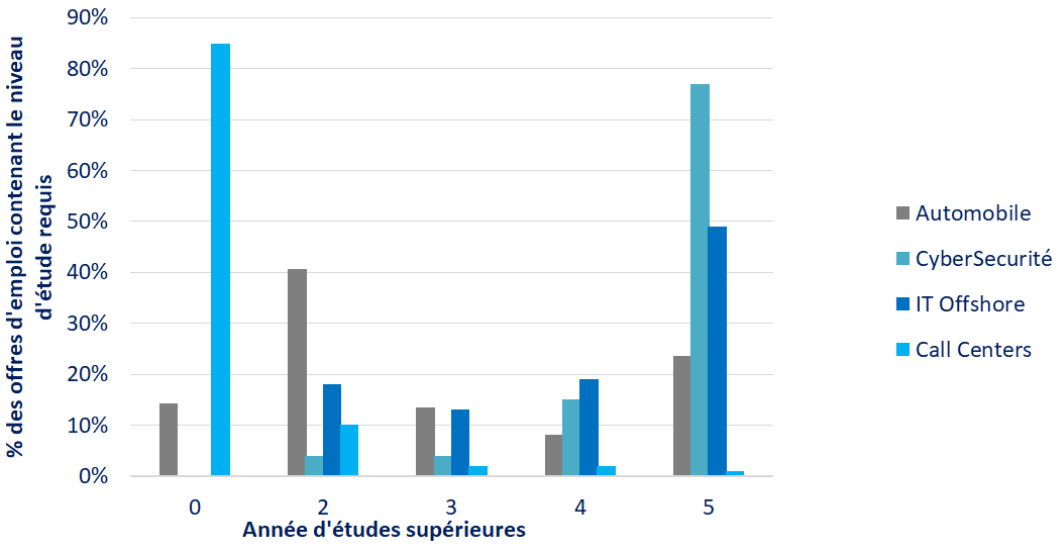
[67] Imane Khaouja, Ghita Mezzour, Kathleen M. Carley, and Ismail Kassou. Building a Soft Skill Taxonomy from Job Openings. Social Network Analysis and Mining 9(1) 1:43, July 2019.

Graphe 8 : Les occupations les plus demandées dans les secteurs étudiés



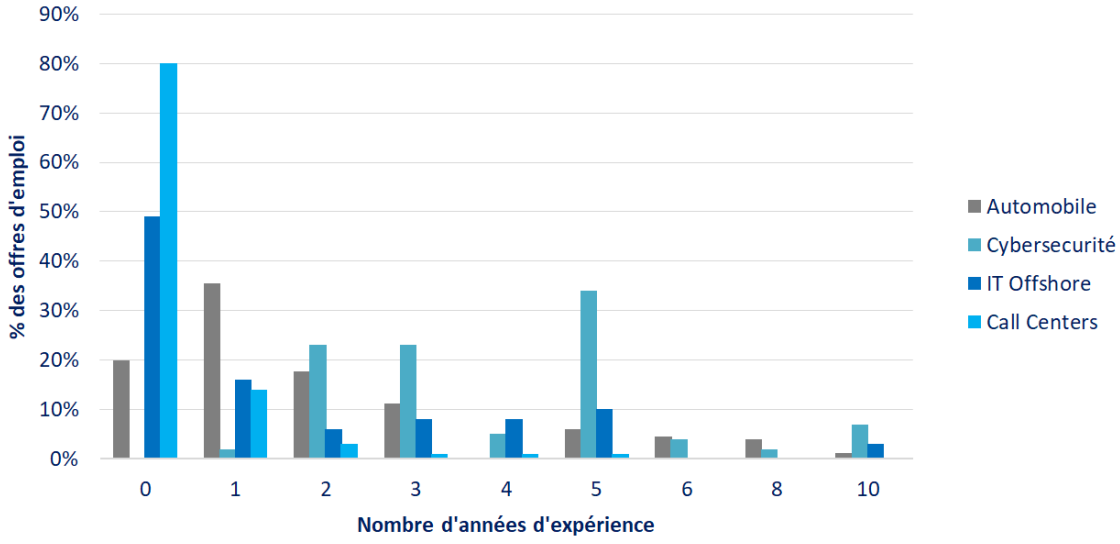
Concernant les études supérieures, le graphe 9 montre les niveaux d'étude les plus demandées par les différents secteurs étudiés. Les points phares qui en ressortent sont que les centres d'appel recrutent surtout des niveaux bac ce qui représente une opportunité pour les jeunes ayant ce niveau d'éducation et sans aucun diplôme professionnel. Quant au secteur de l'automobile, il recrute surtout des techniciens et les bac+5 viennent en second lieu. Pour l'IT offshore et la cyber sécurité, ces secteurs recrutent principalement des bac+5 ce qui reflètent qu'un travail intellectuel de développement s'effectue au Maroc.

Graphe 9: Nombre d'années d'études supérieurs demandées dans les secteurs étudiés



Relativement à l'expérience requise, le graphe 10 met en exergue les années d'expérience demandées dans les secteurs étudiés. L'offshoring (IT offshore et centres d'appel) ne demande pas forcément d'expéreience préalable, ce qui constitue une opportunité importante pour les jeunes fraichement diplômés. Quant à l'automobile, le secteur préfère des profils ayant peu d'expérience au préalable. Cette expérience pourrait être acquise grâce à des stages ou des formations en alternance. Tandis que la cyber sécurité privilégie les profils expérimentés.

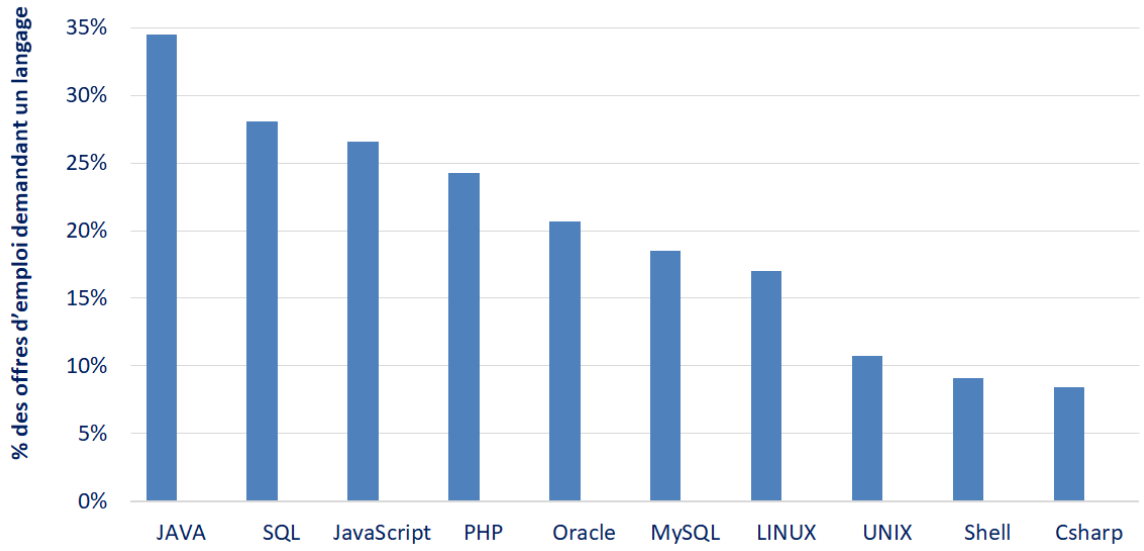
Graphe 10: Nombre d'années d'expérience demandées dans les secteurs étudiés



A partir de l'analyse des langages informatiques les plus demandés, le graphe 11 montre que Java, SQL, Javascript et PHP sont très demandés. Javascript et PHP servent au développement web et sont peu enseignés dans la plupart des universités Marocaines. En outre, le graphe révèle que le classement des langages informatique au Maroc diffère du classement mondial[68] d'où la nécessité d'analyser les besoins locaux pour avoir des recommandations plus adaptées au contexte marocain.

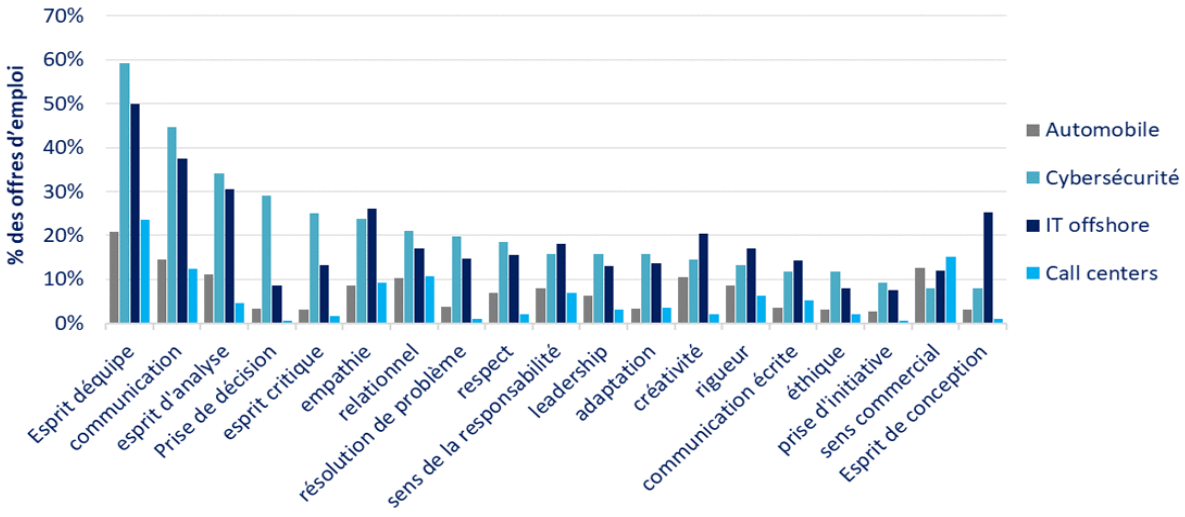
[68] https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2019

Graphe 11: Les langages de programmation demandés dans les secteurs étudiés



Le Graphe 12 révèle que l'esprit d'équipe et la communication et l'esprit d'analyse sont les compétences les plus demandées dans tous les secteurs étudiés. Ces compétences peuvent par exemple être développées à travers des projets que les étudiants effectuent en groupe et présentent ensuite en classe. Les compétences personnelles sont complémentaires aux compétences techniques et jouent un rôle primordial dans les processus de recrutement et promotion (Kautz et Al. 26).

Graphe 12: les compétences personnelles les plus demandées dans les secteurs étudiés

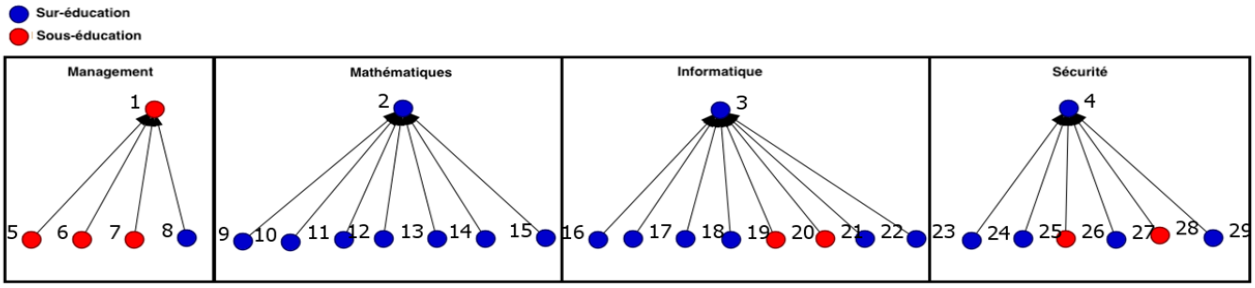


2. ANALYSE DE L'INADÉQUATION DES COMPÉTENCES

Le projet a examiné l'adéquation entre les besoins du marché du travail de la cyber sécurité et les compétences enseignées dans les formations universitaires spécialisés dans ce domaine. Les besoins du marché du travail sont identifiés à travers les offres d'emploi et les compétences enseignées sont identifiées à travers les syllabi des universités.

Une taxonomie hiérarchique des compétences informatiques a d'abord été développée grâce aux données de Dbpedia. Ensuite, la fréquence des mentions des différentes compétences dans les offres d'emploi et les syllabi des universités a été calculée, normalisée et comparée.

Graphe 13 : L'inadéquation entre les compétences du premier niveau de la hiérarchie



Le Graphe 13 présente les résultats de cette analyse en se concentrant sur le premier niveau de la hiérarchie des compétences. Le graphe montre que les compétences mathématiques sont en sur-éducation aux universités, aussi que les compétences liées au management sont en sous-éducation. Nos principaux résultats indiquent que les compétences liées au Pare-feu (Firewall) sont en sous-éducation cependant ces compétences sont très demandées au secteur de la cyber sécurité.

Il est important de noter que le projet a également effectué un questionnaire de 79 professionnels de la cyber sécurité et 35 professeurs qui enseignent dans des programmes sur la cyber sécurité. Le questionnaire a demandé à chacun de ces individus de sélectionner les compétences les plus demandées par le marché de l'emploi. Le questionnaire a révélé des résultats similaires à la méthodologie développée. Ceci indique que cette méthodologie orientée science de données peut être utilisée pour identifier les compétences en inadéquation.

Tableau 3: Nom complet des nœuds présenté dans le Graphe 15

Compétences	Identifiant du noeud	Compétences	Identifiant du noeud
Management	1	Architecture des ordinateurs	16
Mathématiques	2	Matériel	17
Informatique	3	Réseaux informatiques	18
Sécurité	4	Programmation informatique	19
Outil du management	5	Bases de données	20
Gestion du projet	6	Systèmes d'exploitation	21
Management spécifique	7	Logiciels	22
Normes	8	Informatique théorique	23
Algèbre	9	Contrôle d'accès	24
Algorithmiques	10	Biométrie	25
Cryptographie	11	Nucléaire	26
Systèmes dynamiques	12	Risques	27
Analyses mathématiques	13	Sécurité informatique	28
Théorie des nombres	14	Politique de sécurité	29
Statistiques	15		

3. ANALYSE DES PARTIES PRENANTES DE L'EMPLOI AU MAROC

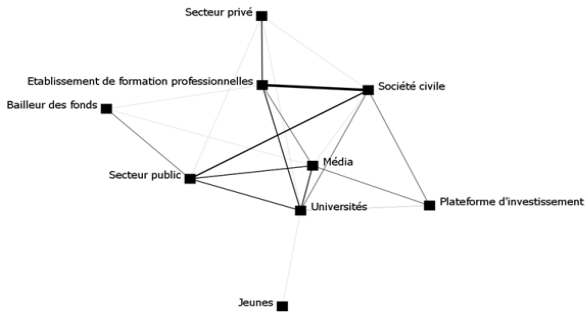
3.1. ANALYSE DE LA COLLABORATION PAR QUESTIONNAIRE

Afin d'analyser le degré de collaboration entre les parties prenantes de l'emploi, le projet a analysé les résultats d'un questionnaire administré par les centres de carrière de l'USAID en collaboration avec des universités Marocaine en 2016. Plus de 250 représentants d'universités, d'établissements de formation professionnelle, de secteur public, de secteur privé, société civile, bailleurs de fonds, plateformes de placement et agences de ressources humaines (RH) à Marrakech, Tanger et Casablanca ont participé au questionnaire. Ce questionnaire s'est intéressé à plusieurs formes de collaboration comme l'échange d'information sur le marché de travail et l'appui financier et technique.

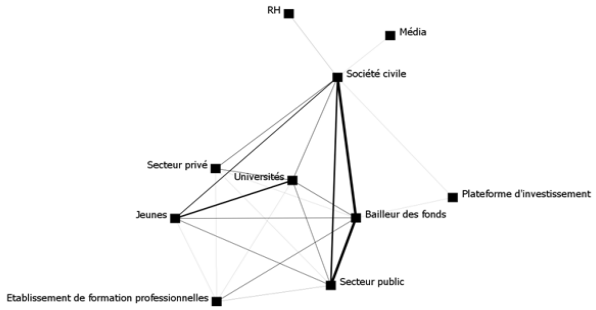
Le Graphe 14 indique un faible niveau d'échange d'informations sur le marché de travail entre les universités et le secteur privé. Ceci indique que les universités n'ont pas forcément les informations nécessaires pour aligner leurs formations au marché de l'emploi. Par contre, les établissements de formation professionnelle semblent mieux liés aux autres parties prenantes.

Le graphe 14 révèle aussi que les jeunes reçoivent très peu d'informations sur le marché de travail des autres parties prenantes ce qui ne leur permet pas de bien choisir les formations à suivre. Paradoxalement, ces jeunes reçoivent de l'appui technique et financier des autres parties prenantes. Ceci indique que le manque d'échange d'information n'est pas forcément dû à un manque de moyens, mais pourrait refléter des différences culturelles entre les parties prenantes.

Graphe 14 : Collaboration entre les parties prenantes de l'emploi



Information sur marché de travail



Appui technique et financier

3.2. ANALYSE DE LA COLLABORATION ENTRE LES PARTIES PRENANTES À TRAVERS LEURS RAPPORTS OFFICIELS

Afin de comprendre les perceptions et les sujets de préoccupation des parties prenantes, le projet a ensuite collecté et analysé les rapports officiels de ces parties prenantes et les articles de presse les concernant. Pour illustrer ceci, le graphe 15 permet de voir une différence entre les perceptions d'une université et d'une entreprise marocaines. En effet, l'université s'intéresse principalement à la formation, à la recherche et aux étudiants. L'entreprise, elle, s'intéresse principalement à son domaine d'activité. On trouve très peu de sujets communs.

Graphe 15 : Nuage des mots extraites à partir des rapports officiels

Entreprise

Université

4. ENSEIGNEMENTS ET RECOMMANDATIONS

Ce projet révèle que les besoins du marché de travail varient par secteur. Les call centers s'intéressent peu aux diplômes universitaires et l'expérience des candidats. Par contre, l'IT offshore et la cyber-sécurité demandent des profils bac+5 et des compétences pointues. Le secteur automobile demande des profils ingénieurs et des profils techniciens. Ce travail a également identifié les compétences techniques et personnelles requises par les différents secteurs. Ces compétences diffèrent souvent des compétences demandées par le marché international d'où la nécessité d'examiner le marché marocain. Par exemple, les langages informatiques les plus demandés sont Java, SQL, JavaScript et PhP. Les compétences personnelles les plus demandées sont l'esprit d'équipe, la communication et l'esprit d'analyse. Le projet a identifié plusieurs compétences en inadéquation entre les formations universitaires et les besoins du marché du travail.

Une analyse des parties prenantes de l'emploi a révélé très peu d'échange d'informations à propos des besoins du marché de l'emploi. Ceci implique que les universités n'ont pas forcément les informations nécessaires pour adapter leurs formations au marché de l'emploi. Pareil, les jeunes ont du mal à choisir des formations avec des débouchés de travail intéressantes.

Ce projet s'est principalement appuyé sur l'analyse du texte pour effectuer différentes analyses et ressortir des informations importantes. L'analyse des offres d'emploi pour identifier les besoins du marché de l'emploi exploite le fait que les recruteurs expriment leurs besoins à travers ces annonces publiées en ligne. Par conséquent, cette approche ne couvre pas le marché informel et les offres véhiculées exclusivement en bouche à oreille. Elle n'est donc pas appropriée aux secteurs peu structurés comme l'agriculture. Pour les secteurs structurés, cette approche peut s'avérer un outil puissant pour identifier les besoins du marché de travail en temps réel et à des coûts intéressants.

Au cours de ce projet, plusieurs réunions avec les parties prenantes ont été organisées afin de les sensibiliser à ces problématiques. L'Université Internationale de Rabat a adapté son curricula sur la base de cette étude afin d'améliorer l'employabilité de ses diplômés.

Remerciement

Ce projet est financé par l'Agence des Etats-Unis pour le Développement International (USAID) sous subvention AID-OAAA-11-00012. Les conclusions présentées dans ce document reflète les opinions des auteurs et ne devraient pas être interprétées comme étant les politiques officielles de l'USAID.

Conclusion générale : Défis et implications statistiques des Big Data

À l'ère des Big Data, la prise de décision ne peut plus se priver de l'intelligence et des bénéfices que ne cessent d'engendrer ces technologies d'analyse pour prendre plus rapidement des décisions stratégiques dans des environnements plus complexes et incertains. De nombreuses entreprises privées ainsi que des organisations nationales et internationales considèrent que les « Big data» n'est pas un simple mot à la mode, mais un outil indispensable pour apporter des réponses à leurs besoin d'efficacité, de pertinence, de portée stratégique et de prise de décision.

En temps de crise mondiale, les "big data" ont gagné d'importance et ont été l'un des outils les plus influents de la réponse mondiale à la pandémie pour prédire les résultats potentiels et sauver des vies. En effet, dans le cas de COVID-19, des données importantes ont été accumulées de tous les points de données du monde entier. La modélisation mathématique a pris ces données et les a utilisées pour identifier les points géographiques critiques de la maladie, créer des modèles de prédiction de décès, fournir des estimations concernant les tests et le besoin de fournitures pour les tests, et guider la prise de décision parmi les décideurs politiques, les prestataires de soins de santé et d'autres acteurs clés. Les informations utilisées sont précieuses et les données importantes peuvent sauver des vies, mais elles ne sont pas infaillibles et perdent de leur valeur lorsqu'elles ne sont pas combinées avec le point de vue de la science et avec la réalité sur le terrain.

Les mégadonnées sont évolutives et peuvent fournir des informations novatrices, en temps réel et plus détaillées pour l'analyse économique et financière. Cependant, les opportunités des Big Data pour chaque pays sont asymétriques et dépendent des caractéristiques du pays et de la disponibilité des systèmes et réseaux générant les Big Data.

Les Big Data peuvent être utiles aux statistiques macroéconomiques et financières et, en définitive, à l'élaboration des politiques grâce à trois caractéristiques essentielles :

- En répondant à de nouvelles questions et en produisant de nouveaux indicateurs ;
- En rattrapant les retards dans la disponibilité des statistiques officielles et en soutenant la prévision des indicateurs existants ;
- En tant que source de données innovante dans la production de statistiques officielles.

Les Big Data sont extrêmement utiles et leur succès ne réside pas dans la mise en œuvre d'une technologie particulière, mais dans la création d'un environnement de personnes et de processus qui exploite les innovations des Big Data. Compte tenu de la diversité des compétences requises pour gérer les Big Data, il offre également aux organisations la possibilité de briser leurs cloisonnements internes, y compris entre les utilisateurs et les producteurs de données et de statistiques.

En outre, la coopération statistique internationale est essentielle pour surmonter les défis liés aux Big Data et pour établir des partenariats durables entre les agences statistiques nationales et internationales, les utilisateurs et les propriétaires de données. Les opportunités, les défis et les implications potentielles sont particulièrement élevés pour les agences nationales de statistique. L'incorporation des Big Data en tant que nouvelles sources de données, complétant ou se substituant aux sources de données traditionnelles, ne sera pas épargnée par les défis méthodologiques, organisationnels et budgétaires. En outre, l'utilisation croissante des mégadonnées dans différents pays montre la valeur de la coopération internationale et de l'apprentissage des autres. Ainsi, et au-delà des défis de l'utilisation de ces Big Data, la communauté statistique internationale devrait travailler mutuellement sur de nouvelles normes pour les statistiques officielles.

De plus, pour se tenir au courant des développements, les agences doivent rechercher de manière proactive des sources de données volumineuses afin de répondre aux besoins de recherche les plus urgents. Certains projets Big Data peuvent également être intégrés aux activités de développement des capacités afin d'aider les membres à renforcer leurs capacités et de tirer parti des sources Big Data disponibles. À l'avenir, la recherche et la compilation des meilleures pratiques - pour des techniques et méthodologies statistiques qui traitent de la véracité et de la volatilité, en particulier - doivent figurer en haut de l'ordre du jour de la communauté des statistiques.

Etant donné que le Big Data n'est pas statique mais dynamique, les systèmes et réseaux générant des Big Data continuent d'évoluer, entraînant avec eux la possibilité des défis pour la statistique. Les Big Data est un complément de statistiques publiques plutôt que substitution. Toutefois des règles de contrôle sont indispensables concernant la qualité des données, la rigueur statistique et la protection des informations personnelle.

/ Références bibliographiques

1. Andrade, S. C., Bian, J. & Burch, T. R. (2009). "Does information dissemination mitigate bubbles? The role of analyst coverage in China". University of Miami Working Paper.
2. Baldacci, E., Buono, D. A. R. I. O., Kapetanios, G. E. O. R. G. E., Krische, S. T. E. P. H. A. N., Marcellino, M. A. S. S. I. M. I. L. I. A. N. O., Mazzi, G. L., & Papailias, F. O. T. I. S. (2016). Big Data and Macroeconomic Nowcasting: from data access to modelling. Luxembourg: Eurostat. Doi: http://dx. doi. org/10.2785/360587.
3. Banbura, M., Domenico G., Michele M. & Lucrezia R. (2013), "Now-Casting and the Real-Time Data Flow," In Handbook of Economic Forecasting, edited by Graham Elliott, Clive Granger, and Allan Timmermann (Amste rdam: Elsevier).
4. Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with Big Data. Federal Reserve Bank of New York Staff Reports. Annual Review of Economics, 10, 615-643.
5. Bortoli, C., Combes, S. & Renault, T. (2018). "Nowcasting GDP Growth by Reading Newspapers". INSEE, Economie et Statistique / Economics and Statistics, 505-506 , 17–33 .
6. Breton, R., N. Swiel and R. O'Neil (2015), 'Using Web Scraped Data to Construct Consumer Price Indices', New Techniques and Technologies for Statistics, Eurostat Conference, 9–13 March 2015.
7. Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G. L., & Papailias, F. Evaluation of Nowcasting/Flash Estimation based on a Big Set of Indicators. Paper prepared for the 16th Conference of IAOS. OECD Headquarters, Paris, France, 19-21 September 2018
8. Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). "Big Data types for macroeconomic nowcasting". Eurostat Review on National Accounts and Macroeconomic Indicators, 1(2017), 93-145.
9. Chen H, Chiang RHL, Storey V. 2012. Business intelligence and analytics: From Big Data to big import. MIS Quarterly 36(4):1165 –1188.
10. Davenport, T. 2006. "Competing on Analytics.", Harvard Business Review. https://hbr.org/2006/01/competing -on-analytics.
11. Dong L. et al. (2017), "Measuring economic activity in China with mobile Big Data", Big Data Lab, Baidu Research, Baidu, Beijing, China.
12. Doornik, J. A. and D.F. Hendry (2015), 'Statistical Model Selection with Big Data', Cogent Economics & Finance , 3(1), 2015.
13. El Alaoui, I. (2018). Transformer les big social data en prévisions-méthodes et technologies: Application à l'analyse de sentiments (Doctoral dissertation, Angers).
14. Elshendy, M., & Fronzetti Colladon, A. (2017). Big Data analysis of economic news: Hints to forecast macroeconomic indicators. International Journal of Engineering Business Management, 9, 1847979017720040.
15. Ferreira, P. (2015), 'Improving Prediction of Unemployment Statistics with Google Trends: Part 2', Eurostat Working Paper.
16. Galbraith, J.W and G. Tkacz (2015), 'Nowcasting GDP with electronic payments data', European Central Bank, Working Paper No 10 / August 2015.
17. Griffioen, R., J. de Haan and L. Willenborg (2014), 'Collecting Clothing Data from the Internet', Statistics Netherlands Technical Report.
18. Hammer, C., Kostroch, M. D. C., & Quiros, M. G. (2017). Big Data: Potential, Challenges and Statistical Implications. International Monetary Fund.

19. Hassani, H., & Silva, E. S. (2015). Forecasting with Big Data: A review. Annals of Data Science, 2(1), 5-19.

20. Hebous, Shafik, and Tom Zimmermann (2016), "Can Government Demand Stimulate Private Investment? Evidence from U.S. Federal Procurement," IMF Working Paper 16/60, (Washington: International Monetary Fund).

21. Heston, S.L. and N.R. Sinha (2014), 'News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns', Working Paper.

22. Hoopes, Jeffrey, Patrick Langetieg, Stefan Nagel, Daniel Reck, Joel Slemrod, and Bryan Stuart, 2016, "Who Sold During the Crash of 2008–9? Evidence from Tax-Return Data on Daily Sales of Stock." NBER Working Paper 22209, National Bureau of Economic Research, (Cambridge, MA).

23. Jiang S. (2014), "Using web scraping price data for price index of e-commerce", United Nations, Big Data Project Inventory.

24. Jiang S. (2014), "Crop survey by farmland: using satellite and aerial remote sensing to help estimate agricultural statistics", United Nations, Big Data Project Inventory.

25. Kapetanios, G., & Papailias, F. (2018). Big Data & macroeconomic nowcasting: Methodological review. ESCoE Discussion Paper, (12).

26. Kautz T, Heckman JJ, Diris R, Weel Bt, Borghans L (2014) Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success. Working Paper 20749, National Bureau of Economic Research. https ://doi.org/10.3386/w2074 9.

27. Koop, G. and L. Onorante (2013), 'Macroeconomic Nowcasting Using Google Probabilities', European Central Bank Presentation.

28. Marc D., "Feasibility study on the use of mobile telephone data for tourism & transportation statistics", Belgium - Statistics Belgium, United Nations, Big Data Project Inventory.

29. Matthew S. et Kenneth K. (2011), "Nowcasting Chinese GDP: Information Content of Economic and Financial Data.", Hong Kong Institute for Monetary Research, HKIMR Working Paper No.04/2011.

30. Misch, F., Olden, M. B., Poplawski-Ribeiro, M., & Kejji, L. (2017). Nowcashing: Using Daily Fiscal Data for Real-Time Macroeconomic Analysis. International Monetary Fund.

31. Nyman, R., D. Gregory, S. Kapadia, R. Smith and D. Tuckett (2014a), 'Exploiting Big Data for Systemic Risk Assessment: News and Narratives in Financial Systems', Working Paper, ECB Workshop on using Big Data for forecasting and statistics, 07-08/04/2014, Frankfurt.

32. Rahal, Charles, 2016, "Unlocking Public Payments Data," Unpublished, (University of Oxford).

33. Reis, F., P. Ferreira and V. Perduca (2015), 'The Use of Web Activity Evidence to Increase the Timeliness of Official Statistics Indicators', Eurostat Working Paper.

34. Robin, F. (2018). "Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys.", INSEE Economie et Statistique / Economics and Statistics , 505-506, 35–63.

35. Richardson, P. (2018). Nowcasting and the Use of Big Data in Short-Term Macroeconomic Forecasting: A Critical Review. Economie et Statistique / Economics and Statistics , 505-506, 65–87.

36. Sakarovitch, B., Bellefon, M. (de), Givord, P. & Vanhoof, M. (2018). "Estimating the Residential Population from Mobile Phone Data, an Initial Exploration". INSEE, Economie et Statistique / Economics and Statistics, 505-506, 109–132.

37. Shi Y. (2014), "Big Data: history, current status, and challenges going forward". Bridge 44(4): 6-11.

38. United Nations Global Working Group (UNGWG). 2017. Big Data. http://unstats.un.org/bigdata. Survey and Project Inventory, http://unstats.un.org/bigdata/inventory.

/Annexe : Synthèse des travaux empiriques utilisant les Big Data pour la surveillance et la prévision économique

Domaine	Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Suivi et prévision du PIB et ses composantes	Galbraith, J. W., Tkacz, G. (ECB)	- Banque Centrale européenne/Canada - Prévision immédiate du PIB avec des données de paiements électroniques	- Données: transactions par carte de débit et par carte de crédit, chèques, PIB décalé, indice du logement, emplois dans les ventes aux particuliers et aux entreprises, indice boursier, FCI, masse monétaire, semaine moyenne de travail (heures), nouvelles commandes, biens durables, stocks, commerce de détail. - Méthodologie: principalement basée sur des régressions linéaires et les modèles factoriels.	En contrôlant les dates de diffusion de chacun des indicateurs, ils génèrent des prévisions immédiates de la croissance du PIB pour un trimestre donné sur une période de cinq mois, soit la période au cours de laquelle un intérêt pour les prévisions immédiates existerait.	Ils constatent que les erreurs de prévision immédiate chutent d'environ 65% entre la première et la dernière prévision immédiate. Parmi les variables de paiement considérées, les transactions par carte de débit semblent apporter les améliorations les plus importantes en termes de précision des prévisions.
	Koop, G. (Université de Strathclyde), Onorante, L. (ECB)	- Banque Centrale européenne/Etats Unis - Prévision macroéconomique immédiate à l'aide des probabilités de Google	- Données: Inflation, Inflation salariale, Chômage, Inflation du prix des produits de base, Production industrielle, Inflation du prix du pétrole, Masse monétaire - Méthodologie : modèle de moyenne dynamique	Dans un exercice empirique impliquant neuf variables macroéconomiques mensuelles majeures aux États-Unis, ils ont constaté que les méthodes DMS apportaient de grandes améliorations à la prévision immédiate.	L'utilisation des probabilités de modèle de Google dans le DMS est souvent plus performante que le DMS conventionnel.
	Matthew S. et Kenneth K. (2011)	-Institut de recherche monétaire de Hong Kong/ Chine - Prévisions actuelles du PIB chinois: contenu informatif des données économiques et financières	- Méthodologie : le modèle factoriel proposé par Giannone, Reichlin et Small (2005) et les critères de Bai et Ng (2002) pour déterminer le nombre de facteurs communs dans le modèle factoriel. - Données: un ensemble de données volumineuses qui contient 189 séries d'indicateurs de plusieurs catégories, telles que les prix, la production industrielle, les investissements en immobilisations, le secteur extérieur, le marché monétaire et le marché financier.	Le modèle identifié génère des prévisions immédiates hors échantillon pour le PIB chinois avec des erreurs de prévision moyennes au carré moins grandes que celles du point de référence Random Walk.	De plus, en utilisant le modèle factoriel, les auteurs constatent que les données de taux d'intérêt constituent le bloc le plus important pour l'estimation du PIB du trimestre en cours en Chine. Les données sur les prix à la consommation et au détail et les indicateurs d'investissement en immobilisations constituent d'autres blocs importants.

	Galbraith, J. W., Tkacz, G. (ECB)	- Banque Centrale européenne/Canada - Prévision immédiate du PIB avec des données de paiements électroniques	- Données: transactions par carte de débit et par carte de crédit, chèques, PIB décalé, indice du logement, emplois dans les ventes aux particuliers et aux entreprises, indice boursier, FCI, masse monétaire, semaine moyenne de travail (heures), nouvelles commandes, biens durables, stocks, commerce de détail. - Méthodologie: principalement basée sur des régressions linéaires	Galbraith et Tkacz (2015) évaluent l'utilité d'un vaste ensemble de données de paiements électroniques comprenant des transactions par carte de débit et de crédit, ainsi que des chèques émis dans le système bancaire, en tant qu'indicateurs potentiels de la croissance actuelle du PIB au Canada. Ces variables capturent un large éventail d'activités de dépense et sont disponibles très rapidement, ce qui en fait des indicateurs actuels appropriés. Bien que chaque transaction effectuée avec ces mécanismes de paiement soit en principe observable, les données sont agrégées pour les prévisions macroéconomiques.	En contrôlant les dates de diffusion de chacun des indicateurs, ils génèrent des prévisions immédiates de la croissance du PIB pour un trimestre donné sur une période de cinq mois, soit la période au cours de laquelle un intérêt pour les prévisions immédiates existerait. Ils constatent que les erreurs de prévision immédiate chutent d'environ 65% entre la première et la dernière prévision immédiate. Parmi les variables de paiement considérées, les transactions par carte de débit semblent apporter les améliorations les plus importantes en termes de précision des prévisions.

Domaine	Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Emploi et statistiques du chômage	D'Amuri, F. et Marcucci, J. (2012)	Banque d'Italie/ Recherche économique et relations internationales/ États-Unis - Le pouvoir prédictif des recherches Google dans la prévision du chômage	- Données: Google Data (tests de racine unitaire et transformations) - Méthodologie: ARMA, ARMAX (combinaisons diverses)	Les auteurs constatent que les modèles enrichis avec l'indice de l'intensité de la recherche d'emploi sur Internet (Google Index, GI) dépassent les modèles traditionnels en prédisant le taux de chômage pour différents intervalles hors échantillon qui commencent avant, pendant et après la grande récession. Les modèles basés sur Google dépassent également les modèles standards dans la plupart des prévisions au niveau des États et par rapport à l'Enquête sur les prévisionnistes professionnels.	Ces résultats survivent à un test de falsification et sont également confirmés lorsque différents mots clés sont utilisés.
	Reis, F. (Eurostat), Ferreira, P. (Eurostat), Perduca, V. (Universite Paris Descartes, CNRS) (2015)	Eurostat/France et Italie - L'utilisation de preuve d'activités web pour augmenter l'actualité des indicateurs de statistiques officielles	- Données: mots-clés liés au marché du travail français et italien. - Méthodologie: régression linéaire	Les traces électroniques laissées par les utilisateurs pendant qu'ils utilisent des services Web pourraient être utilisées comme données en temps réel ou avec un très petit décalage. L'application empirique qu'ils mettent en œuvre est une meilleure prévision immédiate du chômage français et italien.	Des articles dans la littérature ont démontré que ces prédictions peuvent être faites. Cependant, ce type de données devrait être davantage contrôlé quant à sa transparence, sa continuité, sa qualité et son potentiel d'intégration avec les méthodes traditionnelles de la statistique officielle.
	Ferreira, P. (Eurostat)	- Eurostat/France et Italie - Amélioration de la prédiction des statistiques du chômage avec Google Trends	- Données: Google Trends - Méthodologie: Régression linéaire, modèles à facteurs dynamiques	Les modèles de prévision du chômage qui utilisent la variable latente estimée ont donné de meilleurs résultats que les approches proposées dans les travaux précédents, en particulier pendant une période où la tendance a été brutalement modifiée.	Ferreira (2015) utilise un modèle à facteurs dynamiques pour extraire une variable latente des données de Google Trends, ce qui est un bon indicateur de la dynamique du chômage.

Domaine	Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Indices des prix et inflation	Breton, R., N. Swier, R. O'Neil (Office for National Statistics (ONS), UK)	- Office national des statistiques/Royaume Unis. - Utilisation de données du web scraping pour construire des indices de prix à la consommation	- Méthodologie : Web Scraping - Données: la recherche couvre la collecte, la manipulation et l'analyse de données extraites sur le Web (Inflation, IPC, IPP).	L'ONS utilise des données extraites sur le Web pour calculer des indices de prix qui: (i) augmentent le nombre d'articles utilisés, (ii) augmentent le nombre de jours considérés et (iii) augmentent à la fois le nombre d'articles et les jours considérés. La construction de ce type d'indices peut être utile aux économistes et aux décideurs.	Les principaux avantages des données Web récupérées sont les suivants: (i) réduction des coûts de collecte, (ii) couverture accrue (c.-à-d. Plus d'éléments du panier), (iii) fréquence accrue, (iv) production des indices nouveaux ou complémentaires, et (v) une meilleure capacité à répondre aux nouveaux défis.
	Griffioen, R., de Haan, J., Willenborg, L. (Statistiques Pays-Bas)	Collecte de données de l'habillement sur Internet/ Pays-bas	- Méthodologie : Web scraping - Données: IPC, prix des vêtements	Les avantages des prix des vêtements raclés sur le Web sont les suivants: (i) la collecte des prix en ligne est moins chère que la collecte des prix dans les magasins physiques, (ii) compte tenu des coûts de collecte relativement bas, il est incité à s'appuyer sur des «données volumineuses» pour contourner les problèmes de petits échantillons (par exemple, forte variance d'échantillonnage), (iii) la qualité des données en ligne a tendance à être très bonne et (iv) certaines caractéristiques des éléments peuvent être facilement observées. Les principaux inconvénients de ce type de collecte de données sont les suivants: (i) les modifications apportées au site Web peuvent entraîner des problèmes de données, (ii) le choix de la stratégie du web scraping peut affecter les informations collectées et la représentativité des articles, (iii) les informations de pondération ne sont pas disponibles, et (iv) les informations disponibles sur les caractéristiques peuvent être insuffisantes, en fonction de la nécessité d'un ajustement de la qualité.	Le document s'intéresse à la possibilité d'utiliser les prix des vêtements en ligne pour l'analyse de l'IPC. Cette étude entre dans la catégorie du web scraping et présente les résultats et les difficultés de la collecte de prix en ligne sur une période de deux ans.
	Jiang Shu	Chine / Bureau national de statistique...	Utilisation des données du Web-scraping pour l'indice de prix du commerce électronique.	Analyser les données de prix des téléphones portables spécifiques par programme Crawler et établir l'indice de prix quotidien comme référence pour les données de prix mensuelles.	Objectif: exploration Zone de statistiques: statistiques de prix.

Domaine	Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Suivi de l'activité économique	Jiang Shu	Chine / Bureau national de statistique	Utilisation des données du Web-scraping pour l'indice de prix du commerce électronique	Analyser les données de prix des téléphones portables spécifiques par programme Crawler et établir l'indice de prix quotidien comme référence pour les données de prix mensuelles.	Objectif: exploration Zone de statistiques: statistiques de prix
	Jiang Shu	Chine / Bureau national de statistique	Enquête sur les cultures par terre agricole: utilisation de la télédétection par satellite et aérienne pour aider à estimer les statistiques agricoles	Construisez le cadre d'échantillonnage spatial en utilisant les données des enquêtes sur l'utilisation des terres et du recensement de l'agriculture. Ensuite, mettez à jour le cadre d'échantillonnage par télédétection par satellite et aérienne. Avec les échantillons sélectionnés par la méthode d'échantillonnage spatial, nous estimons la zone de plantation des cultures et la production chaque saison.	Objectif: projet pilote destiné à passer en production pour remplacer les données existantes Zone de statistiques: statistiques agricoles
	Marc Debusschere	Belgique / Statistiques de Belgique	Etude de faisabilité sur l'utilisation des données de la téléphonie mobile pour les statistiques du tourisme et des transports	Évaluer la possibilité d'utiliser les données de la téléphonie mobile pour compléter, voire remplacer, les sources de données de produits statistiques, principalement dans les domaines du tourisme ou des statistiques de transport. Explorer la possibilité d'enregistrer des phénomènes non encore accessibles par les méthodes traditionnelles.	Objectif: exploration, projet pilote destiné à passer en production pour améliorer la rapidité d'exécution, projet pilote destiné à passer en production pour compléter les données existantes et projet pilote destiné à passer en production pour remplacer les données existantes Zone de statistiques: statistiques du tourisme.
	Dong L. et al. (2017)	- Suivi de l'activité économique/Chine	- Données géolocalisées générées par l'utilisation des smartphones, d'applications de cartographie en ligne et des médias sociaux. - Les auteurs explorent le potentiel d'utilisation des données mobiles pour mesurer l'activité économique en Chine dans une perspective ascendante.	Premièrement, ils ont construit des indices permettant de jauger les tendances de l'emploi et de la consommation à partir de milliards de données de géo-positionnement. Deuxièmement, il ont avancé dans l'estimation du trafic piétonnier dans les magasins hors ligne à l'aide de données de recherche d'emplacement dérivées de Baidu Maps, qui sont ensuite appliquées pour prévoir les revenus d'Apple en Chine et pour détecter avec précision les fraudes au guichet. Troisièmement, ils ont construit des indicateurs de consommation pour suivre les tendances dans divers industries du secteur des services et les vérifier avec plusieurs indicateurs existants.	À notre connaissance, il s'agit de la première étude à mesurer la deuxième plus grande économie du monde en exploitant des données temporelles spatiales d'une taille et d'une granularité sans précédent. De cette manière, cette recherche fournit de nouvelles approches et de nouvelles perspectives pour mesurer l'activité économique.

Domaine	Auteurs	Secteur/Thème/Pays	Méthodes et données	Principaux résultats	Notes/commentaires
Variables financières	Heston, S. L. (Université du Maryland), Sinha, N. R. A. (Conseil des gouverneurs de la Réserve fédérale)	- Nouvelles versus sentiment: comparer des approches de traitement textuel pour prédire des revenus stock	- Données: nouvelles, retours sur stocks - Méthodologie: sentiment de presse, régressions (transversales) - Heston et Sinha (2014) utilisent un ensemble de données de plus de 900 000 articles pour tester si les informations peuvent prédire les rendements des actions.	Ils constatent que les entreprises sans nouvelles ont des rendements futurs moyens nettement différents des entreprises avec des nouvelles. Confirmant les recherches précédentes, les nouvelles quotidiennes prédisent des rendements boursiers pour seulement 1-2 jours. Mais les nouvelles hebdomadaires prédisent des rendements boursiers pour un quart d'année.	Les reportages positifs augmentent rapidement les rendements boursiers, mais les reportages négatifs ont une réaction longtemps retardée.
	Nyman, R. (UCL), Gregory, D. (Banque d'Angleterre), Kapadia, S. (Banque d'Angleterre), Smith, R. (UCL), Tuckett, D. (UCL).	Angleterre /Nouvelles et narratives dans les systèmes financiers : exploitation des Big Data pour l'évaluation du risque systémique	- Données : Rapports de courtiers, Rapports internes de la Banque d'Angleterre, Archives de nouvelles de Reuters - Méthodologie : l'apprentissage automatique et les composantes principales sont inclus dans la méthodologie afin de calculer les indices de consensus.	Nyman et al. (2014) ont étudié les moyens d'utiliser le Big Data dans la gestion du risque systémique. Leurs données de presse comprennent (i) des commentaires quotidiens sur les événements du marché, (ii) des rapports hebdomadaires de recherche économique et (iii) des nouvelles de Reuters. L'apprentissage automatique et les composantes principales sont inclus dans la méthodologie afin de calculer les indices de consensus basés sur les sources ci-dessus.	Leurs conclusions incluent que les rapports de recherche économique hebdomadaires pourraient potentiellement prévoir l'indice du consommateur du Michigan et que des commentaires quotidiens sur les événements de marché pourraient potentiellement prévoir la volatilité du marché.
	Andrade et al. (2009)	Analyse du rôle des analystes et de la diffusion de l'information dans la perspective de la bulle boursière chinoise de 2007.	Corrélation entre les différentes mesures d'intensité de la bulle et sa couverture par les analystes en tant que mesure de la diffusion de l'information. Indice de recherche Google pour vérifier leur timing et leur intensité.	Relation négative significative entre l'intensité de la bulle et la couverture par les analystes. Forte corrélation positive entre l'indice de recherche Google et le volume de nouveaux comptes.	Cette étude est essentiellement liée aux problèmes de prévision.

DEPF Etudes

CONTACT

Adresse

DEPF
Boulevard Mohamed V. Quartier
Administratif,
Rabat-Chellah Maroc

Téléphone

(+212) 5 37.67.74.15/16

Online

Email : depf@depf.finances.gov.ma
Site web: depf.finances.gov.ma