Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Les mégadonnées comme source de données pour les statistiques officielles : expériences menées à Statistique Pays-Bas

Piet J.H. Daas, Marco Puts, Martijn Tennekes et Alex Priem[1]

Résumé

De plus en plus de données sont produites au moyen du nombre croissant de dispositifs électroniques qui nous entourent et que l'on retrouve sur Internet. La grande quantité de données et la fréquence élevée à laquelle elles sont produites ont donné lieu à l'introduction du terme « mégadonnées ». Compte tenu du fait que ces données rendent compte de nombreux aspects différents de nos vies au quotidien, et en raison de leur abondance et de leur disponibilité, les sources de mégadonnées sont très intéressantes du point de vue des statistiques officielles. Toutefois, les premières expériences obtenues suite aux analyses de grandes quantités d'enregistrements de boucles de détection de véhicules au Pays-Bas, d'enregistrements des détails des appels de téléphones mobiles et de messages des médias sociaux aux Pays-Bas révèlent qu'un certain nombre de défis doivent être résolus pour permettre l'application de ces sources de données aux statistiques officielles. Ces défis, ainsi que les leçons apprises pendant les études initiales, seront traitées et illustrées au moyen d'exemples. De façon plus particulière, les sujets suivants sont abordés : les trois types généraux de mégadonnées définis, la nécessité d'accéder à des quantités importantes de données et de les analyser, la façon de traiter les données qui comportent du bruit et d'aborder la sélectivité (ainsi que notre propre biais concernant ce sujet), comment aller au-delà de la corrélation, la façon de trouver les personnes ayant les compétences adéquates et la bonne attitude pour exécuter ce travail, et comment nous avons traité les problèmes de protection des renseignements personnels et de sécurité.

Mots-clés : mégadonnées, statistiques officielles, défis, leçons apprises.

1. Introduction

Dans notre ère numérique moderne, les données touchent à peu près tous les aspects de nos vies, de la façon dont nous magasinons sur le Web, nous nous déplaçons en voiture ou dans les transports en commun, nous cherchons des renseignements sur les produits et nous communiquons avec nos amis et notre famille. Outre cela, des données sur nos allées et venues sont saisies au moyen de caméras, de téléphones mobiles et de réseaux locaux sans fil. Toutes ces données sont entreposées et peuvent potentiellement être exploitées. Toutefois, dans leur forme brute, ces sources de mégadonnées ne sont pas immédiatement utilisables. On doit pouvoir séparer le signal du bruit, c'est-à-dire avoir l'expertise statistique nécessaire pour dériver de l'information à partir d'une quantité considérable de données, afin d'en extraire la signification. Dans ce cas, des connaissances au sujet de l'inférence statistique à partir de mégadonnées est nécessaire (London Workshop, 2014). Il s'agit d'un domaine d'expertise relativement nouveau, le domaine de l'analyse statistique valide des mégadonnées venant à peine de voir le jour (Fan et coll., 2014). Le défi de ce type d'analyses est d'extraire le signal (le cas échéant) correspondant au sujet d'intérêt à partir d'un ensemble de données important et (très) bruyant (Silver, 2010).

Les mégadonnées sont une source très intéressante pour les statistiques officielles (Glasson et coll., 2013), étant donné qu'elles permettent la production considérable possible de chiffres officiels très pertinents, rapidement et à un coût relativement faible. La façon dont on peut y arriver en pratique est un sujet d'intérêt pour de nombreux instituts nationaux de statistique. Un certain nombre de défis ont été identifiés (abordés de façon plus détaillée dans la section 2). Par exemple, de nombreuses sources de mégadonnées sont composées de données d'observation et, par conséquent, n'ont pas une population cible bien définie, manquent souvent de structure et sont de qualité variée. Cela rend difficile l'application de méthodes statistiques traditionnelles, selon la théorie de l'échantillonnage.

[1]Tous les auteurs sont des employés de Statistique Pays-Bas. Personne-ressource : Piet Daas, CBS-weg 11 Heerlen, Pays-Bas, 6412 EX (pjh.daas@cbs.nl). Les points de vue exprimés dans le présent document sont ceux des auteurs et ne reflètent pas nécessairement les politiques de Statistique Pays-Bas.

Toutefois, ce ne sont pas toutes les sources de mégadonnées qui font face aux mêmes problèmes. En étudiant un certain nombre de sources de mégadonnées, par exemple, les données de capteurs routiers, les enregistrements des détails des appels de téléphones mobiles et les messages des médias sociaux, le groupe d'experts en mégadonnées de Statistique Pays-Bas en apprend davantage au sujet de ces sources, ce qui fonctionne et ce qui ne fonctionne pas, et se familiarise sur l'application possible des mégadonnées aux statistiques officielles. Le présent document fournit un aperçu de ces constatations.

2. Défis

Un certain nombre de défis ont été identifiés et doivent être pris en compte lorsque l'on commence à utiliser les mégadonnées pour les statistiques officielles (Daas et Van der Loo, 2013; Glasson et coll., 2013; Struijs et coll., 2014). Un aperçu des principaux défis figure ci-après.

2.1 Accès

Les instituts de statistique ne sont habituellement pas propriétaires des sources de mégadonnées. Un premier défi consiste donc à obtenir l'accès à des sources pertinentes. Cela nécessite des ententes avec les propriétaires des données et les responsables du traitement, qui ont leurs propres préoccupations concernant les coûts, la confidentialité et d'autres problèmes. Toutefois, ils peuvent aussi profiter de cette collaboration avec des organismes statistiques, par exemple, au moyen de la rétroaction en matière de qualité que fournissent les INS. Les modalités doivent être négociées et être acceptables, tant pour les statisticiens officiels que pour les fournisseurs des données.

2.2 Protection de la vie privée

La protection de la vie privée des personnes est impérative, mais les approches habituelles ne fonctionnent pas toujours lorsque l'on traite des mégadonnées. En outre, lorsque la situation juridique n'est pas claire, les statisticiens peuvent devoir se fonder sur des principes éthiques. La perception qu'a le public des utilisations des mégadonnées revêt une importance cruciale : elle a des répercussions directes sur la confiance à l'égard des statistiques officielles. Ces préoccupations ont été accentuées par les révélations selon lesquelles les services de renseignement figurent parmi les utilisateurs les plus actifs des mégadonnées.

2.3 Méthodologie

De nombreuses sources de mégadonnées sont composées de données d'observation axées sur les événements, qui ne sont pas conçues pour l'analyse statistique traditionnelle. Elles n'ont pas de populations cibles bien définies, de structures pour les données et de garanties de qualité. Cela complique la tâche de l'application de méthodes statistiques fondées sur la théorie de l'échantillonnage (Daas et Puts, 2014a). Par exemple, l'évaluation des problèmes de sélectivité pose un défi (Buelens et coll., 2014). Étant donné qu'un nombre croissant de sources de mégadonnées sont fondées sur des textes ou sont composées d'images, la nécessité d'extraire de l'information de ces types de sources de « données » augmente. Cela nécessite de faire appel à des méthodes d'extraction des données, comme l'extraction de textes et les techniques d'apprentissage machine, avec lesquelles les statisticiens officiels ne sont pas encore très familiers, même si elles existent déjà depuis plusieurs années (Fyhrlund et coll., 2005; Saporta, 2000).

2.4 Interprétation

L'extraction de la signification statistique des sources de mégadonnées n'est pas facile. Un « tweet », un appel téléphonique ou une voiture qui passe dans une boucle de détection ont tous un lien avec des personnes, mais la façon d'interpréter ces signaux n'est pas du tout évidente. Par exemple, l'interprétation des données des téléphones mobiles est entravée par de nombreux problèmes : les personnes peuvent avoir avec elles plusieurs téléphones ou pas du tout, les enfants utilisent des téléphones enregistrés au nom de leur parents, les téléphones peuvent être éteints, etc. Dans le cas des messages des médias sociaux, des problèmes similaires peuvent se poser lorsque l'on tente de déterminer les caractéristiques de leurs auteurs. Des solutions comme la détermination du sexe et de l'âge

des utilisateurs de Twitter, à partir de leur choix de mots, semblent faisables (Nguyen et coll., 2013), mais il reste encore beaucoup à faire (Daas et Burger, 2014).

2.5 Technologie

Un défi évident est le traitement, l'entreposage et le transfert de grands ensembles de données. Les progrès technologiques dans le domaine du calcul de haute performance pourraient résoudre en partie ces problèmes. Le traitement des données à la source, en vue d'éviter le transfert de grands ensembles de données et les données entreposées en double, peut aussi être envisagé (Hager et Wellein, 2010). Les défis technologiques comprennent les mécanismes de sécurité, qui font en sorte, par exemple, que les solutions peu coûteuses dans le nuage ne sont pas une option pour les INS.

2.6 Continuité

Habituellement, les statistiques officielles prennent la forme de séries chronologiques. Pour de nombreux utilisateurs, la continuité de ces séries revêt une importance primordiale. De nombreuses sources de mégadonnées, toutefois, n'ont vu le jour que très récemment, sont toujours en évolution et peuvent disparaître aussi rapidement qu'elles sont apparues. Cela présente un risque pour la continuité et oblige à trouver une façon plus souple de travailler.

3. Études à partir de mégadonnées

3.1 Sources

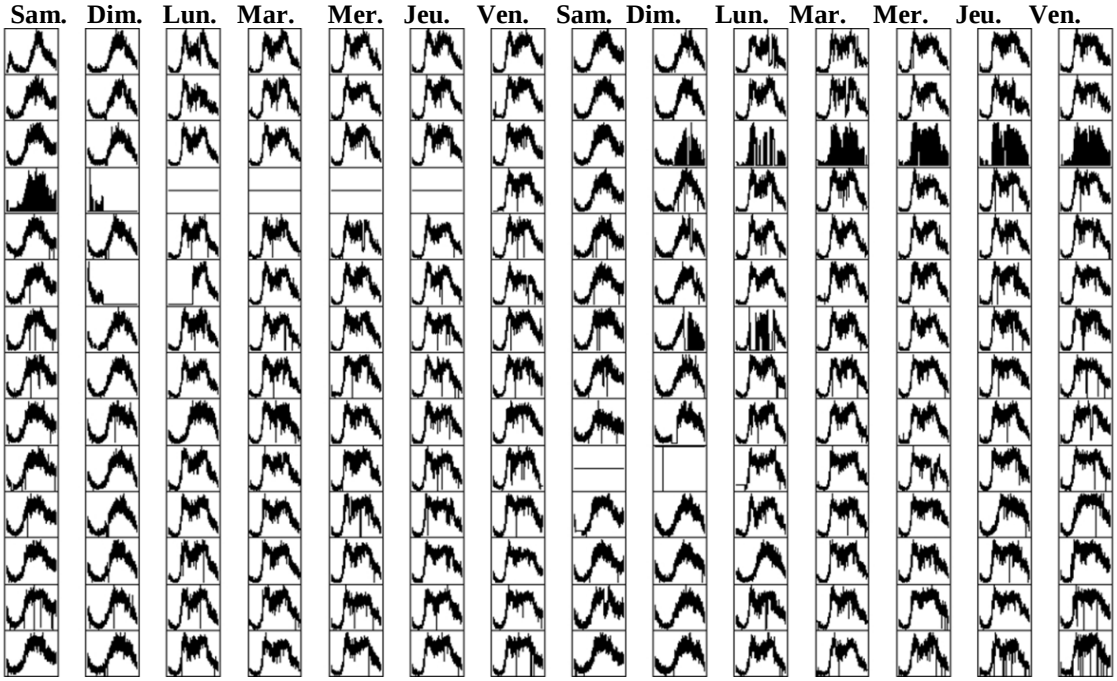
Dans ce chapitre, nous examinons trois exemples typiques de recherches à partir de mégadonnées effectuées à Statistique Pays-Bas. D'autres études connexes à partir de mégadonnées effectuées à notre bureau comprennent des robots Internet, des données de scanneur et des images par satellite. On étudie actuellement d'autres possibilités, comme l'analyse des transactions financières, le premier défi étant souvent d'obtenir l'accès à ces données. La plupart des exemples qui précèdent en sont encore à l'étape de la recherche, mises à part les données de scanneur, qui sont en production depuis maintenant 10 ans. Les robots Internet pour le marché du logement sont sur le point d'entrer en production. Soulignons que les données administratives ne sont habituellement pas considérées comme des mégadonnées, mais que les sources administratives plus volumineuses, comme le registre de la population, les données sur la TVA et les enregistrements des salaires et traitements, pourraient être interprétés de cette façon. L'examen de ces sources plus traditionnelles dans le contexte des mégadonnées pourrait fournir de nouveaux points de vue.

3.2 Données de capteurs routiers

Aux Pays-Bas, il y a plus de 60 000 capteurs routiers, dont 20 000 sont installés sur les autoroutes. Ces capteurs détectent le nombre de véhicules qui passent, selon diverses catégories de longueur, chaque minute. Cela donne lieu à un total de 230 millions d'enregistrements par jour pour les capteurs d'autoroutes seulement. Les données sont recueillies et entreposées par le *National Data Warehouse for Traffic Information* (NDW, www.ndw.nu/en/), un organisme gouvernemental qui fournit les données à Statistique Pays-Bas. Comme les données ne peuvent pas être reliées à des véhicules individuels, les préoccupations en matière de protection de la vie privée ne s'appliquent pas. Cela rend cet ensemble de données attrayant pour l'expérimentation. Le problème le plus important auquel nous avons fait face au moment de l'étude des données de capteurs routiers était le fait que la qualité de données fluctue considérablement. Pour certains capteurs, on ne dispose pas de données pour de nombreuses minutes et, en raison de la nature stochastique des moments d'arrivée des véhicules au capteur routier, il est difficile de calculer directement le nombre de véhicules manquants pendant ces minutes. À cette fin, un filtre d'adaptation a été élaboré et correspond au comportement stochastique des moments d'arrivée des véhicules au capteur (Puts et coll., 2014). La qualité des données varie non seulement par minute, mais aussi par jour (fig. 3.2-1). En corrigeant les données pour tenir compte des données manquantes et en combinant les profils quotidiens fournis par les capteurs sur les mêmes sections de routes, on améliore la couverture et la qualité des données. De cette façon, nous sommes en mesure de produire des indices sur la circulation qui décrivent la situation régionale, au niveau NUTS-3, sur les routes des

Pays-Bas. La combinaison de ces résultats régionaux donne un très bon aperçu de l'état de la circulation routière au pays (Daas et coll., 2014).

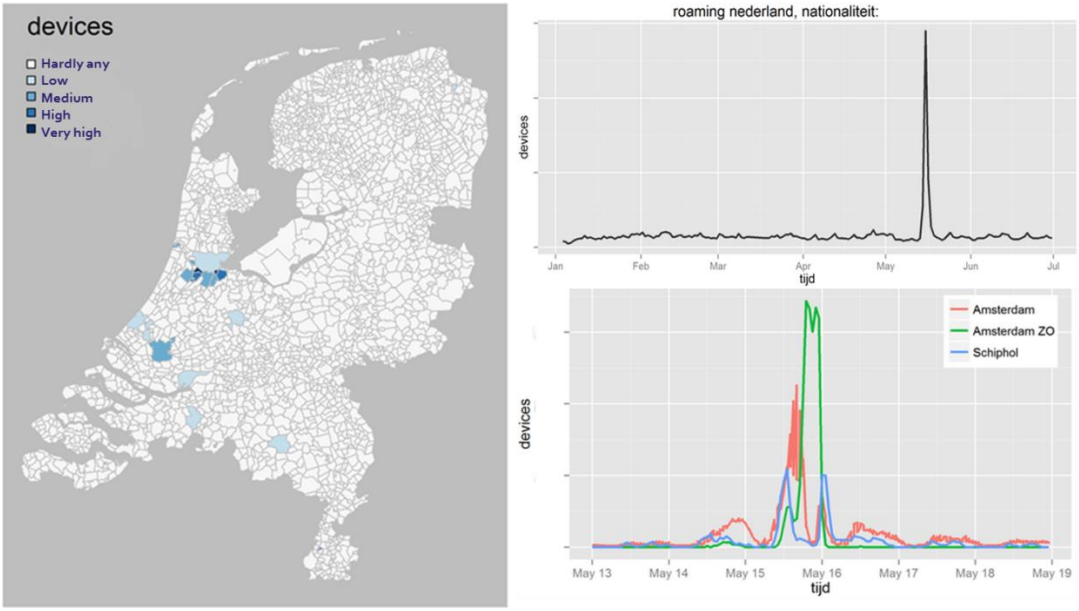
Figure 3.2-1
Profils quotidiens d'un capteur routier sur la digue de l'IJsselmeer (« Afsluitdijk ») pendant 196 jours consécutifs.



3.3 Données de téléphones mobiles

De nos jours, les personnes ont leur téléphone mobile sur elles partout et les utilisent souvent pendant toute la journée. Pour gérer le volume d'appels téléphoniques, une somme considérable de données doit être traitée par les compagnies de téléphones mobiles. Ces données sont très étroitement liées au comportement des personnes; des comportements qui sont intéressants pour les statistiques officielles. Par exemple, le volume est relayé par des antennes téléphoniques réparties au niveau géographique, qui permettent de déterminer où se trouvent les utilisateurs de téléphones. Les antennes de relais, toutefois, peuvent changer plusieurs fois pendant un appel. Grâce à un contrat avec un tiers, Statistique Pays-Bas a eu accès aux données des enregistrements détaillés des appels (EDA) d'une compagnie de téléphones mobiles des Pays-Bas, dont la part du marché représente environ le tiers du marché de la téléphonie mobile aux Pays-Bas. Les données des EDA totalisent 115 millions d'enregistrements par jour et comprennent des renseignements sur les utilisateurs néerlandais et les utilisateurs en itinérance du réseau. Les microdonnées rendues anonymes des EDA ont été traitées par une compagnie intermédiaire spécialisée, selon les spécifications demandées par Statistique Pays-Bas. Seuls les résultats agrégés ont été envoyés à Statistique Pays-Bas, conformément aux dispositions de protection de la vie privée. Plusieurs utilisations ont été étudiées pour les statistiques officielles, y compris le tourisme récepteur (Heerschap et coll., 2014) et la population de jour (Tennekes et Offermans, 2014). Dans la figure 3.3-1, on présente un exemple de données d'EDA appliquées au « tourisme » récepteur. Dans cette figure, l'activité des téléphones mobiles est associée à un des pays d'Europe participant à la finale de l'Europa League, le 15 mai 2013, au stade d'Amsterdam. Le résultat le plus frappant est le fait que l'activité des téléphones mobiles de ce pays particulier autour de cette date est beaucoup plus élevée que l'activité dans le reste de la période étudiée. Cela rend bien compte des touristes qui visitent notre pays pour un événement particulier au cours d'une très courte période (Heerschap et coll., 2014). Il est très probable que la majorité de ces visiteurs ne sont pas inclus dans les statistiques officielles sur le tourisme fondées sur l'hébergement.

Figure 3.3-1
Activité des téléphones mobiles enregistrée dans un pays d'Europe autour de la période de la finale de l'Europa League de l'UEFA au stade d'Amsterdam, en 2013. L'activité relative dans toutes les régions des Pays-Bas et dans des régions particulières sont présentées, y compris l'activité globale aux Pays-Bas pendant la première moitié de l'année.



La figure 3.3-2 comprend un exemple d'études sur la population de jour. Les « allées et venues de jour » est un sujet dont on sait très peu de choses jusqu'à maintenant en raison du manque de sources, contrairement à la « population de nuit », qui est fondée sur les registres officiels (de résidence). La figure montre la population de jour à midi, un lundi de mai, dans les cinq plus grandes municipalités des Pays-Bas. Des résultats pour une municipalité de travailleurs typique (Haarlemmermeer, où l'aéroport Schiphol est situé) et une municipalité de navetteurs typique (« Almere ») sont aussi inclus. Les couleurs indiquent le nombre de personnes qui partent, demeurent ou arrivent. Les points indiquent la population officiellement enregistrée. Ce travail est un exemple d'une nouvelle statistique possible fondée sur les mégadonnées qu'un INS pourrait produire (Tennekes et Offermans, 2014).

3.4 Messages des médias sociaux

Plus de trois millions de messages de médias sociaux publics sont produits sur une base quotidienne aux Pays-Bas. Ces messages sont disponibles pour quiconque ayant accès à Internet. Les médias sociaux sont une source de données où des personnes partagent volontairement de l'information, discutent de sujets qui les intéressent et communiquent avec leur famille et leurs amis. Afin de déterminer si les médias sociaux représentent une source de données intéressante pour les statistiques, les messages des médias sociaux des Pays-Bas ont été étudiés de deux points de vue : le contenu et les sentiments. Les données de source des médias sociaux ont été fournies par la compagnie Coosto (www.coosto.com/uk/), qui recueille systématiquement tous les messages des médias sociaux des Pays-Bas et attribue des scores de sentiments, notamment. Des études du contenu des messages sur Twitter aux Pays-Bas (la principale source de messages dans les médias sociaux publics aux Pays-Bas à ce moment-là) ont révélé que près de 50 % de ces messages étaient constitués de « babillages inutiles ». Les autres portaient principalement sur les activités pendant les temps libres (10 %), le travail (7 %), les médias (5 %) et la politique (3 %). Les babillages moins sérieux nuisaient à l'utilisation de ces messages plus sérieux (Daas et coll., 2012). Ces babillages ont aussi eu des répercussions négatives sur les études d'extraction de texte. Les études des données de Coosto ont révélé que les sentiments dans les messages des médias sociaux des Pays-Bas étaient fortement corrélés avec la confiance des consommateurs (fig. 3.4-1). Ce phénomène est principalement affecté par les variations dans

les sentiments de tous les messages Facebook publics aux Pays-Bas (r = 0,85). L'inclusion de diverses sélections de messages publics sur Twitter a amélioré cette association et la réaction aux changements dans les sentiments (r = 0,89). Les sentiments observés étaient stables sur une base mensuelle et hebdomadaire, mais les chiffres quotidiens affichaient un comportement très volatil.

Figure 3.3-2
Population de jour estimée à midi, un lundi de mai, dans les cinq plus grandes municipalités des Pays-Bas.

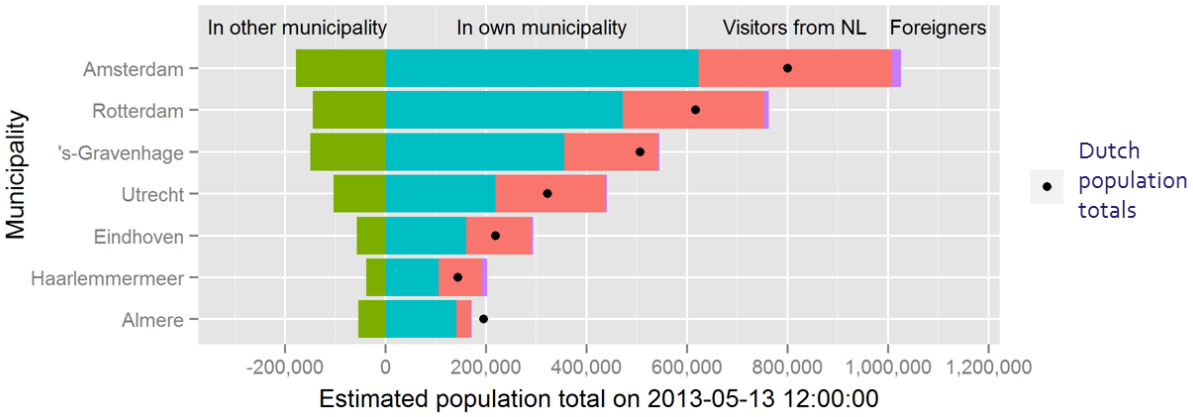
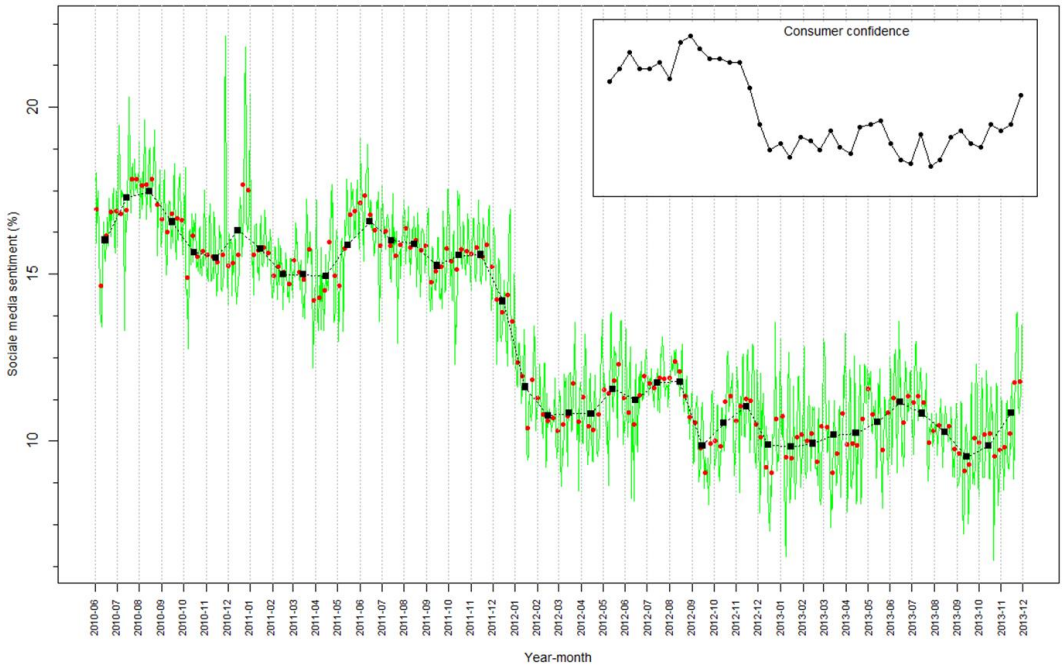


Figure 3.4-1
Élaboration d'agrégations quotidiennes, hebdomadaires et mensuelles des sentiments dans les médias sociaux, de juin 2010 à novembre 2013, en vert, rouge et noir, respectivement. Dans l'insertion, on montre l'évolution de la confiance des consommateurs pour la même période.



Ainsi, il peut devenir possible de produire des indicateurs hebdomadaires utiles des sentiments, même le premier jour ouvrable après la semaine étudiée. Les études de causalité de Granger ont révélé qu'il est plus probable que les changements dans la confiance des consommateurs précèdent ceux dans les sentiments dans les médias sociaux que le contraire. Une comparaison de l'évolution de divers agrégats de sentiments sur sept jours et des séries mensuelles

sur la confiance des consommateurs a confirmé ce résultat et a révélé que le décalage qui touche les sentiments dans les médias sociaux est le plus probablement de l'ordre de sept jours. Ces résultats de la recherche et d'autres sont conformes à la notion selon laquelle les changements dans la confiance des consommateurs et les sentiments dans les médias sociaux sont affectés par un phénomène sous-jacent identique. Une explication de ce phénomène se trouve dans Appraisal-Tendency Framework (Han et coll., 2007), qui porte sur la prise de décisions par les consommateurs. Dans ce cadre, on déclare qu'une décision de consommation est influencée par deux types d'émotions, à savoir accessoire et intégrale. Dans ce cadre, l'émotion intégrale est pertinente pour la décision en question, tandis que l'émotion accessoire ne l'est pas. Selon cette théorie, la confiance des consommateurs est probablement influencée principalement par l'émotion accessoire, étant donné qu'elle n'est pas mesurée non plus en rapport avec la décision proprement dite d'acheter quelque chose. Cela laisse supposer que les sentiments dans les messages des médias sociaux peuvent être le reflet de l'émotion accessoire de la partie de la population qui est active dans ces médias sociaux. Étant donné la nature générale de cette dernière affirmation, on pourrait appeler cela l'« humeur » de la nation dans le contexte de la prise de décisions en matière de consommation. Des résultats plus détaillés de cette étude sont décrits dans Daas et Puts (2014b).

4. Conclusions

L'avènement des mégadonnées présente de nouvelles possibilités pour les statistiques officielles, mais pose aussi de nouveaux défis. Des défis importants pour les statistiques officielles sont les suivants : i) tenir compte de la sélectivité des mégadonnées, ii) éditer les données à une grande échelle, et iii) réduire le volume de données sans perdre (trop) d'information. Étant donné qu'il existe différents types de sources de mégadonnées, p. ex., produites par des humains, produites par un capteur et fondées sur des transactions, chaque source devrait être étudiée et jugée selon ses propres mérites. Le plus important est d'axer les travaux sur les données, en laissant de côté la perspective axée sur un échantillon qui est si familière aux statisticiens (Daas et Puts, 2014a). Une approche axée sur les données peut empêcher que des constatations intéressantes soient jugées contre le mauvais cadre. À l'heure actuelle, le domaine de la recherche à partir de mégadonnées est tout nouveau, et on dispose d'une expérience limitée seulement pour estimer les répercussions des mégadonnées sur les statistiques officielles. Dans une certaine mesure, on pourrait faire un parallèle avec l'avènement de l'échantillonnage d'enquête (Bethlehem, 2009). Au moment où les statistiques officielles ont pris forme, au milieu du XIX^{e} siècle, au début seules les approches de dénombrement de type recensement étaient considérées comme valables. Autour de 1895, les premières idées ont été formulées pour les statistiques fondées sur un échantillon, mais il a fallu plusieurs décennies avant que le paradigme désormais dominant de l'échantillonnage d'enquête soit accepté et bien établi. Dans l'ensemble, les défis énoncés ci-dessus devront être pris en compte. De façon plus particulière, il faut de nouvelles dispositions législatives, ainsi que des statisticiens ayant de nouvelles compétences et un nouvel état d'esprit (« des experts des données »), de nouvelles méthodes et des installations de calcul appropriées (London Workshop, 2014; ASA-Working group, 2014). Les travaux profiteraient aussi d'une collaboration internationale intensifiée entre les fournisseurs de données, les scientifiques et les statisticiens officiels.

Remerciements

Les auteurs souhaitent exprimer leur reconnaissance à leurs collègues de Statistique Pays-Bas, soit Edwin de Jonge, Joep Burger, Bart Buelens, Jan van den Brakel, May Offermans, Barteld Braaksma et Peter Struijs, pour avoir stimulé les discussions et fourni des remarques constructives. Ces travaux n'auraient pas été possibles sans le soutien du programme de l'innovation de Statistique Pays-Bas.

Bibliographie

ASA-working group (2014). *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society*. Rapport d'un groupe de travail de l'American Statistical Association. http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf.

BETHLEHEM, J. G. (2009). *The rise of survey sampling*. Document de travail 09015 de Statistique Pays-Bas. Statistique Pays-Bas, La Haye/Heerlen, Pays-Bas.

BUELENS, B., P. DAAS, J. BURGER, M. PUTS et J. VAN DEN BRAKEL (2014). *Selectivity of Big Data*. Document de travail 201411 de Statistique Pays-Bas, La Haye/Heerlen, Pays-Bas.

DAAS, P. J. H. et J. BURGER (2014). *Profiling Big Data sources to assess their selectivity*. Résumé pour la New Techniques and Technologies for Statistics Conference 2015, Bruxelles, Belgique.

DAAS, P. J. H. et M.J.H. PUTS (2014a). « Big Data as a source of statistical information ». *The Survey Statistician* 69: 22 à 31.

DAAS, P. J. H. et M.J.H. PUTS (2014b). *Social Media Sentiment and Consumer Confidence*. European Central Bank Statistics Paper Series 5, Francfort, Allemagne.

DAAS, P., M. PUTS, S. OSSEN et M. TENNEKES (2014). *Processing and methods for Big Data: a traffic index based on huge amounts of road sensor data*. Document présenté à la Conference of European Statistics Stakeholders, Rome, Italie.

DAAS, P. J. H., M. ROOS, M. VAN DE VEN et J. NERONI (2012). *Twitter as a potential data source for statistic*". Document de travail 201221 de Statistique Pays-Bas, La Haye/Heerlen, Pays-Bas.

DAAS, P. J. H. et M.P.J. VAN DER LOO (2013). *Big Data (and official statistics)*. Document au 2013 Meeting on the Management of Statistical Information Systems, Paris, France, Bangkok, Thaïlande. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf.

FAN, J., F. Han et H. LIU (2014). « Challenges of Big data analysis ». *National Science Review* 1: 293 à 314.

FYHRLUND, A., B. FRIDLUND et B. SUNDGREN (2005). "
« Using Text Mining in Official Statistics ». *Knowledge Mining, Proceedings of the NEMIS 2004 Final Conference, Studies in Fuzziness and Soft Computing 185*: 201 à 211.

GLASSON, M., J. TREPANIER, V. PATRUNO, P. DAAS, M. SKALIOTIS et A. KHAN (2013). *What does "Big Data" mean for Official Statistics?*. Document pour le High-Level Group for the Modernization of Statistical Production and Services. http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614.

HAGER, G. et G. WELLEIN (2010). *Introduction to High Performance Computing for Scientists and Engineers.* Boca Raton: Chapman & Hall/CRC Computational Science.

HAN, S., J.S. LERNER et D. KELTNER (2007). « Feelings and consumer decision making: the appraisal-tendency framework ». *Journal of Consumer Psychology* 17: 158 à 168.

HEERSCHAP, N. M., S.A. ORTEGA AZURDUY, A.H. PRIEM et M.P.W. OFFERMANS (2014). *Innovation of tourism statistics through the use of new Big Data sources*. Document présenté au Global Forum on Tourism Statistics, Prague.
http://www.tsf2014prague.cz/assets/downloads/Paper%201.2_Nicolaes%20Heerschap_NL.pdf.

London Workshop (2014). *Statistics and Science*. Rapport sur le London Workshop on the Future of the Statistical Sciences. http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf.

NGUYEN, D-P., R. GRAVEL, R.B. TRIESCHNIGG et T. MEDER (2013). TweetGenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter* 4: 4 à 9.

PUTS, M., M. TENNEKES et P. DAAS (2014). *Using Road Sensor Data for Official Statistics: Towards a Big Data Methodology*. Document présenté à la Strata 2014 Conference, Barcelone, Espagne.

SAPORTA, G. (2000). *Data Mining and Official Statistics*. Document présenté à la Quinta Conferenza Nationale di Statistica, Rome, Italie.

SILVER, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail —but Some Don't*. New York: Penguin Group.

STRUIJS, P., B. BRAAKSMA et P. DAAS (2014). « Official Statistics and Big Data ». *Big Data & Society*, *April–June:* 1 à 6.

TENNEKES, M. et M.P.W. OFFERMANS (2014). *Daytime population estimations based on mobile phone metadata*. Document présenté aux Joint Statistical Meetings 2014. Boston, États-Unis.