Nations Unies ECE/CES/2019/23

Conseil économique et social

Distr. générale
9 avril 2019
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante-septième réunion plénière
Paris, 26-28 juin 2019
Point 2 b) de l'ordre du jour provisoire
Nouvelles sources de données − accessibilité et utilisation
Deuxième séance : aptitudes requises pour exploiter de nouvelles sources de données

Nouvelles compétences requises pour travailler avec les mégadonnées − enseignements tirés jusqu'à présent

Note du Bureau de statistique de la Slovénie

Résumé

Le présent document décrit les enseignements tirés au cours des six dernières années par le Bureau de statistique de la Slovénie en ce qui concerne les compétences requises pour travailler avec les mégadonnées. Il commence par dresser la liste des compétences acquises, les plus importantes étant celles qui sont liées aux caractéristiques des nouvelles sources de données, aux méthodes avancées de traitement des données et à la communication avec les détenteurs de données. Il aborde ensuite les modes d'acquisition des compétences, l'apprentissage par l'action étant le plus important, puis l'échange de données d'expérience au sein de la communauté statistique internationale. Il se termine par un regard vers l'avenir et mentionne le maintien des compétences comme étant la tâche la plus difficile.

Ce document est présenté pour examen à la deuxième séance (« Aptitudes requises pour exploiter de nouvelles sources de données ») du séminaire organisé en 2019 dans le cadre de la Conférence des statisticiens européens sur le thème « Nouvelles sources de données – accessibilité et utilisation ».

GE.19-05941 (F) 150419 080519
1905941

I. Introduction

1. Les mégadonnées et leur utilisation ont été l'un des sujets les plus brûlants de la statistique officielle ces dernières années. Le Bureau de statistique de la Slovénie (SURS) s'emploie depuis plus de six ans à recueillir, traiter et analyser les sources de mégadonnées. Au cours de cette période, un projet a été mené à bien grâce à l'inclusion d'une source de mégadonnées, à savoir les données scannées concernant les prix, dans la production statistique courante. D'autres projets consistent notamment à analyser le potentiel que représentent les données sur la localisation et les transactions provenant des téléphones mobiles, les offres d'emploi en ligne, les prix en ligne, les images satellitaires, les données relatives aux transactions financières et les données provenant des capteurs de trafic routier. Tous ces projets présentent les mêmes difficultés, par exemple l'accès, les considérations éthiques, la méthodologie, la technologie et la qualité. Cependant, il existe aussi une difficulté qui semble être la plus horizontale car elle concerne tous les projets et toutes les phases statistiques, à savoir les capacités et les connaissances nécessaires.

2. Le présent document détaille les enseignements tirés par le SURS en ce qui a trait aux capacités et aux connaissances requises pour mener à bien des projets relatifs aux mégadonnées. Il dresse tout d'abord la liste des nouvelles connaissances ou capacités qui ont été acquises, puis s'intéresse à la manière dont cette acquisition s'est produite et conclut en envisageant l'avenir. Dans ce document, le terme « compétences » est utilisé pour désigner les connaissances et les capacités.

II. Ce qui a été appris

3. En mettant en œuvre des projets relatifs aux mégadonnées, le SURS a fait naître de nouvelles compétences et a amélioré ou renforcé les compétences existantes. Certaines compétences étaient nouvelles pour certains membres du personnel et dans certains domaines de la statistique, mais pas pour d'autres membres ou dans d'autres domaines ; par exemple, certains statisticiens spécialisés ont commencé à utiliser des outils informatiques plus avancés pendant les projets, tandis que de leur côté, certains informaticiens ont approfondi leurs connaissances méthodologiques statistiques. La diffusion des compétences déjà existantes au sein du Bureau a été un effet important du travail réalisé. Par ailleurs, de nombreuses compétences ne sont pas nouvelles pour le « monde extérieur », mais le sont pour le SURS ; elles existent depuis plusieurs décennies, mais ce n'est que récemment qu'elles ont commencé à occuper une place plus importante dans les statistiques officielles slovènes (par exemple, la modélisation des données). Les compétences nouvelles ou améliorées les plus importantes acquises par le SURS au fil de ses travaux sur les mégadonnées sont les suivantes :

a) Plus de sources de données sont connues. De nouvelles sources ont été reconnues et leurs caractéristiques étudiées (propriétaires/détenteurs, fréquence, qualité, contenu, accessibilité). Il s'agit par exemple des informations générées par les détaillants (prix, quantités, codes des produits vendus), par l'utilisation du téléphone mobile (lieux, transactions), par les systèmes d'observation de la Terre, par les dispositifs de comptage des véhicules, etc. ;

b) On connaît mieux la législation applicable aux différents types de données, principalement la législation relative à la protection des données à caractère personnel et la législation relative aux communications électroniques ;

c) Davantage de modes de collecte des données sont utilisés. Par exemple, les données scannées concernant les prix sont collectées à partir des fichiers de données transmis directement par les entreprises. Dans le passé, une telle méthode n'était employée que pour collecter des données auprès de l'administration publique, mais pas auprès des entreprises. En outre, des techniques d'extraction de données sur Internet sont désormais utilisées pour recueillir des données sur les prix dans les boutiques en ligne et sur les offres d'emploi sur les portails Web dédiés à l'emploi. Pour cela, il a été nécessaire d'acquérir des connaissances sur la composition des sites Internet, le fonctionnement des navigateurs, la

programmation des chercheurs et explorateurs Web spécifiques et généraux et l'optimisation de la recherche par mot-clé selon la structure HTML ;

d) De nouvelles méthodes de travail avec les fournisseurs de données ont été mises au point. Traditionnellement, les entreprises étaient censées fournir des données conformément à la forme et au contenu des données prescrits. Désormais, avec les données scannées concernant les prix, chaque entreprise fournit des données d'une manière et sous une forme qui sont adaptées au mieux à ses possibilités et à son mode de fonctionnement. Le SURS a ainsi appris à être encore plus convivial avec les fournisseurs de données ;

e) La connaissance des entreprises s'est améliorée : quel est leur modèle commercial, quelles sont les questions qui peuvent leur poser problème (révéler des secrets d'affaires ou les données des clients, etc.), comment elles perçoivent l'administration publique, quel est leur environnement commercial et législatif, comment elles fonctionnent, quels sont leurs centres d'intérêt ;

f) Le changement du paradigme fondamental de la production de statistiques a été reconnu et progressivement accepté. Traditionnellement, le travail du statisticien commence par les résultats attendus. Une fois que les besoins des utilisateurs sont connus et les résultats définis, l'examen des sources de données et des meilleurs moyens de collecter et traiter les données commence (des résultats attendus aux moyens mis en œuvre). Dans le cas des mégadonnées, c'est généralement l'approche inverse qui prévaut. D'abord, les sources de données d'entrée sont disponibles, et ce n'est qu'ensuite que le statisticien commence à analyser si ces sources de données ont une valeur statistique (des moyens aux résultats). Avec cette nouvelle façon d'aborder la production de statistiques, le SURS dispose de compétences accrues pour sélectionner et intégrer les données les plus « parlantes » et pour décider de l'orientation des recherches ;

g) Il est plus courant de voir les choses dans une large perspective. Avec les sources de mégadonnées, le SURS apprend à faire preuve d'ouverture d'esprit lorsqu'il envisage les utilisations possibles. Par exemple, la question des utilisations possibles des images satellitaires dans la production de statistiques a fait l'objet d'un débat approfondi au sein du SURS, avec une large participation du personnel invité à présenter ses propositions ;

h) De nouvelles méthodes ont été maîtrisées. Avec les données scannées concernant les prix, la méthode de calcul de l'indice des prix à la consommation pour les aliments et les boissons à partir de produits représentatifs a été abandonnée au profit d'une prise en compte de l'ensemble des produits. On a ainsi appris à calculer également de nouveaux types d'indices ;

i) Les pratiques d'intégration des données sont mieux connues. L'intégration de données provenant de différentes sources n'est pas nouvelle pour le SURS puisque les dossiers administratifs sont déjà utilisés depuis plusieurs décennies. La nouveauté consiste à intégrer les mégadonnées et les données directement collectées ;

j) Les méthodes d'optimisation des programmes permettant de traiter de grandes quantités de données et d'utiliser intensivement les processeurs sont maîtrisées. Par exemple, de nombreux efforts ont été consacrés à la compréhension et à l'élaboration de programmes multifils, qui permettent d'accélérer la collecte, l'analyse, le traitement et l'agrégation des données. Des méthodes fondées sur ce principe sont employées pour analyser, modifier, transformer et agréger les données des dispositifs de comptage des véhicules ; en raison de leur taille (plusieurs dizaines de gigaoctets), les boucles standard prendraient beaucoup plus de temps. Les programmes multifils sont également utilisés pour l'extraction et le stockage des contenus de plusieurs centaines de milliers de sites Web ;

k) La fouille de textes et l'apprentissage machine, méthodes employées pour structurer des données non structurées et les classer avec différents algorithmes, sont de mieux en mieux connus. Ces méthodes sont utilisées à titre expérimental pour classer les produits de consommation selon la nomenclature COICOP, pour apparier les unités économiques et pour structurer les offres d'emploi publiées sur Internet ;

l) La modélisation des données couplée à l'apprentissage machine pour la prévision et la compilation des statistiques est une méthode beaucoup mieux connue. Cela s'accompagne d'une amélioration des capacités de compréhension et d'élaboration de

modèles, ainsi que de compréhension des limites et des implications de la combinaison de modèles. Par exemple, la prévision des agrégats économiques par régression linéaire à l'aide des données des dispositifs de comptage des véhicules et du chiffre d'affaires a été introduite. La méthode de compilation des offres d'emploi à partir de données non structurées par régression logistique, régression linéaire, méthode du plus proche voisin et méthode des ensembles AdaBoost a été mise au point ;

m) De nouveaux outils et bibliothèques de programmation sont utilisés et le recours aux outils existants est plus fréquent : technologies (.NET, ASP.NET, ML.NET frameworks), outils (MS Visual Studio, SQL Server Data Tools − SQL Server Integration Services), langages (C#, R, Python), bases de données (MS SQL Server, Oracle), bibliothèques Python (TensorFlow, Scikit-Learn, Keras), etc. ;

n) En raison du haut niveau d'instabilité des outils et des modèles économiques des fournisseurs de ces outils (dont certains, par exemple, après avoir été disponibles en accès libre pendant un certain temps, ont soudainement été remplacés par des versions entièrement commerciales, ou, alors qu'ils étaient disponibles hier, ne le sont plus aujourd'hui), le SURS a appris à être plus flexible et adaptable à ces changements, afin que ceux-ci ne causent pas de perturbations majeures dans son travail ;

o) On connaît mieux la coordination et le lien entre toutes les compétences susmentionnées, de même qu'il est davantage reconnu qu'il est plus utile pour le personnel de posséder plusieurs compétences différentes (codage, mathématiques, méthodes statistiques, par exemple).

III. Comment cela a été appris

4. D'un point de vue organisationnel, le SURS a abordé les mégadonnées de la même manière que les autres travaux de développement, qui sont considérés comme faisant partie intégrante du travail de chaque statisticien. Il n'y a pas de service spécial qui s'occuperait uniquement de l'innovation, du progrès, des améliorations, etc. Le SURS estime que la séparation des activités de développement et des tâches ordinaires conduit à une ségrégation, à du cloisonnement, à une méconnaissance des questions pratiques par les « statisticiens de développement » et du progrès par les « statisticiens ordinaires », et à des problèmes supplémentaires de mise en œuvre des changements. Il en va de même du travail avec les mégadonnées. Ce travail est maintenant organisé dans le cadre de projets établis formellement avec la participation d'employés de différents services administratifs.

5. Plus précisément, les compétences nécessaires pour travailler avec les mégadonnées ont été acquises (ou sont encore en cours d'acquisition) de différentes manières, à savoir :

a) L'apprentissage par la pratique. La réalisation de projets et de travaux pratiques est de loin le moyen le plus important d'acquérir des compétences. Du personnel de différents services (spécialistes des méthodes statistiques, informaticiens, mathématiciens) fait partie des équipes de projet et une bonne coopération entre les membres de ces équipes est essentielle pour que puisse avoir lieu un échange efficace d'idées, de connaissances, etc. De plus, certains statisticiens participent à plusieurs projets, ce qui est utile pour transférer des solutions d'un projet à l'autre ;

b) La participation à des activités internationales : Équipe spéciale des mégadonnées de la CEE, Équipe spéciale du SSE (Système statistique européen) sur les mégadonnées, ESSNet BigData1 (extraction d'offres d'emploi sur Internet, premières estimations, méthodes), ESSnet BigData2 (offres d'emploi, transactions financières), Groupe d'experts du SSE sur les données scannées ;

c) Les visites d'étude dans des bureaux de statistique ayant plus d'expérience dans l'utilisation des sources de mégadonnées, et l'étude des documents et des pratiques de ces organismes. Ces visites et cette étude ont particulièrement servi dans le domaine de l'utilisation des données scannées ;

d) La participation physique ou virtuelle (séminaires, cours ou ateliers en ligne) à des conférences, ateliers, cours et autres activités internationales portant notamment sur

les thèmes suivants : les mégadonnées : traitement et analyse efficaces de données très volumineuses et non structurées à des fins de statistique officielle ; l'utilisation de R dans la statistique officielle ; un statisticien peut-il devenir un scientifique des données ? ; introduction aux mégadonnées et à leurs outils ; les sources de mégadonnées − Internet, les médias sociaux et l'analyse de textes ; les sources modernes de mégadonnées : téléphone mobile et autres capteurs ; la collecte automatique des prix en ligne ; la prévision macroéconomique ; science des données : la régression linéaire ; l'analyse des données symboliques ; atelier sur l'analyse des données à des fins de statistique officielle ; apprentissage machine et économétrie ; Conférence mondiale de l'ONU sur les mégadonnées, etc. ;

e) L'auto-apprentissage à partir des ressources disponibles sur Internet (manuels, guides, vidéos, plateformes de développement de logiciels, forums, par exemple) ;

f) Assister à des conférences universitaires. Le personnel du SURS a assisté, dans le cadre d'études à temps plein à la Faculté d'informatique et des sciences de l'information, à plusieurs conférences sur le traitement des mégadonnées ;

g) La coopération avec les instituts. Le SURS coopère principalement avec l'Institut « Jožef Stefan », en particulier dans le domaine des algorithmes avancés pour la classification des textes par langue ;

h) Les connaissances sont également acquises pendant les études, c'est-à-dire avant qu'une personne n'intègre le SURS. Actuellement, les études les mieux adaptées aux besoins sont les études en informatique, sciences de l'information, mathématiques, physique, sciences sociales et statistique. Lorsqu'une personne possède une connaissance et une compréhension de base de la programmation, de la programmation par objets et des structures de données, les connaissances en apprentissage machine, en extraction de données et en fouille de textes peuvent être acquises relativement vite, au moins sous une forme élémentaire.

6. L'un des enseignements les plus importants tirés au sujet de la formation est qu'il ne sert à rien d'accumuler des connaissances juste au cas où quelqu'un en aurait besoin. Auparavant, de nombreux membres du personnel suivaient des cours sur les mégadonnées, mais n'utilisaient pas les connaissances acquises durant ces cours ou quittaient le Bureau avant d'avoir participé aux projets relatifs aux mégadonnées. Chacun peut désormais acquérir des connaissances de l'une ou l'autre des façons susmentionnées lorsqu'il est chargé d'accomplir une tâche pour laquelle ces connaissances sont nécessaires.

IV. Perspectives d'avenir

7. En ce qui concerne l'avenir des compétences, trois points valent d'être mentionnés.

8. Premièrement, les compétences manquantes : Le besoin le plus urgent est d'améliorer les compétences dans les domaines complexes de la fouille de textes, de la compréhension de l'apprentissage machine, de la modélisation des données et de l'intégration des mégadonnées. De plus, il est aujourd'hui nécessaire d'acquérir des compétences en matière de traitement des données par informatique répartie en groupes, réseaux neuronaux, arbres de décision et forêts aléatoires. Ces compétences s'acquerront de toutes les façons énumérées plus haut.

9. Deuxièmement, la répartition des compétences : Jusqu'à présent, seul un nombre limité d'employés ont été associés aux travaux sur les mégadonnées et acquièrent des compétences dans ce domaine. À l'avenir, il sera nécessaire de veiller à ce que les compétences soient aussi réparties entre les autres membres du personnel. Cela deviendra plus facile à mesure que de nouvelles sources de données seront intégrées avec succès dans la production statistique courante ; autrement dit, chaque projet réussi rendra les étapes suivantes plus faciles. Par conséquent, pour les gestionnaires, l'un des plus grands défis sera de savoir comment motiver le personnel pour qu'il s'adapte aux changements que la numérisation apporte à la statistique officielle.

10. À ce sujet, il convient également de mentionner les progrès très rapides des sciences en question. Pour quiconque, il est déjà difficile de bien connaître toutes les branches de l'apprentissage machine et de la fouille de textes. Dans le secteur privé, un processus de spécialisation de ces professions a déjà démarré. Néanmoins, pour un expert, il reste nécessaire de bien connaître ne serait-ce que les bases de tous les éléments et de suivre l'évolution récente. Il faut donc s'attendre à ce que, dans le domaine de la statistique officielle aussi, les scientifiques des données se spécialisent davantage. Le nombre d'experts nécessaires pour couvrir tous les éléments du travail sur les mégadonnées s'en trouvera encore augmenté.

11. Troisièmement, le maintien des compétences : En raison de la forte demande de scientifiques des données et, par conséquent, du niveau élevé des rémunérations dans le secteur privé, il est très difficile de retenir le personnel qui possède les compétences requises. Au moins à moyen terme, conserver le personnel sera, sinon le principal, du moins l'un des principaux défis en matière de ressources humaines au SURS.

V. Conclusion

12. La volonté d'aborder de nouveaux sujets et de rechercher de nouvelles idées, sans craindre les échecs, en étant curieux, en ayant une attitude positive face aux changements constants et en faisant preuve d'enthousiasme pour ce travail est la qualité optimale pour travailler avec les mégadonnées. Cela étant, les compétences doivent être développées pour permettre pleinement d'atteindre les résultats escomptés. Les deux enseignements les plus généraux sur les compétences acquises par le SURS pour mener à bien les projets relatifs aux mégadonnées sont les suivants. Premièrement, avec les mégadonnées, peu de choses diffèrent des autres activités du Bureau ; les compétences sont, dans certains cas, un peu plus spécifiquement axées sur les sources de mégadonnées, mais sont les mêmes dans une grande majorité des cas. Deuxièmement, l'acquisition de compétences n'est pas un problème grave. Elle exige un effort intellectuel, du temps et de l'argent, mais cela mis à part, il existe de nombreuses possibilités d'acquérir des connaissances et des compétences. Le problème réellement grave est aussi le plus classique : comment détecter les nouveaux employés et les employés existants, comment les inciter à acquérir de nouvelles compétences et à assumer de nouvelles responsabilités, et comment les retenir afin de conserver les compétences au sein du Bureau.