



# IIS

Institut Interfédéral de  
STATISTIQUE



## EDITORIAL

Philippe Mauroy  
Vice-Président du Conseil d'administration

### Les institutions statistiques belges soumises à un audit européen

Chaque jour nous sommes submergés par une vague d'informations, de faits et de statistiques. Les institutions publiques de statistique jouent un rôle important à cet égard. Leur objectif est de donner accès à chacun à des statistiques qui sont de qualité élevée, fiables, comparables et crédibles.

D'autre part, les utilisateurs doivent pouvoir vérifier que cette mission est remplie et que les données sont de qualité. Ce n'est que de cette manière que les institutions statistiques pourront convaincre leurs utilisateurs d'avoir confiance.

Pour augmenter encore cette confiance, il y a eu un audit européen en Belgique en décembre 2021. Cet audit, ou examen par les pairs, a été effectué auprès de différentes institutions qui produisent des statistiques européennes en Belgique. Un groupe d'experts internationaux a examiné dans quelle mesure ces statistiques sont en conformité avec le Code de bonnes pratiques de la statistique européenne. Cet examen donnera lieu à des recommandations qui seront publiées au printemps 2022.

Dans cette newsletter, vous trouverez un article expliquant les principes de l'examen par les pairs dans le système statistique européen.

Les thèmes suivants, qui présentent les travaux de l'Institut interfédéral de statistique, sont également abordés dans ce numéro : les indicateurs de suivi des objectifs de développement durable, l'amélioration des statistiques sur les logements dans le cadre du Censu ou encore les Olympiades européennes de Statistique. Enfin, un article présentera comment la technique du machine learning peut être utilisée pour la production de statistiques publiques.

Nous vous souhaitons une agréable lecture.

## IIS NEWSLETTER

### TABLE DES MATIERES

- p. 1** [Editorial](#)
- p. 2-3** [Les peer reviews \(examens par les pairs\) dans le système statistique européen](#)
- p. 4-5** [Indicateurs de suivi des Objectifs de développement durable](#)
- p. 6-7** [Censu Logements – Collaborer pour plus de statistiques de meilleure qualité](#)
- p. 8-9** [Les Olympiades européennes de Statistique mettent les jeunes au défi](#)
- p. 10-11** [L'apprentissage automatique pour les statistiques publiques](#)





## EUROPEAN STATISTICAL SYSTEM

# Les peer reviews (examens par les pairs) dans le système statistique européen



Martine Lefèvre



Principal outil utilisé pour analyser les progrès réalisés dans la mise en oeuvre du [code de bonnes pratiques de la statistique européenne](#), l'examen par les pairs est défini dans le [Handbook of Statistical Organization \(2021\)](#) comme une « Evaluation of a statistical organization, statistical process and/or statistical output by one or more people with similar competencies to the producer. Often used for quality evaluation of a national statistical system, statistical process and/or statistical output by peers from within the same statistical organization or from another statistical organization" (p. 551).

Au niveau du SSE, depuis l'adoption du code de bonnes pratiques en 2005, deux cycles d'examens ont eu lieu et un troisième se déroule actuellement. Ils concernent l'ensemble des Etats membres de l'UE, les pays de l'AELE (Islande,

Liechtenstein, Norvège et Suisse) et Eurostat.

Ce texte ci-après dresse un état des lieux de la thématique, sachant qu'au niveau de l'IIS, les autorités statistiques désignées sont soumises au respect de ce code à travers l'art.36 de l'[Accord de coopération du 15 juillet 2014](#), signé entre l'Etat fédéral et les entités fédérées.

### [Le code de bonnes pratiques de la statistique européenne](#)

Elément indispensable du système d'information de toute société démocratique<sup>1</sup>, la statistique officielle est donc régulée au niveau européen par un code de bonnes pratiques. Toutes les autorités statistiques responsables du développement, de la production et de la diffusion de statistiques officielles se sont engagées à observer ce code.

<sup>1</sup> Nations unies (2014), [premier principe fondamental de la statistique officielle](#).



Comme l'indique l'art. 11 du [règlement européen relatif aux statistiques européennes](#): « Le code de bonnes pratiques a pour objet d'assurer la confiance du public dans les statistiques européennes, en définissant la manière dont celles-ci doivent être développées, produites et diffusées en conformité avec les principes statistiques énoncés à l'art. 2, §1, et les meilleures pratiques statistiques internationales.»

Les principes statistiques visés sont l'indépendance professionnelle, l'impartialité, la fiabilité, le secret statistique et le rapport coût-efficacité.

Sous l'égide du [Comité du système statistique européen](#), ce code, adopté en 2005, a été revu en 2011 ainsi qu'en 2017. Pierre angulaire du cadre qualité commun du SSE, cet instrument d'autorégulation est désormais fondé sur seize principes, qui portent sur l'environnement institutionnel et organisationnel, les processus et les résultats statistiques, assortis chacun d'un ensemble d'indicateurs utilisés lors de l'examen de leur mise en œuvre.

### Le suivi de la mise en œuvre du code avec les examens par les pairs

Ces examens, gérés par Eurostat, se déroulent en plusieurs étapes : autoévaluation par chaque autorité statistique de son respect des principes du code, analyse des documents par une équipe d'experts qui réalise ensuite une visite dans le pays concerné avec un examen approfondi sur la mise en œuvre du code et fournit un rapport avec des recommandations d'amélioration. Ce rapport est soumis à l'autorité statistique concernée pour approbation et pour la rédaction des actions à mettre en œuvre pour suivre les recommandations. Au niveau des pays, le suivi de la mise en œuvre des actions d'amélioration est réalisé par Eurostat ; tandis qu'au niveau de celui-ci, il est réalisé par le [Conseil consultatif européen pour la gouvernance statistique](#), dont la mission est : « d'exercer un contrôle indépendant sur le SSE en ce qui concerne la mise en œuvre du code de bonnes pratiques de la statistique européenne ». L'ensemble de ces travaux et leurs suivis sont publiés [sur le site d'Eurostat](#).

### - 1re revue par les pairs (2006 – 2008)

La première revue a porté sur le degré de conformité des INS aux principes du code relatifs à leur environnement institutionnel et à leurs pratiques de diffusion. Le [rapport de la Commission au Parlement et au Conseil \(octobre 2008\)](#) en synthétise les résultats et prévoit une autre série d'examens par les pairs dans les cinq années suivantes.

En 2012, le Comité du système statistique européen a formulé des recommandations en faveur de la mise en place d'un nouveau cycle d'examens, tenant compte notamment du nombre limité de principes du code observé lors du premier cycle, de la révision du code de 2011, de l'adoption conjointe d'un [Cadre d'assurance qualité](#)<sup>2</sup>, des développements en matière de gouvernance de la statistique européenne, dont la publication du règlement sur les statistiques européennes en 2009, et des conclusions du [rapport spécial de la Cour des comptes européenne de 2012](#) sur les statistiques européennes.

### - 2e revue par les pairs (2013 – 2015)

Un deuxième cycle d'examens par les pairs a donc été réalisé au sein du SSE de 2013 à 2015. Il s'est différencié du premier sur plusieurs points dont l'élargissement de l'examen à l'ensemble des principes du code et l'intégration dans le processus d'un certain nombre d'[autres autorités statistiques nationales \(ONAs\)](#). Il a également été suivi par un [rapport de la Commission au Parlement européen et au Conseil \(mars 2016\)](#).

Dans son [rapport annuel de 2020](#) (p. 36), le Conseil consultatif européen pour la gouvernance statistique indiquait qu'à la suite de cette 2e revue par les pairs, 707 recommandations avaient été proposées, conduisant à la formulation de 910 actions d'amélioration par les institutions statistiques concernées. Il notait qu'à la fin 2019, 118 actions restaient à réaliser.

### - 3e revue par les pairs (2021 – 2023)

Sur la base des résultats de la revue précédente et après la révision du code en 2017, les responsables de la statistique européenne ont décidé, dès 2018, de réévaluer la conformité des autorités statistiques nationales, des ONAs et d'Eurostat avec ce code révisé. Cette 3e revue étend son champ d'application aux ONAs, sélectionnés par l'INS avant la visite d'évaluation par les pairs sur la base d'un processus spécifique.

Elle prend en compte l'impact de la pandémie de Covid 19 sur la statistique officielle, ainsi que la complexité croissante de l'environnement de la production statistique. Elle poursuit deux objectifs :

- prouver aux diverses composantes du SSE et aux parties prenantes externes à celui-ci que le SSE est fondé sur des principes fondamentaux et qu'il fonctionne dans le respect de ceux-ci :

- aider les INS, les ONAs sélectionnés et Eurostat à améliorer et à développer davantage leur système statistique par des recommandations d'amélioration prospectives qui pourraient aller au-delà du code actuel; mais également en stimulant les autorités gouvernementales pour qu'ils soutiennent la mise en œuvre des recommandations.

Ces objectifs ciblent à la fois les parties prenantes internes et externes au SSE. Cette dimension externe est importante étant donné le besoin de disposer de statistiques fiables et de renforcer la confiance du public dans les statistiques officielles. Pour atteindre ces objectifs, une stratégie de communication particulière accompagnant la réalisation de cette 3e revue a été décidée début 2021.

[L'ensemble de la documentation sur la mise en place et les modalités d'organisation de cette troisième revue](#) est publié sur le site d'Eurostat ; avec l'indication de la planification des examens et du calendrier prévu pour la publication des rapports. Au niveau de la Belgique, la revue s'est déroulée du 6 au 10 décembre 2021 et le rapport est attendu en avril 2022.



<sup>2</sup> Guide et compléments pour la mise en œuvre du code de bonnes pratiques de la statistique européenne.

# Indicateurs de suivi des Objectifs de développement durable

Alain Henry

Un groupe de travail de l'IIS est chargé de coordonner le suivi des *Sustainable Development Goals* sur la base d'indicateurs. Un ensemble de 116 indicateurs a été défini, dont 81 sont actuellement publiés sur le [site « indicateurs » du Bureau fédéral du Plan](#). Cet ensemble pourrait évoluer à l'avenir, en particulier suite à la consultation organisée en 2022 auprès des principaux Conseils d'avis fédéraux.

## Les SDG et leurs indicateurs de suivi

En 2015, l'ONU a adopté les *Sustainable Development Goals* (SDG, objectifs de développement durable), un ensemble de 17 objectifs ambitieux à l'horizon 2030. Ces 17 SDG sont répartis en [169 sous-objectifs](#). Le suivi du progrès dans la réalisation de ces objectifs est organisé à partir de [231 indicateurs](#) définis par la Commission statistique de l'ONU.

Avec l'adoption des SDG, l'ONU invite les pays à suivre les progrès vers ces objectifs à l'aide d'indicateurs. En Belgique, l'IIS a été chargé d'organiser le suivi de ces indicateurs et a créé à cette fin en 2016 le groupe de travail (GT) « Indicateurs pour les SDG ». C'est un des quatre groupes de travail permanents chargés des travaux préparatoires pour le Conseil d'administration de l'IIS. Le groupe de travail est composé de représentants de Statistics Belgium, de la Banque nationale de Belgique, du Bureau fédéral du Plan (BFP) et des autorités statistiques régionales (Statistiek Vlaanderen, IWEPS et IBSA). Le BFP en assure la présidence et le secrétariat.

Le rôle de ce GT est de coordonner les travaux sur les indicateurs de suivi des SDG entre les partenaires de l'IIS et plus particulièrement de formuler des propositions sur le maintien d'une base de données de

ces indicateurs et sur le développement de nouveaux indicateurs. Il s'agit aussi de suivre les travaux de l'ONU et d'Eurostat sur le sujet et, si nécessaire, de formuler des avis complémentaires et de définir des positionnements internationaux sur les questions statistiques liées aux indicateurs pour les SDG.

## Méthode pour constituer les listes d'indicateurs

Selon le mandat initial du GT, il s'agissait de partir de la liste des indicateurs de l'ONU pour identifier les indicateurs déjà disponibles en Belgique et dont le suivi pouvait être rapidement organisé. En s'inspirant de la méthode utilisée par Eurostat pour définir les indicateurs au niveau européen, le GT a réparti les indicateurs en différentes catégories, au regard des données existantes en Belgique. Selon le consensus entre les experts participants, certains indicateurs ont été exclus du travail du GT, car étant de nature non statistique (par exemple, est-ce que la Belgique dispose d'une stratégie de réduction des risques liés aux catastrophes naturelles – indicateur 1.5.3 de l'ONU) ou ne s'appliquant pas à la Belgique (par exemple la moyenne mondiale pondérée des taux de droits de douane, indicateur 17.10.1). Les autres indicateurs ont alors été répartis entre ceux pour lesquels les données étaient disponibles et ceux pour lesquels un travail supplémentaire était nécessaire.

## Résultats: des ensembles d'indicateurs

Suite à ces travaux, le Conseil d'administration de l'IIS a approuvé en février 2021 un ensemble de 116 indicateurs comme liste de suivi des indicateurs des SDG en Belgique.

Sur ces 116 indicateurs, 81 sont actuellement publiés sur le [site « indicateurs de développement durable » du BFP](#). Pour de nombreux indicateurs, des ventilations par

catégorie de population sont également proposées. Ainsi, 37 indicateurs sont ventilés selon le sexe, 40 selon la région, 19 selon le revenu, 11 selon le niveau d'éducation, etc. Une comparaison avec la moyenne européenne ou des pays voisins est fournie pour 73 indicateurs.

Il faut noter que parmi ces 81 indicateurs, 37 ne sont pas inclus dans la liste de l'ONU. Ils couvrent des problématiques pertinentes en Belgique, mais qui ne sont pas reprises au niveau mondial. Il s'agit par exemple de l'obésité des adultes (sous-objectif 2.2 sur la malnutrition), du report ou de l'annulation de soins de santé pour raisons financières (sous objectif 1.4 sur l'accès aux services de base), de la consommation d'énergie primaire (sous objectif 7.3 sur la productivité de l'énergie) ou de la congestion routière (sous objectif 9.1 sur l'accès à des infrastructures de qualité).

Cet ensemble de 81 indicateurs pourrait être élargi aux 116 indicateurs mentionnés ci-dessus. En plus de ces 116 indicateurs, 31 indicateurs supplémentaires sont publiés par l'ONU pour la Belgique. Ces indicateurs pourraient également être ajoutés si les méthodes utilisées par l'ONU étaient identifiées pour établir un lien direct avec les sources de données belges. Enfin, le GT a identifié 87 autres indicateurs qui demanderaient un travail de recherche supplémentaire avant de pouvoir être ajoutés à cet ensemble. Mais pour réaliser ces différents travaux, des moyens humains supplémentaires devraient être dégagés.

### Utilisation de ces indicateurs

En 2017, une sélection de 34 indicateurs a aussi été présentée dans l'annexe statistique de l'[Évaluation nationale volontaire](#) présentée par la Belgique à l'ONU à l'occasion du Forum politique de haut niveau sur le développement durable.

Les indicateurs de suivi des SDG sont régulièrement utilisés au BFP. Le progrès de la Belgique vers les SDG est mesuré avec une sélection de 51 indicateurs. Pour chaque indicateur, le BFP évalue si la prolongation des tendances actuelles permet d'atteindre son objectif, fixé sur la base des SDG et/ou d'autres engagements pris par la Belgique.

Les derniers bilans ont été publiés en juin 2019 dans le [Rapport fédéral](#)

[sur le développement durable](#), en juillet 2021 dans un article [Atteindre tous les SDG en 2030 demande une rupture de tendance](#), titre qui résume bien les conclusions qui peuvent être tirées des bilans successifs réalisés depuis 2005, ainsi que dans le [Rapport Indicateurs de développement durable](#) de février 2022.

Le BFP utilise également ces indicateurs lors de la préparation des Rapports fédéraux sur le développement durable, tant dans la partie évaluation que la partie prospective de ces rapports.

### Travail futur

Ces listes d'indicateurs sont en évolution constante, au fil des modifications apportées par l'ONU, des nouvelles données qui deviennent disponibles, des nouvelles problématiques à couvrir (par exemple les questions de santé mentale suite à la pandémie de Covid-19), etc. Le rôle du GT est aussi de suivre ces évolutions pour en informer le Conseil d'administration de l'IIS.

Au niveau belge, deux événements sont appelés à influencer ces listes d'indicateurs. Tout d'abord le Plan fédéral de développement durable approuvé en octobre 2021 par le gouvernement fédéral prévoit d'organiser un débat sur le choix des indicateurs de suivi des SDG. Une proposition des experts de l'IIS et des administrations fédérales sera soumise aux principaux Conseils d'avis fédéraux.

Ensuite, la Belgique devrait préparer pour mi-2023 une deuxième Evaluation nationale volontaire. Elle inclura vraisemblablement, comme l'édition de 2017, une annexe statistique. Les demandes formulées relativement à cette annexe orienteront également les travaux de l'IIS sur les indicateurs de suivi des SDG.



# Census Logements - Collaborer pour plus de statistiques de meilleure qualité

Saskia Hin

Fin 2018, la collaboration de l'IIS sur le Census Logements a été mise en place pour améliorer les statistiques sur les logements dans le cadre du Censur 2021. Les différences observées par le passé entre le nombre de logements et le nombre de ménages ont été examinées plus en détail afin d'identifier les causes et d'améliorer les mesures. Par ailleurs, les besoins en nouveaux indicateurs et variables ont été répertoriés. Cela a conduit à formuler des recommandations pour le prochain Censur.

## Objectifs

Fin 2018, l'IIS a mis en place une collaboration sur le Census Logements en vue de préparer la mise en œuvre du Censur 2021 en Belgique. Le Censur 2021 est organisé dans toute l'UE. Eurostat coordonne les travaux et fournit ainsi des définitions et métadonnées uniformes. Il collecte également les [résultats du Censur](#). En plus des indicateurs obligatoires, les pays ont la possibilité de collecter d'autres indicateurs, selon leurs propres définitions.

La collaboration de l'IIS s'est penchée spécifiquement sur le logement, un des thèmes du Censur. Il s'agit d'une collaboration entre Statbel, la VSA, l'IWEPS et l'IBSA. L'IWEPS a assuré la présidence et le secrétariat du groupe de travail. Cette collaboration poursuivait deux objectifs :

- Premièrement, il a fallu comprendre pourquoi, dans certains secteurs statistiques ou communes, le nombre de logements était nettement inférieur au nombre de ménages lors du Censur 2011 afin d'améliorer la méthode d'identification des logements ;
- Deuxièmement, le groupe de travail a examiné de nouveaux indicateurs et/ou variables statistiques qu'il serait bon de développer dans le cadre du Censur 2021. En particulier, il a été nécessaire de disposer d'indicateurs qui donnent des informations sur la propriété et les profils des propriétaires ainsi que sur les caractéristiques des logements.

## Première phase : demandes de données pour analyse

La première étape a consisté à obtenir des données nécessaires. Plusieurs pistes ont été explorées à cet effet. L'approche souhaitée d'une demande commune de tous les partenaires du projet s'est finalement avérée trop complexe d'un point de vue juridique. Les différents partenaires ont donc introduit des demandes séparées. Fin 2019, les différents partenaires ont pu commencer à travailler sur les données. Il s'agissait d'informations provenant du cadastre, du registre national de la population et de la Banque Carrefour des Entreprises (BCE).

## Identification des logements

Lorsqu'on procède au couplage des données du registre national et celles du cadastre, on constate régulièrement que le nombre de ménages à une adresse est supérieur au nombre de logements à la même adresse selon le cadastre. Cela peut être dû par exemple aux ménages qui ont quitté le pays mais qui restent encore enregistrés au registre national. Cela peut être aussi dû à certains logements qui ne sont pas enregistrés au cadastre. Il s'agit par exemple des logements qui sont transformés en plusieurs appartements et ces derniers ne sont pas enregistrés au cadastre.

Afin d'éviter que le nombre de logements ne soit ainsi sous-estimé, des « logements virtuels » ont été créés lors du Censur. La méthodologie utilisée lors du Censur 2011 pour créer ces logements virtuels a été examinée par le groupe de travail de l'IIS.

Le problème identifié, à savoir le nombre de logements est inférieur au nombre de

ménages, varie fortement d'une région à l'autre. En Région flamande, le problème se posait uniquement à Anvers, Louvain et Gand. En Région de Bruxelles-Capitale, il concernait toutes les communes, même s'il était plus important dans les communes centrales. En Région wallonne, le problème concernait un mélange de communes aux caractéristiques très différentes (par exemple, aussi bien des communes urbaines que des communes rurales).

Les différences géographiques dans la création de ces logements virtuels ont été examinées en détail. Le processus et les hypothèses pour identifier des logements virtuels ont été évalués. Sur la base de cette étude, le groupe de travail a formulé des recommandations pour améliorer encore la méthode de détermination de ces logements virtuels lors du prochain Censur.

En outre, un questionnaire a été envoyé aux collègues des autres pays européens qui organisent un recensement administratif des logements. Les différences entre le cadastre et le registre de la population sont nettement moins importantes dans les autres pays européens que dans notre pays. Dans les autres pays européens, les rapports sont généralement réciproques entre les administrations qui produisent des données telles que le cadastre et les instituts de statistique. Les instituts de statistique utilisent non seulement les données des administrations, mais les administrations tiennent également compte des suggestions de ces instituts. En Belgique, les instituts de statistique utilisent également les données des administrations. Cependant, les administrations n'améliorent pas leur fonctionnement sur la base des suggestions éventuelles des instituts de statistique. Le code de bonnes pratiques de la statistique européenne, auquel se sont engagés les différentes autorités régionales et fédérale, mentionne à ce propos la ligne directrice suivante : « Lorsque les statistiques européennes sont fondées sur des données administratives ou autres, les définitions et les concepts utilisés à des fins non statistiques sont une bonne approximation de ceux qui sont

employés en statistique. » Une consultation mutuelle pour une meilleure coordination est donc utile et peut s'avérer bénéfique tant pour les instituts de statistique que pour les administrations publiques.

### Nouveaux indicateurs et variables

Fin 2020, la collaboration entre les mêmes partenaires a été prolongée afin d'identifier de nouveaux indicateurs et variables sur le thème du logement qui sont prioritaires pour les régions. Les dimensions sur lesquelles a porté le travail sont la propriété et les profils des propriétaires (le nombre de logements possédés par le propriétaire ou la présence d'usufruit) ainsi que les caractéristiques des logements (le certificat de performance énergétique et l'inoccupation). Le groupe de travail a discuté de ces priorités et établi un classement commun de celles-ci. Les définitions ont été élaborées en détail pour assurer une description sans équivoque et garantir la comparabilité des indicateurs entre les régions. Fin septembre 2021, une liste des indicateurs et variables souhaités a été consolidée et incluse dans un rapport interne destiné au conseil d'administration de l'IIS. La faisabilité des différents indicateurs et variables a été évaluée par l'équipe de Statbel, ce qui a débouché sur un classement de faisabilité.

### Et quoi d'autre ?

Les objectifs du projet de collaboration ont donc été atteints et le projet a été formellement clôturé. Néanmoins, les partenaires de l'IIS souhaitent maintenir la plateforme pour poursuivre les échanges de connaissances sur le thème du logement.

Après la fourniture à Eurostat des indicateurs et variables obligatoires pour le Census, Statbel déterminera, en fonction du temps et des ressources disponibles, quels indicateurs et variables qui sont souhaités par l'IIS pourront être effectivement produits et dans quels délais. Statbel pourrait ainsi demander dans ce cadre un éventuel détachement de collaborateurs des partenaires régionaux. En 2022, un groupe de travail de l'IIS composé des mêmes partenaires sera de nouveau mis en place afin d'assurer le suivi de ce processus dans le cadre d'une collaboration continue entre Statbel et les partenaires régionaux.



# Les Olympiades européennes de Statistique mettent les jeunes au défi

Wendy Schelfaut

Dans le cadre des Olympiades européennes de Statistique, les élèves des deuxième et troisième degrés de l'enseignement secondaire relèvent le défi de tester leurs connaissances et leurs capacités à interpréter des statistiques officielles. Les Olympiades sont organisées en Belgique depuis l'année scolaire 2018-2019 par Statbel, Statistiek Vlaanderen et l'IWEPS, en collaboration avec Eurostat et INE Espagne.

## 782 élèves participent

Les Olympiades européennes de Statistique sont une compétition atypique. Contrairement aux autres olympiades, les élèves travaillent plutôt en équipes. Un professeur accompagne ses élèves tout au long de la compétition. Ils décident ensemble si les élèves travaillent individuellement, en duo ou en équipes de trois. Lors de cette année scolaire 2021-2022, 322 équipes se sont inscrites aux Olympiades européennes de Statistique: 246 équipes du troisième degré et 76 du deuxième degré de l'enseignement secondaire. Deux tiers des équipes participent en néerlandais (213 équipes), un tiers en français (109 équipes).

## Théorie, recherche et compétences créatives

Les Olympiades européennes de Statistique testent différentes compétences des élèves participants: leur connaissance théorique des statistiques, leurs aptitudes en matière de recherche et leurs compétences créatives. Toutes les épreuves se déroulent en ligne. Les élèves sont accompagnés d'un professeur tout au long des épreuves. La branche de celui-ci n'a pas d'importance. Il peut s'agir d'un professeur de mathématiques, d'économie, de géographie mais également de langues ou d'autres matières plus créatives. En effet, les compétences de toutes ces matières sont couvertes et, le plus souvent, les enseignants de ces matières au sein de l'école travaillent ensemble pour soutenir leurs équipes le mieux possible.

En Belgique, les Olympiades de Statistique se déroulent en deux phases. Un premier test en ligne (janvier 2022) évalue trois compétences: les connaissances de base en statistique, les compétences en matière d'analyse et l'interprétation de rapports statistiques. Ce test est éliminatoire: les équipes réussissent si elles obtiennent au moins 50%. Parmi les équipes qui ont réussi, seul le top 20 par communauté et par degré participe à la deuxième épreuve. Donc au maximum 20 équipes néerlandophones du 2e degré et 20 équipes néerlandophones du 3e degré, et idem pour les communautés francophone et germanophone.

La deuxième épreuve (février - mars 2022) évalue les compétences créatives; les élèves peuvent-ils communiquer sur les statistiques de manière correcte et créative dans une vidéo? Toutes les équipes sélectionnées ont pour mission de créer une vidéo de maximum 2 minutes sur un sujet déterminé à l'avance, en anglais ou du moins avec des sous-titres anglais. Ce thème est le même pour tous les pays européens, étant donné que les lauréats participeront à la compétition au niveau européen.

Début mai 2022, aura lieu la cérémonie belge de remise des prix et l'annonce des quatre équipes belges (deux par degré) qui participeront à l'étape européenne avec leur vidéo.

## Culture statistique

Les statistiques font partie intégrante de notre vie quotidienne. Nous sommes submergés de chiffres et de données provenant des sources les plus diverses. Il existe des instituts officiels de statistique comme Eurostat pour l'Europe, Statbel pour la Belgique et Statistiek Vlaanderen, l'Iweps et l'IBSA pour la Flandre, la Wallonie et la Région de Bruxelles-Capitale, respectivement.

Ils s'efforcent, en tant qu'offices publics de statistique, d'améliorer les connaissances statistiques du grand public. Les Olympiades européennes de Statistique mettent les élèves au défi de travailler avec des statistiques, d'apprendre à interpréter des données et de présenter leurs conclusions.





## Partenaires nationaux et internationaux

Les Olympiades européennes de Statistique sont une compétition unique en leur genre. Le succès de la compétition est assuré par une coopération tant à l'intérieur qu'à l'extérieur de nos frontières. Au niveau international, les partenaires centraux sont Eurostat et INE Espagne. Ils sont responsables de la coordination internationale, assurent le soutien technique et organisent la phase européenne de la compétition. En outre, des consultations régulières ont lieu avec les offices statistiques de tous les pays participants: la Bulgarie, la Croatie, Chypre, la Finlande, la France, la Grèce, la Hongrie, l'Islande, l'Irlande, l'Italie, le Liechtenstein, la Lituanie, le Luxembourg, la Norvège, la Pologne, le Portugal, la Slovénie, l'Espagne et, bien entendu, la Belgique.

En Belgique, la coopération autour des Olympiades européennes de Statistique trouve son origine dans l'Institut interfédéral de Statistique (IIS). En 2018, Statbel, Statistiek Vlaanderen et l'IWEPS ont décidé d'unir leurs forces et d'organiser les Olympiades dans tout le pays. Les trois institutions travaillent ensemble pour assurer la plus large communication possible sur les Olympiades. Les trois partenaires sont également impliqués dans les évaluations du contenu par le biais de jurys communautaires et d'un jury national, ainsi que dans la cérémonie nationale de remise des prix à la fin du processus.

Outre la coopération au sein de l'IIS, les Olympiades européennes de Statistique en Belgique sont également soutenues par d'autres partenaires: le *Departement Economie, Wetenschap en Innovatie* et le *Departement Onderwijs* de la Communauté flamande, le Ministère de la Communauté germanophone, le Département Enseignement de la Fédération Wallonie-Bruxelles, la *Vlaamse Vereniging voor Wiskundeleraars* et la Société royale belge de Statistique.



### Pour plus d'informations?

La communication sur les Olympiades européennes de Statistique se fait par plusieurs canaux:

Le site web: <http://www.olympiadesstatistique.be>

Une newsletter bimensuelle; inscription possible via le site web

Une page Facebook: <https://www.facebook.com/EuropeanStatisticsCompetitionBelgium>

Vous pouvez également joindre l'équipe qui coordonne l'organisation à l'adresse e-mail [statolympiade@economie.fgov.be](mailto:statolympiade@economie.fgov.be)

# L'apprentissage automatique pour les statistiques publiques

Michael Reusens

## Introduction

L'apprentissage automatique est un sujet brûlant dans le monde des statistiques publiques. Cet article donne des explications sur l'une des formes d'apprentissage automatique, à savoir l'apprentissage supervisé, et examine les opportunités et les défis de l'apprentissage automatique pour les statistiques publiques.

## Apprentissage automatique

L'apprentissage automatique recouvre des algorithmes informatiques qui s'améliorent automatiquement en apprenant sur la base de l'expérience et de données<sup>1</sup>. Ces algorithmes construisent un modèle basé sur des données, qui permet de faire des prévisions ou de prendre des décisions toujours meilleures. L'apprentissage automatique fait partie de l'«intelligence artificielle».

L'apprentissage automatique comporte plusieurs sous-branches qui offrent des opportunités pour les statistiques publiques: apprentissage supervisé, non supervisé et par renforcement. L'«apprentissage supervisé» est celui qui connaît le plus d'applications. L'apprentissage supervisé est une forme dans laquelle le modèle est entraîné sur la base de données étiquetées. Ce modèle est ensuite utilisé pour prévoir les caractéristiques (étiquettes) des nouvelles observations. Par exemple, un modèle pourrait être développé pour prévoir le prix d'une maison sur la base de certaines de ses caractéristiques, notamment la surface habitable, la commune, le nombre de chambres et l'année de construction. La figure 1 montre comment se présente typiquement un set de données de ce type (voir tableau) pour l'apprentissage supervisé.

À partir d'un grand nombre d'entrées, un modèle qui prévoit le prix d'un logement neuf en fonction de ses caractéristiques descriptives est construit. La figure 1 montre également que le modèle est continuellement amélioré sur la base de nouvelles observations : dès que le prix d'une nouvelle maison est connu, cette nouvelle observation peut automatiquement

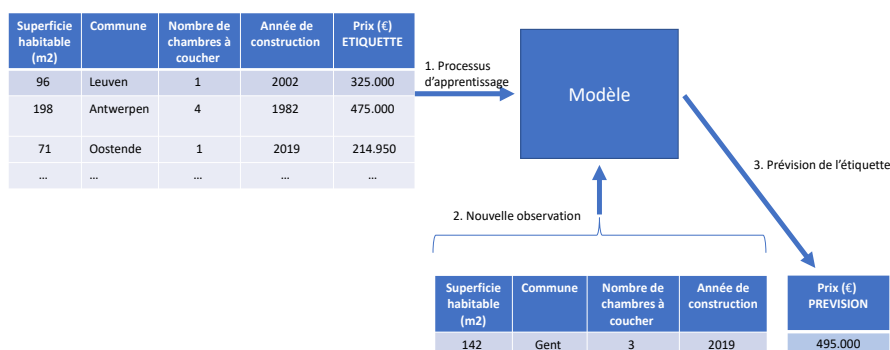


Figure 1 : Représentation schématique de l'apprentissage et de l'application du modèle d'apprentissage automatique

être utilisée pour améliorer l'algorithme informatique. Le modèle dont nous parlons ici de manière schématique pourrait, par exemple, être appliqué pour estimer automatiquement les prix des logements.

Il existe de nombreux modèles d'apprentissage supervisé, notamment la régression logistique, les arbres décisionnels et les réseaux neuronaux. Ces modèles diffèrent les uns des autres par la complexité des schémas qu'ils peuvent apprendre, le degré de compréhension que vous pouvez avoir d'un schéma entraîné, la quantité de données nécessaire pour apprendre et la puissance de calcul requise.

1 Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.

## Opportunités pour les statistiques publiques

L'exemple ci-dessus ne contient que des variables structurées. Le monde des statistiques a des décennies d'expérience dans le traitement et l'analyse de ce type de données. Les statistiques publiques sont traditionnellement produites sur la base d'enquêtes et de sources de données administratives structurées, telles que les données du registre national. Les techniques d'apprentissage automatique peuvent également traiter ces données.

En outre, le monde de l'apprentissage automatique a développé des **méthodes très performantes pour les 'données non structurées'**, telles que les images ('vision par ordinateur'), le texte libre ('traitement automatique du langage naturel') et le son ('audition par ordinateur'). De telles statistiques apparaissent de plus en plus fréquemment :

- Le département *Landbouw en visserij* (gouvernement flamand) analyse la [croissance des variétés cultivées](#) à l'aide d'algorithmes de 'vision par ordinateur'.
- Le *Bureau of Labor Statistics* (USA) catégorise automatiquement les [accidents du travail](#) décrits en texte libre et en tire des statistiques.
- Le *Centraal Bureau voor Statistiek* (NL) prévoit la [probabilité de déménagement](#) de chaque habitant sur la base de données de registre.

Les techniques d'apprentissage automatique présentent les avantages suivants par rapport à la 'production statistique classique' sur la base de données administratives :

- Les prévisions obtenues par apprentissage automatique permettent d'obtenir **plus d'informations** à partir des **données (administratives) disponibles**.
- Les méthodes de travail sur les données non structurées permettent d'**identifier de nouveaux phénomènes** et d'**améliorer les statistiques existantes**.

Les techniques d'apprentissage automatique peuvent également présenter certains avantages par rapport aux enquêtes :

- **Réduction de la charge de réponse** car les caractéristiques ne doivent pas être demandées mais sont dérivées de données disponibles.
- Des statistiques **plus fréquentes et produites en temps opportun** car les données sont souvent disponibles en temps réel, alors que les enquêtes sont réalisées périodiquement.
- Des statistiques **moins chères** par rapport au coût élevé des enquêtes, dans les cas où les données nécessaires peuvent être obtenues à moindre coût.

## Défis pour les statistiques publiques

Les opportunités offertes par l'apprentissage automatique et les nouvelles sources de données dépendent de la capacité à relever les défis suivants.

- **Biais** dans les nouvelles sources de données. Les enquêtes sont conçues de manière à être aussi représentatives que possible du groupe visé, par exemple les adultes âgés de 18 à 45 ans. Les nouvelles sources de données sont souvent dérivées de processus/activités non statistiques existants, tels que les messages sur les médias sociaux, les sites web des entreprises ...
- **Transparence** des méthodes. Les modèles d'apprentissage automatique ont souvent une quantité importante de contenu qui est opaque. Il est dès lors difficile de communiquer pourquoi une étiquette a été attribuée à une observation.
- Des **questions de confidentialité et d'éthique** se posent rapidement lorsqu'il s'agit d'analyser des données à caractère personnel à grande échelle, par exemple si certains groupes de population risquent d'être désavantagés par le modèle. Le règlement général sur la [protection des données](#) est un fil conducteur important pour l'utilisation correcte des données à caractère personnel.
- **Compétences en matière d'apprentissage automatique**. Pour optimiser les opportunités offertes par ces nouvelles techniques, il faut accroître l'expertise au sein des institutions de statistique.
- **Accès** aux données nécessaires. Dans de nombreux cas, les données requises ne sont pas détenues par les services statistiques, mais par d'autres entités et organisations privées. Il est souvent très difficile d'obtenir l'accès à ces données pour la production de statistiques.

## Le centre de science des données de Statistiek Vlaanderen

Statistiek Vlaanderen développe un centre de science des données (data science hub) depuis 2019. Il se concentre sur l'accès à de nouvelles sources de données et l'application de l'apprentissage automatique pour produire des statistiques nouvelles et de meilleure qualité sur la Flandre. Par exemple, le centre travaille à l'élaboration de nouvelles statistiques sur l'industrie 4.0 au moyen du webscraping, et sur l'apprentissage tout au long de la vie à partir de données de Twitter. En outre, il organise des formations pour accroître les compétences en matière d'apprentissage automatique au sein de Statistiek Vlaanderen. Ces activités sont réalisées en coopération avec des partenaires au sein de l'Autorité flamande, des institutions scientifiques et d'autres services statistiques confrontés aux mêmes défis.

# COLOPHON

Ont participé à ce numéro :

Alain Henry – Expert - Bureau fédéral du Plan

Saskia Hin – Experte - IBSA

Martine Lefèvre – Attachée scientifique - IWEPS

Philippe Mauroy – Directeur général a.i. - Statbel

Michael Reusens – Data Science Coördinator - Statistiek Vlaanderen

Wendy Schelfaut – Directrice du service « Communication & Diffusion » - Statbel

L'IIS est le résultat d'un accord de coopération de 2014.  
L'IIS doit permettre d'améliorer en continu la qualité des statistiques publiques, venant en soutien des politiques publiques.



[www.iis-statistics.be](http://www.iis-statistics.be)