Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Statistics Netherlands and AI

by Barteld Braaksma and May Offermans

Release date: November 5, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Statistics Netherlands and AI

Barteld Braaksma and May Offermans[1]

Abstract

The ways in which AI may affect the world of official statistics are manifold and Statistics Netherlands (CBS) is actively exploring how it can use AI within its societal role. The paper describes a number of AI-related areas where CBS is currently active: use of AI for its own statistics production and statistical R&D, the development of a national AI monitor, the support of other government bodies with expertise on fair data and fair algorithms, data sharing under safe and secure conditions, and engaging in AI-related collaborations.

Key Words: Artificial Intelligence; Official Statistics; Data Sharing; Fair Algorithms; AI monitoring; Collaboration.

1. Introduction

In the coming years, Artificial Intelligence (AI) will be used more and more by government (ministries, local and regional government bodies as well as executing organisations) to tackle major societal challenges, such as poverty policy, the energy transition and climate change or combating society-undermining criminality. Statistics Netherlands or CBS is actively involved in this development as described in its position paper on AI (Dutch only).

CBS has a unique position in the Netherlands, laid down by law. As the national statistical office, CBS is the data hub of the Dutch government. By law, CBS has access to all government sources and to a growing amount of big data from private parties for statistical use. CBS collects, combines and processes the data and shares the results with society in the form of reliable and valuable information.

The ways in which AI may affect the world of official statistics are manifold and CBS is actively exploring how it can benefit from AI as an institute and use AI within its societal role. In this paper we describe a number of AI-related areas where CBS is currently active: use of AI for its own statistics production and statistical R&D, the development of a national AI monitor, the support of other government bodies with expertise on fair data and fair algorithms, data sharing under safe and secure conditions, and engaging in AI-related collaborations.

2. AI projects at Statistics Netherlands

First and foremost, CBS already explores and uses several AI methods to produce statistics. These techniques are being used especially when processing new (big) data sources or when dealing with large numbers of variables. Some cases related to e.g. automated coding are already in production, but more often AI projects are still in the R&D stage and published as experimental statistics or in research reports. Below we describe a number of recent cases. Further examples can be found on the CBS innovation site.

2.1 Innovative hotspots

Once every two years, Statistics Netherlands carries out an innovation survey among companies with at least 10 employees. But is it also possible to gain insight into innovation in companies with fewer than 10 employees? This

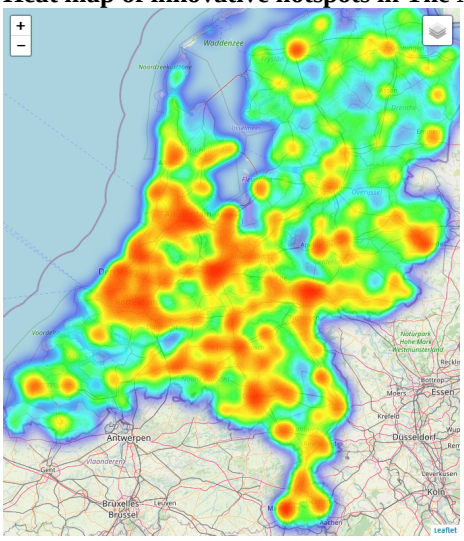
[1]Barteld Braaksma, Statistics Netherlands, P.O. box 24500, NL-2490HA The Hague, The Netherlands (b.braaksma@cbs.nl); May Offermans, Statistics Netherlands, P.O. box 4481, NL- 6401CZ Heerlen, The Netherlands.

would enable, for example, the mapping regional clusters of innovative companies in the Netherlands, so-called innovative hotspots, with a greater degree of detail. Together with the company Innovatiespotter, CBS has carried out a study for the Ministry of Economic Affairs and Climate. The approach developed uses a combination of web scraping, text mining and AI. Combining these techniques with data from sources such as company websites and the CBS innovation survey allows to provide a more detailed picture.

In the first stage, websites were identified for as many companies as possible from the General Business Register of CBS- assuming that the vast majority of innovative companies will have a website.. Subsequently, keywords were determined that are characteristic of sites of innovative companies. This enables the classification of websites as belonging to an innovative or non-innovative company. The results were validated in various ways, for example with a detailed study in the Eindhoven region.

The map below (presented in interactive form in this article-Dutch only on the CBS website) shows the concentrations of innovative companies in the Netherlands in a so-called heat map. A user can choose between a distribution based on the number of companies or based on the number of employed persons. In the latter case, the areas with relatively large companies stand out. For the purpose of the heat map, all innovative companies are grouped in areas of 1 by 1 km, with a minimum of 5 companies per area. This method prevents traceability to individual companies. It does, however, mean that areas with few innovative companies disappear from the map.

Figure 2.1-1
Heat map of innovative hotspots in The Netherlands

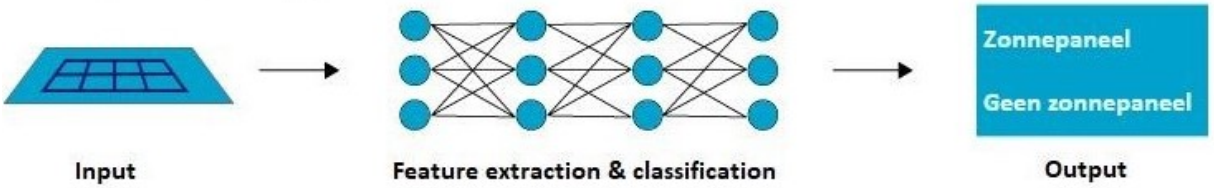


With this method, CBS believes it has found a way to provide more detailed statistics on innovation in the Netherlands. In addition, it also seems possible to look specifically at e.g. a demarcation (population and location) of start-ups. The method has been tried in other countries (e.g. Belgium and Sweden) and on other topics (notably on determining the size of the European drone industry and detecting companies involved in AI development). Results were mixed and it appeared that the method is not ‘magical’- a critical eye and common sense are indispensable to interpret results.

2.2 Detection of solar panels

In the past, CBS statistics on solar power were based on a survey among about 300 suppliers and importers of solar panels. Recently, a new method has been adopted whereby the installed capacity is derived from administrative registers. A consequence of the new method is that figures can be reported at better granularity. The registration of solar panels, however, is not mandatory in the Netherlands. The registers therefore do not fully cover the installed base of solar panels.

Figure 2.2-1
Scheme of the deep learning approach used to detect solar panels



In the research project 'Deep Solaris', CBS collaborates with the Dutch Open University and the statistical offices of Flanders and Germany to get a complete and detailed picture of installed solar panels by using machine learning models on aerial photographs, resulting in an interactive map with the locations of solar panels in sample regions Flanders (BE), North Rhine-Westphalia (DE) and Limburg (NL). From these locations, regional statistics with numbers of solar panels can be made. Various machine learning techniques were investigated, such as random forest, support vector machine and convolutional deep neural networks. The best performing models, Xception and VGG16, achieve training and test accuracy of over 90%. The results of the experiments show that it is possible to use AI successfully to classify solar panels from aerial photographs. See the article on the CBS innovation site for further details.

2.3 Detecting cyber crime from police reports

With the digitization of society, interest in cybercrime is growing. Given the complexity and constantly evolving nature of cybercrime, the total extent of cybercrime is unknown at this time. In order to gain better insight, it is important first of all to be able to estimate the number of cyber-related crimes, or crimes in which ICT has played a role. To obtain a first estimate, existing police reports were analyzed. That is not easy: only a small part of the relevant offenses can be directly derived from them. A further distinction must be made manually between digital offenses and traditional forms of crime. However, this manual method is very time consuming and prone to errors.

CBS has therefore developed an automated method for investigating official police reports. In order to find out in which registered crimes a cyber-aspect plays a role, the official reports from 2016 were examined by means of AI-based text analysis, in close collaboration with the police. In this way, the size of the registered cybercrime can be estimated better and considerably faster than with the manual method.

To find out whether there is a cyber-aspect to a reported crime, the content of the text itself has to be searched. To enable the computer to search for cybercrime, the following three steps were followed:

- Define cybercrime
- Develop the algorithm: classify relevant keywords and train the model
- Test and optimize the model

The final model was applied to the full set of 820 thousand official police reports for 2016. The result was that in 2016, more than 72 thousand police reports involved cybercrime, which amounts to almost 9 percent. The share of cybercrime differs greatly per type of offense from the Standard Crime Classification used by Statistics Netherlands. For example, in the category of fraud almost all analyzed police reports can be classified as cybercrime. For traffic offenses, the percentage of cybercrime is virtually nil. Further details can be found here (Dutch only).

2.4 Green and grey areas in the inner city

The amount of green and grey, or the ratio between vegetation and paving, has impact on the local climate in a city. Municipalities try to formulate policies to influence this ratio. But how do you find out what the ratio is and whether it changes? With support from the company Ellipsis Drive, CBS carried out research into the use of AI methods to aerial images. The main question was 'can image recognition using deep learning be used to classify land surface as paved or green?' To answer this question, aerial photographs, both ordinary RGB images and infrared recordings, with a resolution of 25 cm from the National Geo Register were used, covering the years 2016-2018. The

Regentessekwartier district in The Hague was chosen as a test bed. Training and test sets were created for this district. Different variants of neural networks were applied to the available RGB photos. It appeared necessary to look critically at the limitations of the methods and the influence of natural phenomena like shadows and withered green surfaces. Moreover, parts of surfaces were found to be hidden behind tall buildings in photos taken at an angle.

Figure 2.4-1
The district used for detecting green and gray. Left the training area, right the (adjacent) testing area

In addition to the AI method, a more classical method based on infrared photos was used. This method has previously been used in the UK by the Office of National Statistics. The red color areas of the infrared photos (caused by chlorophyll) were identified using an automated filtering method. This method turned out to be about as accurate as manual annotation, which is obviously much more time consuming.

The model results gave a green area of 17% for both 2016 and 2017, which had increased to 23% in 2018. Random manual inspection of subareas (128x128 meters) showed that the results were largely correct. In part, however, the increase was found to be due to artifacts that eluded proper detection such as withering (the summer of 2018 was very hot). The results from neural networks and RGB photos turned out to be quite useful, but the more classical approach with filtered infrared photos worked all in all slightly better. The neural network approach, however, is promising in that it allows for performing more complex analyses, such as recognizing more types of greenery. Further research is needed to determine whether these statements hold up more broadly. The technical research report (Dutch only) provides further details and is available upon request.

2.5 Victims of high-impact crimes

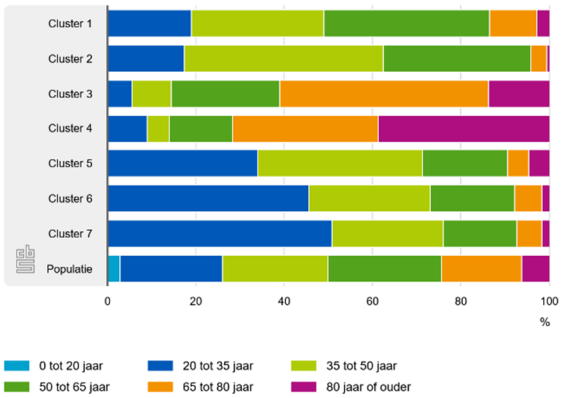
High impact crimes is a collective term for offenses that have a major impact on victims and their immediate environment. Examples of high impact crimes are burglaries, street robberies and violent crimes. The Dutch Ministry of Justice and Security is strongly committed to combating and preventing high impact crimes. In order to develop targeted policies they need better information. The Institute for Financial Crime asked CBS to work on a joint pilot study for the Ministry to investigate possibilities to add data-driven knowledge to current trend reports.

The pilot study, as described here (Dutch only), focused on similarities between victims of domestic burglary. As a study population, Dutch adults registered as victims of domestic burglary in 2017 were chosen. Background characteristics of these persons and their homes were added from various administrative registers available at CBS. In addition, various criminological variables were included in the study, such as distance from the victim of a domestic burglary to the suspect (if known), criminal history of suspect and victim, and the residential location.

Theoretically, there is no unambiguous classification for dividing victims of domestic burglary into homogeneous groups. Such a classification, however, could be of value to policymakers if we can construct it. Therefore we decided to look into cluster analysis, an unsupervised learning method widely used in big data research. The aim of this method is to find groups (clusters) of people that are as similar as possible in terms of characteristics. A feature of the cluster analysis method is that no substantive knowledge is used to form the clusters: the data must speak for itself.

The pilot study was mainly concerned with determining the possibilities and possible broader applicability of the methodology, rather than developing new methods. A key question concerned the translation of results to non-statisticians. Seven homogeneous clusters of victims could be distinguished, each containing about 10 to 20 percent of all victims. Given the large number of variables considered, it was not clear at a glance which characteristics are distinctive for each cluster. Can the abstract results be made interpretable for domain experts and thus provide insights for policy issues? In order to gain better insights, visualizations were made for each characteristic. The figure below shows e.g. the distribution of age categories for each cluster compared to the whole adult Dutch population.

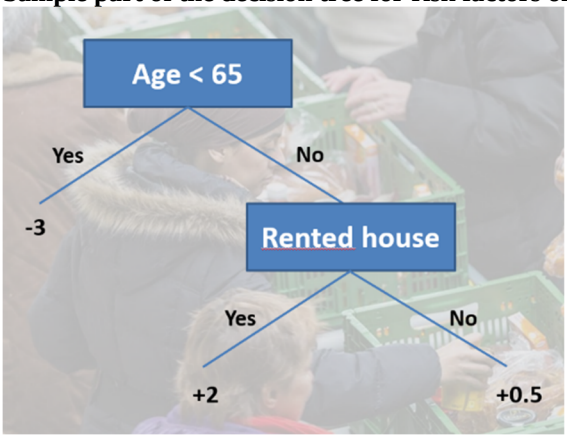
Figure 2.5-1
Clusters of victims of burglary by age group distribution, compared to the distribution of the adult Dutch



2.6 Predicting the risk of poverty

With a view to effective poverty policy, CBS has been commissioned by the Dutch Ministry of Internal Affairs and Kingdom Relations to investigate how well it is possible to estimate the probability that someone will remain poor or become poor on the basis of (combinations of) characteristics, and for which aspects the combination of population-wide register data and machine learning can lead to new insights about risk factors for transitions into and out of poverty. To this end, more than 500 potential risk factors for poverty in the fields of demographics, economics, health and crime have been examined based on data from 2011–2018. The study was carried out together with the Ministry and the cities of Amsterdam and Heerlen, where both statisticians and policy makers were involved. The results of the study are presented here (Dutch only).

Figure 2.6-1
Sample part of the decision tree for risk factors on poverty



Given the complexity of the problem and the large number of variables involved, an AI method was developed. The method is based on XGBoost, a well-known decision tree model. This was combined with a global surrogate model, to get better insights in the complexities and structure of the overall results, and SHAP, a game theoretic approach to

estimate the contribution of individual factors to the overall risk of poverty. All in all, the method performed quite satisfactorily for aggregate results, with the caveat that the approach seems less suitable for predictions at the individual level- which was never our goal as statisticians. Several roads for follow-up research are being explored.

3. AI monitoring

Artificial Intelligence is expected to have a major impact on economy and society. The Netherlands therefore has the ambition to shape a leading ecosystem for AI research, development and implementation. AI is among the emerging technologies stimulated by the Growth Fund, a large investment fund aimed at stimulating the earning capacity of the Dutch economy. In a first installment, grants amounting to EUR 276 million have been allocated for AI by the Growth Fund, which are expected to be matched by the same amount from other stakeholders. A dedicated new foundation called AI-Ned (= AI in the Netherlands) is being established to channel the available funds and report on their use and impact. Adequate monitoring was an explicit condition from the Growth Fund, where specific attention should be paid to guaranteeing a human-centric approach to AI.

In short, AI-Ned requires monitoring of the state of AI in the Netherlands, to account for its activities and to determine action points for interventions and adjustments. To fulfil its reporting obligations, the AI-Ned foundation has asked CBS to develop, together with the research organization TNO, a proposal for a comprehensive and nation-wide AI monitor. The purpose of this monitor is to gain insight into the speed, direction and impact of the various AI developments on the economy and society, including insights into the structure and dynamics of the AI ecosystem. A main reason for involving CBS is that its institutional position allows to set up and maintain an authoritative, independent and impartial central information source that is easily accessible for all interested users.

More specifically, CBS and TNO were asked to provide coherent quantitative information on four key aspects:

1. Key Performance Indicators of the AI-Ned Foundation
2. The impact of AI on the Dutch economy and society, in a broad sense
3. The structure and dynamics of the Dutch AI ecosystem
4. The Dutch AI activities in an international perspective

In order to achieve the desired results, a coherent conceptual indicator framework has to be designed that provides comprehensive insights into the state of AI in The Netherlands. The draft framework should be discussed with key stakeholders, to check it satisfies their needs and guarantee that it actually contributes to insights for interventions and adjustments. If not, there should be room for modification. Next, the indicator framework should be populated with available data. Some indicators may already be available. For example, as part of its existing publications on the digital economy, CBS has developed statistics on how AI is applied in, and developed by, the business community. Not all data has to be provided by CBS. For example, it is expected that the publishing company Elsevier/RELX will contribute data on scientific achievements from their repositories and expertise. It may, nevertheless, appear that not all desired indicators can be adequately measured from existing sources and methods. Then choices must be made whether or not to fill gaps, using new methods or tapping into new and/or alternative data sources.

Preparatory studies for the AI monitor were launched end-2021, while a first release of the AI monitor is foreseen for mid-2022. In our view, good monitoring requires more than just supplying figures. That is why we also devote a great deal of attention to interpretation of the statistics in the AI monitor and background articles to put them in context.

4. Fair data and fair algorithms

When working with AI, especially in domains that directly affect humans, ethical questions come into play. This is further strengthened by the European and Dutch focus on human-centric AI with clear call to use AI in a way that observes fundamental human rights. When allocating funds to AI, the Growth Fund mentioned above therefore emphasized the importance of so-called ELSA labs, dealing with Ethical, Legal and Societal Acceptance of AI. This connects very well with the core values of CBS, and, for that matter, official statistics. Therefore CBS is already participating in three ELSA lab proposals and seeks to become a member of the ELSA board. CBS is also investigating

aspects of accountability, explainability and transparency in relation to AI methods. A particular aspect that CBS has paid particular attention to is fairness. Below we describe the CBS perspective on fair data and fair algorithms, not only in its own statistics production but especially in supporting other government bodies.

Data is always at the heart of AI; After all, algorithms have to be trained with data. To get a fair algorithm, the training data must be representative and free from bias whenever possible. An algorithm reflects the patterns in the underlying data and thus it is essential to ensure that the data is as fair as possible- or at least sufficiently understood such that bias and selectivity effects can be treated and neutralized. This is an observation that is close to trivial for a statistician, but general awareness of the importance of data used in algorithms is only just emerging. With its statistical knowledge on data and data handling, based on scientific insights, as well as its formal position articulated in the Dutch Statistics Act, CBS can play an important role in the AI domain especially when it comes to understanding the data used.

In Dutch society, more and more decisions are being made by (semi-)automated systems that use algorithms. Local authorities and other government bodies use them in many different areas of life: for example, to assess the likelihood of a student dropping out of school before they complete their education. It is essential for the social acceptance of the use of AI by governments that algorithms are transparent and explainable. This also applies to the algorithms used by CBS. But how do you make sure that this approach does not lead to unfair situations? And how can you explain the workings of such an algorithm? What is or is not fair and desired behaviour of an algorithm may be the subject of a political, administrative and/or societal debate and thus eludes statistics. What statistics can do, however, is contribute to a better understanding of the underlying mechanisms, e.g. how a model behaves and in particular how data influences the outcome of a model. CBS has therefore actively explored the question of algorithmic fairness in connection with related issues like accountability, explainability and transparency. Some excellent studies that also attracted attention from media and others were carried out by students doing an internship at CBS, see e.g. this article and references mentioned there.

A key project in this area carried out by CBS concerned research into fairness of AI methods in the social domain, e.g. for detection of fraud with social benefits. In this project, CBS developed an ‘AI starter kit’ to support the dialogue between stakeholders, such as IT specialists and policy makers, as well as a dashboard to test algorithms against formal fairness criteria. The project attracted quite some attention and increased awareness of algorithmic fairness. At the same time, it gave rise to an internal discussion at CBS on what exactly should be the role of a statistical institute in the societal debate around ethical AI use. The discussion has not finished yet at the time of writing and is likely to continue for some time, given the complexity and rapidly evolving nature of the topic. CBS is now developing corporate guidelines on what we consider the scope of our statistical domain, acknowledging that there still is a large grey area. One example of a guideline is that CBS does not consider it part of its role to conduct audits or provide formal approval on (AI-based and other) algorithms developed by and for government bodies. On the other hand, CBS is willing to provide methodological advice and support for government algorithms in selected cases, in particular when it comes to help understand quality aspects (including bias and selectivity) of the data being used.

5. Data sharing and synthetic data

Contrary to older ‘rule-based’ forms of AI, modern AI is data-driven and heavily depends on available data. Consequently, there is a large interest in making as much data as possible available for AI use. At the same time, many relevant data sources cannot be freely shared. Commercial interests, privacy issues, the sheer size of big data and other concerns stand in the way. There is a clear need for suitable methods that allow seamless access to and sharing of data, under controlled conditions, as a key enabler for developing and deploying modern AI applications.

Privacy is always central to CBS's work. Society trusts CBS with its data, because it trusts the way in which it is handled. All (survey, register and other) data that arrives at CBS is immediately pseudonymized. Only after this process do researchers and statisticians start working with the data. So they never see data from individuals without a pseudo-key, in accordance with CBS guidelines. Data published by CBS is always aggregated to a level at which it is no longer possible to identify persons or companies. CBS demonstrably complies with applicable legislation and guidelines, and it is actively committed to both enabling data use and protecting confidentiality as good as possible. CBS is therefore investigating the possibilities of advanced Privacy Preserving Techniques, in collaboration with TNO, universities, the international statistical community and other partners. Eventually, these state-of-the-art

methods should allow e.g. remote execution of sophisticated algorithms, including record linkage of physically distributed data sets. A clear driver for this area of research is the foreseen demand from an AI perspective.

Another approach for making more data available and shareable for AI is by creating synthetic data sets. This typically calls for a trade-off between usability and confidentiality protection, depending on the use case at hand. Also in this area, CBS is collaborating with a broad range of partners. AI plays a double role here: on the one hand, synthetic data may be used in AI applications but on the other hand, AI methods such as generative adversarial networks or GANs (the technology behind deep fakes) may be used to may be used to generate synthetic data sets.

6. AI collaborations

CBS is actively seeking collaborations in the field of AI. The examples presented above already show this. The focus is clearly on collaboration within government, where CBS aims for a role as data hub and data partner. We do not, however, *a priori* restrict which partners we allow and pursue a 'triple helix' strategy, where we collaborate with government bodies, knowledge institutions and the private sector depending on the situation- even broadening to a 'quadruple helix' approach when collaboration with individual citizens, e.g. through citizen science, comes into play.

An example of a successful partnership within government is the Fair Algorithms project mentioned above. This project was carried out in collaboration with the city of Amsterdam and several other cities, the Union of Dutch municipalities, the University of Amsterdam and Code for NL (a Dutch network of civic tech professionals and enthusiasts), and funded by the innovation budget of the Ministry of Internal Affairs and Kingdom Relations. Each of the six projects described in Chapter 2 also involves some form of collaboration.

CBS actively participates in the Dutch AI coalition (NLAIC) and is a member of the strategy team of the NLAIC, formally elected as representative of the public sector. The NLAIC is a public-private partnership in which government, industry, educational and research institutions as well as civil society organizations work together. The aim of the coalition is to stimulate, support and where necessary organise Dutch activities in AI. The NLAIC wants to put the Netherlands in a vanguard position in the field of knowledge and application of AI for prosperity and well-being. The NLAIC functions as a catalyst for development of AI applications in our country. In addition, the NLAIC was instrumental in taking the initiative for the AI monitor mentioned above. CBS contributes to a number of NLAIC working groups and is part of the leadership in those on public services, data sharing and human-centric AI.

In addition to national activities, which are often (but not always) channeled through the NLAIC, CBS participates in international AI activities both inside and beyond the statistical community. As an example of the latter, CBS is partner of the TAILOR network funded by the European Commission; TAILOR stands for Foundations of Trustworthy AI-Integrating Learning, Optimisation and Reasoning. As part of this network, CBS has co-organised a (first) Theme Development Workshop on the Public Sector that should feed into the AI Roadmap for the European Union. It goes without saying that official statistics gets attention on this roadmap. As an example of collaboration inside the statistical community, the work on Machine Learning under the umbrella of the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) should be mentioned. This work started with a position paper prepared by the statistical institutes of Canada, Finland, Italy, Mexico and The Netherlands; then led to a formal two-year HLG-MOS project and continues as a global Machine Learning Community led by ONS from the UK.