Background document on the Machine Learning Project

Prepared for the 2019 Workshop on the Modernisation of Official Statistics

Written by Claude Julien (project manager) with assistance from numerous team members

Executive Summary

One of the biggest threats to National Statistical Organisations (NSOs) producing official statistics is that of quickly becoming irrelevant in today's fast-paced and ever increasingly complex society, economy and environment. NSOs are threatened by an increasing number of public and private organisations who produce and promote statistics in a more timely and accessible manner, attracting the attention of policy makers and other users. These data producers use approaches and methods beyond those traditionally used by NSOs,such as machine learning and artificial intelligence. In fact, these approaches and methods are no longer that new anymore and the importance of integrating them into the production of official statistics has now been widely recognized by NSOs. At its November 2018 workshop, the HLG-MOS further brought this point forward by supporting a proposal from the Blue-Sky Thinking Network to launch a Machine Learning Project.

Based on mutual interest and building on existing national developments, the objective of the project is to advance the research, development and application of machine learning techniques (ML) to add value (relevance, timeliness, quality, efficiency) to the production of official statistics. Building on the engagement of 38 participants from 18 organisations in 14 countries, pilot studies are being conducted to demonstrate the value-added of ML; identify best practices in the implementation of ML techniques; and share knowledge and IT tools. The pilot studies are conducted on specific statistical business processes (coding, editing and imputation) and the use of non-survey/census source of data (satellite images). Some of the studies are used to test some of the features of a quality framework that is being developed to underpin the best practices in the integration of ML solutions.

Since being launched in mid-March and fully functioning in May, over 40 documents (presentations, working papers, ML scripts, software documentation) have been shared within the team. Most of the work was discussed a sprint held in September and now continue to be shared and discussed at monthly meetings. The work of the team is also shared with 30 other individuals from 14 national or international organisations who either support it or wish to follow its progress.

One of the recurring themes from the discussions is that integrating machine learning into official statistics requires more than simply building machine learning systems. In fact, a number of participants noted that they had already developed otherwise successful machine learning solutions, but had been unable to implement them into production processes because of a variety of organizational and structural impediments including uncertainty over who should be responsible for building, evaluating, and maintaining these highly interdisciplinary systems. If extended for a second year, the project would explore how different NSOs are organized to integrate ML in their production processes, and report on the different practices, sources of impediments and propose successful practices.

In six months, the project team has collectively gained and shared considerable knowledge in how and where ML can add value in the production of official statistics. The engagement and enthusiasm of the participants has generated momentum that will, if the project is extended for a second year, lead to the delivery of the results of numerous studies, summaries and recommendations on the value added of ML, best practices in the implementation of ML, documented ML scripts, a review of challenges and practices in the integration of ML in NSOs and a workshop to further share the great work of the team. All this supported by a quality assurance best practices and reference documents gathered along the way.

1. Background

The mission of the HLG-MOS is to work collaboratively to identify trends, threats, and opportunities in modernising statistical organisations. One of the biggest, if not the biggest, threat to statistical organisations that produce official statistics is that of quickly becoming irrelevant in today's fast-paced and ever increasingly complex society, economy and environment.

National statistical organisations (NSOs) are being challenged to be more responsive to the increasing need for more relevant, timely, detailed and accessible statistical information and data services that can be trusted and used to make clear and effective data-driven policy decisions. NSOs are also pressured to meet these expectations highly-efficiently within existing budget levels. They are also challenged by the ever-increasing amount of data available in a wide variety of sources, formats and level of quality.

NSOs are finally threatened by the presence of an increasing number of public and private organisations, big or small, who produce and promote statistics in a more timely and accessible manner that attracts the attention of policy makers and many other users, in spite of relevance or quality deficiencies, at times. These organisations are able to produce these statistics for various reasons including: development or quick access to alternative data method and their adoption in the production process, such as machine learning algorithms; greater IT capacity (space and power); and by imposing fewer constraints on quality, transparency, ethics and privacy.

At the same time, these are the areas where NSOs hold a competitive advantage. They are more transparent by publishing details on data sources, methods and various indicators. They have considerable collective expertise in efficiently integrating different sources of data and they have a legal obligation to respect privacy and protect against disclosure. Furthermore, the capacity to do so not only lies within each NSO, but increasingly through a network of professionals around the world that are brought together through collaborative initiatives such as the HLG-MOS.

In addition to counting their individual and collective expertise, statistical organisations must have an adaptive culture to remain relevant by responding to the timely data needs of stakeholders, both in terms of statistical information and services, in a continuously responsible manner. One of the pillars on which NSOs have fulfilled their mandate is in the development of expertise in sound methods and processes integrated in the production of official statistics. As mentioned above, this pillar is challenged by the proliferation of data demands, data sources and data producers who are increasingly able to more easily link the sources to the demands.

Most of these data producers use different approaches and methods than those traditionally used in statistical organisations. Many of the data sources require these new approaches and methods, for instance, such as machine learning and artificial intelligence. However, they are not that new anymore. The importance of bringing them into the production of official statistics has now been widely recognized by statistical organisations. At its November 2018 workshop, the HLG-MOS further brought this point forward by supporting a proposal from the Blue-Sky Thinking Network to launch a Machine Learning Project. The BSTN's position paper justified the project as follows:

> “The interest in the use of Machine Learning (ML) for official statistics is rapidly growing. For the processing of some secondary data sources (including administrative sources, big data and the Internet of Things) it seems essential to look into opportunities offered by modern ML techniques, while also for primary data ML techniques might offer added value, as illustrated in the ML position paper mentioned above. Although ML seems promising there is only limited experience with concrete applications in the UNECE statistical community, and some issues relating to e.g. quality and transparency of results obtained from ML still have to be solved.”

To maximise what ML can bring to statistical organisations and the speed at which it can be brought, one could state this purpose much more boldly. Interest in ML is beyond “rapidly growing”. Interest in ML is one of many elements that will ensure the long-term relevance of statistical organisations by enabling them to produce better information more quickly, while potentially using less of the resources at their disposal. It is not a question of “looking into it”, but rather how best to integrate it in the production of official statistics just like any other sound method. The question is not if it can be adopted, but (1) where can it be integrated to quickly replace current processes, especially manual ones, (2) where could it be integrated to support or complement current processes to improve them and (3) where is there no choice but to integrate it (then becoming a question of accessing and integrating it quickly).

The BSTN’s observation that “there is only limited experience with concrete applications in the UNECE statistical community” is accurate. It may be partly due to the lack of boldness in considering ML or any other similar significant change, which is not atypical of statistical organisations. This could be mostly due to a lack of understanding and concrete demonstrations of applicability in the context and culture of producing official statistics. While it is essential to have an agile adaptive culture to survive, statistical organisations must implement change in a *responsible* manner to maintain trust in its products and services. The challenge is to do this in a timely manner to remain relevant.

2. HLG-MOS Machine Learning Project

The project proposed by the BSTN was strongly supported by the HLG-MOS. It included four work packages:

1. Refine the scope and define the business case
2. Pilot studies
3. Quality issues
4. Lessons learned

The project manager was hired and the first team members (11 from 6 organisations) joined in March. A kickoff meeting was held virtually on March 19. A first virtual sprint was held on April 4 at which time the team quickly had grown to 23 participants from 13 organisations, demonstrating the interest in this type of development. The focus of the sprint was to initiate the planning the project. Most of the participants had little to no experience with machine learning. On the other side of the ML maturity spectrum, one participant had gone through the development and full implementation of machine learning in classification and coding.

The members were polled on what they and their organisations expected from the ML project. The most common words mentioned were around learning, sharing, applying and quality. This supported the observation made by the BSTN on the keen interest in ML, yet with very low adoption in the production process also supported the need to collaborate on pilot studies to facilitate learning and sharing.

2.1 Scoping and Planning

Based on the initial work of the BSTN, the expectations of the HLG-MOS and the interests of the team members, it was decided to conduct pilot studies on the following themes (work package 1): coding and classification (C&C), edit and imputation (E&I) and the use of imagery. It was also decided to develop a quality framework to underpin the use of ML (work package 2). It is important to note that the pilot studies were needed to facilitate learning, sharing and application of ML within the group. As important as the knowledge gained and shared by the team members are the conclusions, recommendations and ML scripts on each pilot study theme and the best practices, quality framework and organisational considerations needed to develop and implement ML solutions that will be gathered and shared to all organisations.

A first face-to-face sprint was hosted by the Office for National Statistics in the UK from May 13 to 15 at the London and Newport offices. At that time, the project team had grown to 27 participants from 14 organisations. The sprint was attended by 12 team members from 9 countries. Several presentations were given on the current status of ML projects in participating organisations. The main objectives of the project were discussed one last time and were agreed upon. They are as follows:

Based on mutual interest and building on existing national developments, the objective of the project is to advance the research, development and application of machine learning techniques to add value to the production of official statistics. To achieve this objective the Machine Learning (ML) project team will aim to:

- *Investigate and demonstrate the value added by ML in the production of official statistics, where "value added" is measured as an increase in relevance, better overall quality or reduction in costs.* (through level 1 knowledge mentioned above)
- *Advance the capability of ML to add value to the production of official statistics.* (level 2 knowledge)
- *Advance the capability of national statistical organisations to use ML in the production of official statistics.* (level 3 knowledge)
- *Enhance collaboration between statistical organisations in the development and application of ML.*

The participants confirmed the relevance of the three pilot study themes. In their opinion:

- These topics are already being considered in many organisations
- Two themes cover existing production processes (C&C and E&I)
    - C&C is considered more readily applicable use of ML; the main question is why it hasn’t already been used more widely
    - E&I is considered a potential use of ML; the first question being to determine if it adds value compared to other existing processes

- The other theme covers the use of less traditional data, i.e. satellite or aerial imagery
    - In this case, ML is considered essential to make productive and efficient use of the data source; the main objectives are making ML solutions more accessible while demonstrating how it enables the use of imagery data to produce new statistical information

The scope and activities of the work packages and pilot studies were also discussed and agreed upon at the May sprint. The work of the project began.

2.2 Project Participants

Before reporting on the progress of the project, it is important to introduce its team members, their organisations and their expertise. Since the beginning, interest in the project continued to grow, reaching 38 participants from 18 organisations in 14 countries. One organisation joined the project as a follow-up to a topic that was suggested by the BSTN.

As they joined the project team, participants were asked about their main area of expertise and other areas of expertise. While the most common main area of expertise within the group is statistics, when all areas of expertise are combined, the team brings together a rich mixture of talents needed to advance the use of machine learning (statistics, informatics, subject-matter analysis and data science).

In addition, 31 persons from 14 organisations and 9 countries are either supporting the work of team members or are following the developments of the project team.

In addition to participating directly or indirectly on the project, several team members, additional collaborators or followers are involved in other international activities that touch on the use of machine learning. A list of these activities is provided in Appendix 1. These connections allow the project to be kept informed of the developments from other groups and vice-versa.

3. Project Progress[1]

Soon after the May sprint, the scope and plans of several pilot studies were set and work teams were formed. Their work is coordinated by pilot study theme leaders (C&C, E&I and Imagery) and work package leaders (pilot studies, quality and lessons learned). The leaders report on a monthly basis at a Webex meeting at which team members, followers and additional collaborators are invited. The project manager communicates project updates and action items.

The following countries are collaborating on the pilot studies:

- Coding and classification on product descriptions – Poland, USA, Canada, UNECE
- Coding and classification on industry – Serbia, Mexico, Canada, Norway, Australia
- Coding and classification on Web sentiment – Belgium, Poland, Mexico
- Coding and classification Quality Assurance – USA, Canada, Mexico

[1] Note that this document in meant to describe the progress of the ML project and not to provide complete or, even, accurate technical details on each of its pilot studies. These will be provided in separate reports.Some technical shortcuts are purposely taken by the author.

- Edit and imputation – Germany, UK, Italy, Poland, Belgium
- Imagery – Mexico, UNECE, Australia, Netherlands, Switzerland
- Quality – Canada, Germany, Australia, Italy

After the summer period, a face-to-face sprint took place from September 18-20 in Belgrade. It was attended by 16 team members from 11 countries, representatives from Russia and the Czech Republic, and several employees from the Statistical Office of the Republic of Serbia. The sprint was also attended by at least 8 other persons at different times by Webex.

Fifteen presentations and four working papers highlighted significant progress achieved on many fronts in a relatively short period of time. The discussions during the presentations and throughout the meeting facilitated collaboration and led to significant advancement on some studies and initiated other collaborations. In addition, now that the project is at full speed (given the time that participants can dedicate to it), it was decided to add a presentation and discussion on one pilot study at the end of the monthly progress meeting. One took place in October and another is scheduled for November.

3.1 Classification and coding

Among the organizations participating in the project the only ML application currently being used for text coding and classification in production is for the **US Bureau of Labor Statistics** (BLS) Survey of Occupational Injuries and Illnesses (SOII). This survey involves collecting hundreds of thousands of written narratives describing cases of occupational injury and illness each year and then assigning 6 codes to each to indicate various characteristics of the incident. The development of the ML application started in 2012 and has been gradually implemented and improved. It was put in place to improve the quality of the coding in comparison to the existing manual process. Each year, the ML system is trained using all previously coded data, except for a small representative sample of narratives that are coded by experts. This gold standard data set is used to assess both the accuracy of the ML system and the manual coding as well. The latter is a key element in demonstrating the quality of the ML system.

As narratives are collected, the ML system evaluates each narrative and estimates the probability that each possible code is correct. If the probability for the highest probability code exceeds a predetermined threshold, the ML system assigns that code, otherwise the characteristic is left for manual coding. The threshold is set using the ML and manual coding accuracy rates[2] on the gold standard dataset to ensure that, on average, the accuracy of the ML assigned codes are higher that the manually assigned codes on the subset of data being autocoded. Through continuous improvements to the ML system over the years, the percentage of codes that are assigned by ML system (i.e. automated coding) has reached 81%[3] and have been proven to make 40% fewer errors than humans on the same narratives. Among the characteristics that are coded, the percentage of assigned values that

[2] A code assigned by a coder or ML is accurate when it agrees with the value assigned by experts, assumed to be the truth. In some instances, the disagreement (or error) is due to ambiguous input data.

[3] “BLS use of autocoding has expanded significantly over time. In 2014, only 5 percent of codes and only occupation codes were assigned by machine learning. By 2018 automatic coding had been expanded to include all five primary coding tasks (occupation, nature, part, source, and event) with the model assigning approximately 81% of these codes.” Ref: https://www.bls.gov/iif/autocoding.htm

agree with the experts range between 52% to 84% from coders, and between 70% and 92% from ML.

It is important to note that all narratives still receive some manual coding because one of the characteristics is not currently processed by ML and must therefore be coded manually. Manual coders are instructed to review and correct, if necessary, all ML codes when completing the coding of the narrative. The percentage of ML codes that agree with the coders during the review range from 92% to 98%. As the ML system and its accompanying quality assurance processes stabilize, the review requirements will likely be relaxed.

Starting from the BLS knowledge and experiences, ML scripts and advice were shared with other organisations. Statistics Poland and Statistical Office of the Republic of Serbia (SORS) are experimenting it to code product descriptions to ECOICOP and business activity to NACE, respectively.

The pilot study from **Statistics Poland** is conducted on 17000 product descriptions scraped from the Internet. They are coded to 61 ECOICOP categories. The data has been anonymized and shared with team members. It has been processed by different organisations through several ML methods, achieving accuracy rates up to 92.6%.

Furthermore, the dataset, originally in Polish, has been translated to English and French and also coded through several ML methods, still achieving accuracy rates in the very high 80% (refer to the table in Appendix 2). The purpose of this extension is to assess the feasibility of creating a common dataset in different languages and providing the means of experimenting different ways of coding it using ML as a learning tool. This learning package could be initially delivered through a workshop. Statistics Poland have also developed an interactive coding tool that could be useful in delivering this workshop.

The pilot study conducted by Statistics Poland was further extended in collaboration with product experts to include descriptions received directly from stores. The descriptions were coded using the same ML approach to over 100 categories and achieved accuracy rates in the high 80%.

The **Statistical Office of the Republic of Serbia** is experimenting on coding descriptions of economic activity to NACE on over 17000 records. Their first attempts achieved an accuracy rate of 74% at the 2-digit level and 63% at the 3-digit level. The results indicate the potential of using ML to at least partially automate their current manual coding process or increase the accuracy of their coded data (like at the BLS above). Further work is being conducted to improve the performance of the ML solution and determine the amount of manual coding required to ensure the same level of quality or better.

While Statistics Poland and SORS are experimenting ML coding to potentially replace or assist current manual process, other organisations on the project are investigating or implementing ML solutions in operations currently using index-based automated processes that require up-to-date reference files. For example, the current automated system used at **INEGI** is able to code 95% of most of their variables with an accuracy rate of 99%. However, they can only code 76% of industry and occupation descriptions with 95% accuracy. A first attempt at coding all descriptions with ML achieved 87.7% accuracy on industry and 82.0% on occupation. Refer to the table in Appendix 3. However, to achieve a similar level of accuracy as the current process, the ML solution can only code 60% of the industry descriptions. They expect that, once ML is improved and adapted to their production process, coding time could be reduced by up to 50 percent. To confirm its value-added in terms of costs and quality, they are considering using ML as a quality corroboration mechanism in their 2020 Population and Housing Census.

Statistics Flanders (Belgium**)** is looking into the use of social media data, as one of the big data sources, to measure the sentiment of a population (e.g. concerning their perceived quality of life, 'beyond GDP') as a valuable alternative for survey questions. Building on the work from Statistics Poland in the context of the Eurostat Big Data ESSNet, Statistics Flanders is experimenting with ML to categorize tweets. One of their challenges is to integrate in their ML scripts a way to make a distinction between Dutch and Flemish tweets. They are testing several ways to create test and training data sets and should complete this by the end of the year. A coding GUI has been developed in Python. Finally, they have hired a ML specialist to set up a Data Science Hub. The specialist will also assist them in applying alternative ML algorithms to the sentiment case.

On another front, **Statistics Canada** has shared the code and documentation on their generic coding system (G-Code) in which it has recently introduced an ML solution (FastText). **Statistics Norway** is experimenting with a similar solution with limited success so far, but, through some collaboration with Canada have been able to achieve better results. The **Australian Bureau of Statistics** have also shared information on their Intelligent Coder using ML (model-based) approaches.

3.2 Edit and imputation

The **Office for National Statistics** (ONS - UK) is in the process of combining three surveys:

- Living Cost and Food – LCF
- Survey of Living Conditions – SLC
- Wealth and Asset Survey – WAS

to form a much larger survey, the Household Financial Survey - HFS.

The LCF currently run an extensive manual editing process where every household and person level data is examined for inconsistency, error and missing data. All this is then manually corrected to form the post-edit data set. The SLC and WAS run an automated editing process to detect outliers and inconsistencies.

A new editing solution for income data has to be built for all 3 surveys to form the HFS survey. The system currently used for the LCF can not be upscaled and the SLC/WAS method is deemed not to be accurate enough. ML is being tested to assess its ability to predict LCF records that actually require a correction of income data rather than review every record. The test is based on prior LCF manually edited records.

Up to now, the ML setup has been able to predict 85.7% of the records that required some change (i.e. 85.7% of 17%), while reducing the percentage of records to manually review to

31.3% rather than 100%. With some tuning, the ML solution can predict 100% of the records that required some change, but the review rate would then increase to about 42%.

Furthermore, with ML a decision tree can be visualised that describes edit rules for the prediction of a required review. These rules can then be used to explain to subject matter experts what the ML system is doing. On one hand this can provide some reassurance to the experts; one the other hand some of the derived edit rules may not make full sense to them. This situation can be mitigated by limiting the size and complexity of the decision tree, but that can severely reduce the predictive power of the ML solution.

The **Italian National Institute of Statistics** (Istat) has analysed the potential use of ML for the specific area of editing, meant as the detection of error phase. So far, there are not so many applications in this area and the analysis gathered some ideas and/or hints about the potential us according to general agreed Generalized Statistical Data Editing Model. It can be considered as a base to propose a framework to enhance tests in this area. It has been shared with the ONS to support their pilot study described above.

ISTAT is also experimenting with ML techniques for the imputation of the Attained Level of Education (ALE) in the Base Register of Individuals (BRI). BRI is the result of the integration of data from different sources and it is the basis of the next Permanent Italian Census. The output of the model is a probability distribution for the 8 ALE classes. Much like in coding, for imputation, ML proposes values and their respective probability of being accurate. The experiment compared the imputation using the most probable value (MPV) to the random selection among all probable values (RV). Results are very promising: MPV results with a high predictive accuracy (82%) but a not satisfactory distributional accuracy, while RV gives origin to a lower predictive accuracy (72%) and a better distributional accuracy.

In comparison to standard techniques such as Log Linear models, MV produced more accurate predictions and more accurate distributions on the whole population (approx. 309,000 persons). The table in Appendix 4 provides accuracy measures for the whole population and five sub-populations. RV imputations show higher prediction accuracy in 5 of 6, and higher distributional accuracy in X of 6. Furthermore, ML techniques allow a more automatic imputation process.

The **Federal Statistical Office of Germany** (Destatis) is conducting tests and sharing results on different machine learning methods for regression based imputation. Their experiments also show several promising but prima facie counterintuitive results. They observe that their method, which is different from the MLP experimented by Istat, reach good results in regression imputation (without random effect) in terms of predictive and distributional accuracy (i.e. there seems to be no trade-off between the two anymore). These results and resulting questions motivated them to motivated them to seek collaboration from Statistics Netherlands (CBS) where similar results have been observed.

Statistics Poland is experimenting with several standard and ML methods on data from their tourism program. They are conducting their test on financial data (expenditures) and counts of people (e.g. members of sport clubs). These data are currently either imputed manually or not at all. Their tests also show that ML methods produce more accurate predictions.

The Flemish Institute for Technological Research (VITO – Belgium) in experimenting with ML on an imputation project where they want to fill in energy consumption per sector/carrier for the Energy Balance of Flanders of the current year with forecasts based on over 50 economic indicators. Because data for different sectors and energy carriers are received at different times, early estimates require missing data to be imputed. Their first results using several ML approaches have demonstrated the value of the approaches that can use the full set of economic indicators rather than a subset of indicators with higher correlations. However, even these better ML approaches are slightly worse than simply predicting the current quarter consumption with the same quarter the previous year. More models will be tested, including more standard ones. Since the study is currently conducted using publicly available aggregate data, it has been proposed to call on other members within the team to run other models and share the experience.

In summary, the assessment of the accuracy of the ML techniques tested for edit and imputation are more positive than initially expected. Further investigations are needed to confirm and complete the assessment of the value added of ML in the edit and imputation process:

- How to assess the accuracy of imputation in a proper way?
- What is the best way to measure distributional accuracy that takes into account not just large populations, but smaller and more specific subpopulations as well?
- How often constraints on the variables are not respected when using regression imputation, e.g. predicting 105 hours of work per week
- ML is often expected or believed to be less costly to setup and maintain than current standard approaches. This needs to be assessed and confirmed.
- The team is collectively experimenting a variety of ML methods for imputation on a different types of data indifferent contexts. In contrast, the project would need at least another pilot study on editing to build on the study being conducted by the ONS.

3.3 Imagery

As mentioned in section 2, ML is considered essential to make productive and efficient use of huge sources of data, such as satellite imagery. However, compared to the traditional data where production process is already well defined, the scope where ML can be applied to satellite data is not so clear because workflow needed to use these types of new data is not clearly understood. Satellite images are highly complicated data products that require several steps of pre-processing and it is difficult to grasp where the involvement of data scientists/statisticians starts, joins and ends. To address this problem and thereby facilitate the use of ML for satellite imagery, the project is conducting the following activities:

The **Instituto Nacional de Estadística y Geografía** (INEGI - Mexico) and the **UNECE** have collaborated in drafting a pipeline that describes the process needed to use satellite data for statistical production. It generally follows the phases of the Generic Statistical Business Process Model (GSBPM) and is composed of 13 activities (refer to Appendix 5). The activity in the Modelling phase, where ML is mostly used, will be further expanded with the developments of the other pilot studies and the work package on quality. In additional to better understand the whole process needed to use satellite data for producing official statistics, the pipeline will also propose a common language to facilitate collaboration among experts (e.g. statistician, EO expert) in different fields and sharing of knowledge and experiences among different organizations.

INEGI is also experimenting ML approaches that combine Landsat satellite data and census data to build a model to monitor the growth of cities of Mexico (Urban and Rural), which would generate more timely input for :

- Cartography update (more efficient prioritization)
- Estimation of the population in non-census years
- Statistics related to SDG Indicator 11.3.1(Ratio of land consumption rate to population growth rate) and SDG Indicator 15.3.1 (Proportion of land that is degraded over total land area).

The objective of the ML model is to classify 1 km by 1 km land parcels as urban, rural or neither. The current model when compared to the census 2011 benchmark achieves an overall accuracy of 76%. The model is still being improved.

The **Statistics Netherlands** (CBS) are also experimenting with combining satellite data and their rich system of social statistics dataset to:

- Produce more detailed statistics by predicting the labels for areas with little or no statistical information due to low sample size or poor data quality.
- Produce more timely statistics by nowcasting the labels for periods with little or no statistical information yet.
- Improve small area estimation or time series models by learning features that can serve as auxiliary information.

They have started with Superview satellite images with a resolution of 0.5 m and publicly available square statistics on persons, households and dwellings per 500 m × 500 m square (25-ha grid) and 100 m × 100 m square (1-ha grid). Superview satellite images and square statistics use the same coordinate reference system. They have downloaded several satellite images and started cropping them into 25-ha squares and labeling each square with several statistics, such as the level of urbanicity. Refer to the images in Appendix 6.

To learn predictive features from the images, a convolutional neural network (CNN) will be trained. The exact architecture and choice of hyperparameters is yet to be determined. Special IT infrastructure might be needed to meet the demands for large storage capacity (images), computational power (CNN) and to include privacy-sensitive data such as geotagged household income.

The **Swiss Federal Statistical Office** (FSO) land use statistics are an invaluable tool for long-term spatial observation with an acquisition period that has been gradually reduced from 12 years (in 1979) to 6 years today. At present, internal resources are almost entirely allocated to visual interpretation, at the expense of other activities. In light of this, certain artificial intelligence algorithms could help though by facilitating land use and cover classification and by improving change detection. The objective of the pilot project “Arealstatistik Deep Learning” (ADELE) is to gain a better understanding of this field in preparation for the gradual development of a computer prototype. The results and knowledge obtained so far demonstrated the innovation potential for the FSO in using artificial intelligence to process images. The experience gained will be leveraged when much more complex neural networks are implemented and aerial images covering the whole of Switzerland are used.

In addition to using satellite data to produce new statistics, the pilot studies will further contribute to the advancement of ML solutions by:

- Providing concrete examples to apply the pipeline described above;
- Sharing ML solutions and best practices for others to use in the Modelling phase of the pipeline; and,
- Developing quality assurance processes to monitor the performance of the model when applied with little or no ground truth information, e.g. how to measure prediction accuracy and take this measurement into account when disseminating the estimates or indicators. This work is being conducted in collaboration with the work package on quality.

3.4 Quality

NSOs producing Official Statistics have a responsibility to their data users to inform them of data quality. In response many quality frameworks exist for traditional methods, however they do not cover some aspects of ML. The goal of this work package is to propose a quality framework for evaluating ML when used in standard statistical business processes, e.g. coding, and as the main statistical process, e.g. analysis of satellite data. The work package will also bridge the gap between it and existing traditional frameworks. This will allow NSOs to compare outputs from traditional and ML methods and to inform users of data quality when ML is used to produce outputs.

Following in the direction suggested by the BSTN last December and in response to the needs expressed at the planning sprint in May and confirmed at discussions on the pilot studies in September, the work package has first put the emphasis the accuracy dimension of the framework. Concepts in the proposed framework have been proposed and are being evaluated in the pilot studies being performed in work package 1. Members are realizing the quality challenges beyond demonstrating the value-added of ML, e.g. QA support in production. They are also identifying other dimensions where ML may add value, these will be included in the proposed framework and evaluated in the pilot studies.

3.6 Integration of Machine Learning

One of the recurring themes from the discussions at the sprints and monthly meetings is that integrating machine learning into official statistics requires more than simply building machine learning systems. In fact, a number of participants noted that they had already developed otherwise successful machine learning solutions, but had been unable to implement them into production processes because of a variety of organizational and structural impediments including uncertainty over who should be responsible for building, evaluating, and maintaining these highly interdisciplinary systems. Another comment observation is that in pursuing their pilot studies, team members are finding out similar or complementary developments within their own organisations. Some of their work has even generated internal collaboration. The goal of this work package, which hasn’t fully started yet, is to explore how different NSOs are organized to integrate ML in their production processes, and report on the different practices, sources of impediments and propose successful practices. It will be the main topic at a meeting prior to the November workshop.

4. Project Outputs

4.1 Sharing information

In addition to collaborating on several pilot studies, the team members have been conducting literature reviews and sharing numerous reference documents. Building on the papers provided by the BSTN, over 80 papers have been gathered and shared within the project so far.

As mentioned in section 2, all documents produced by the project (working documents, official documents, project updates) have been made available to a list of people who are either assisting team members in their work, involved in other international activities or are looking for an opportunity to get more involved.

Some team members have also presented their work within their organisations and to other organisations. The ML project was introduced at the ModernStats World Workshop 2019. Members from the Office for National Statistics presented the ML project and their work on edit and imputation to numerous employees of the Bureau of Labor Statistics (Washington), Destatis (Wiesbaden) and the International Monetary Fund. The presentations were very well received and led to other people requesting access to the work of the project through the followers and additional collaborators. The project manager and work package leaders presented the project to senior managers at Statistics Canada. They also met with the Chief Statistician of Canada to discuss the objectives of the project. Some of the work conducted on imputation will be presented to the Istat Advisory Committee and two members have been invited to present their work at the Sixth International Conference on Establishment Statistics (ICES VI).

4.2 Future plans

Soon after the September sprint, the project manager and work package leaders met in Ottawa from October 9 to 11. The purpose of the meeting was to take stock of the progress, set plans for future work and prepare for the November workshop.

Considerable progress has taken place since the project actually started in May. If the project were to end by the end of the year, this progress report could be enhanced and complemented by small reports or updates to the current working documents, i.e. mostly PowerPoint presentations. If the project were to be extended to a second year, the project manager and work package leaders agreed on the following project deliverables and milestones:

- A report on each pilot study. This report would describe the study, its value proposition to the organisation, its technical details, its results in demonstrating the value-added of ML, accompanying actions or challenges in advancing to use of ML in the organisation, and future work. The reports would be delivered by February 29 2020 by the team members conducting the studies. The reports would also be accompanied by documented ML scripts or applications for others to use. The team members would be supported by the pilot study theme members.

- A report on each pilot study theme. This report would summarize the value-added and best practices in using ML in the theme (coding, E&I and Imagery), advances in the implementation in organisations and recommend future work. These reports would be accompanied by selected reference documents among those gathered by the team. The reports would be delivered by April 30 2020 by pilot theme leaders with the support of team members conducting the studies and the work package leaders.
- A report on the integration of ML in the production process. This report would summarize the value-added of ML for NSOs, best practices in developing and maintaining them in production and… The report would be accompanied by a draft quality framework on key aspects in the use of ML and a report on how different NSOs are organized to integrate ML in their production processes, sources of impediments and successful practices. This report would be delivered by September 30 2020 by the work package leaders supported by pilot study leaders and the project manager.
- A workshop on machine learning that would provide an actual application for participants to get a hands-on learning experience, e.g. coding, and key lessons learned. This would be delivered by November 2020. This would be a team effort.
- A project wrap-up and recommendations on any next steps. This would be delivered by the project manager supported by work package leaders and the UNECE Secretariat.

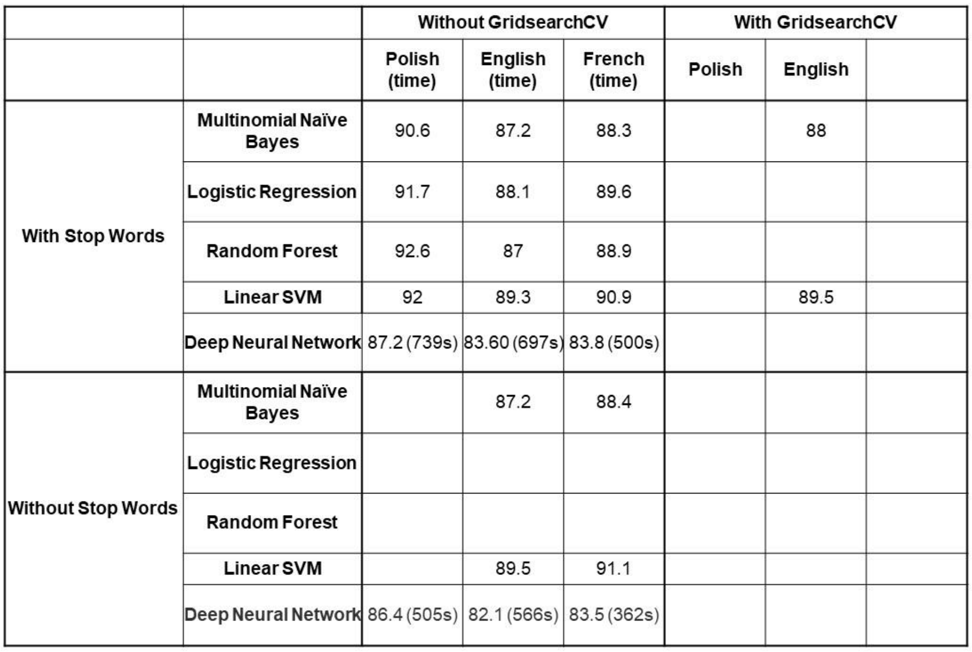
Appendix 1

List of international activities in which ML project participants are involved

- UN Global Working Group for Big Data - Satellite Imagery and Geo-spatial data
- ESSNet Big Data: from exploration to exploitation
- Copernicus – Europe’s eyes on Earth
- UN-GGIM Inter-Agency and Expert Group on the Sustainable Development Goal Indicators (IEAG-SDGS) Working Group on Geospatial
- CES In-depth review of satellite imagery / earth observation technology in official statistics
- HLG-MOS Specialized topic on Statistical Data Editing
- HLG-MOS project – Strategic Communication (Phase2)
- European Union’s Horizon 2020 - MAKSWELL - Work Package 3

Appendix 2[4]

Comparison of Accuracy Scores (%) across different languages and ML approaches (Poland, US-BLS, Canada and UNECE)

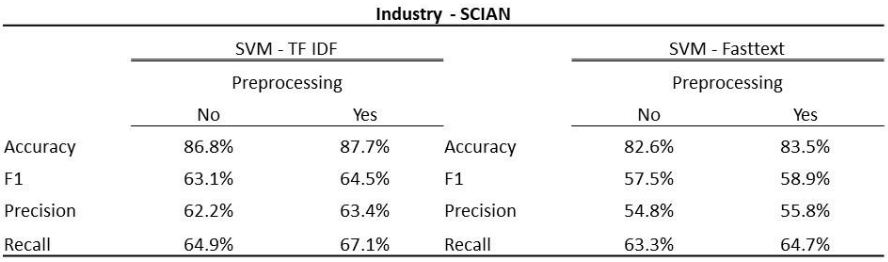


		Without GridsearchCV			With GridsearchCV		
		Polish (time)	English (time)	French (time)	Polish	English	
With Stop Words	Multinomial Naïve Bayes	90.6	87.2	88.3		88	
	Logistic Regression	91.7	88.1	89.6			
	Random Forest	92.6	87	88.9			
	Linear SVM	92	89.3	90.9		89.5	
	Deep Neural Network	87.2 (739s)	83.60 (697s)	83.8 (500s)			
Without Stop Words	Multinomial Naïve Bayes		87.2	88.4			
	Logistic Regression						
	Random Forest						
	Linear SVM		89.5	91.1			
	Deep Neural Network	86.4 (505s)	82.1 (566s)	83.5 (362s)			

[4] All the numbers provided in this document are not final. They are provided to provide some concrete indications of early results. Final, complete and accurate results will be provided in the technical reports on each pilot study.

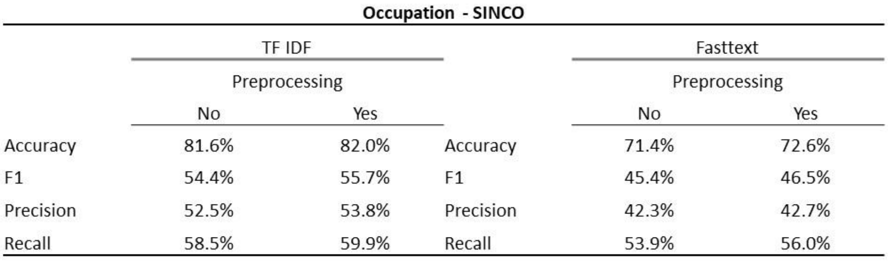
Appendix 3

Performance metrics on coding of industry and occupation (Mexico)



Industry - SCIAN

	SVM - TF IDF			SVM - Fasttext	
	Preprocessing			Preprocessing	
	No	Yes		No	Yes
Accuracy	86.8%	87.7%	Accuracy	82.6%	83.5%
F1	63.1%	64.5%	F1	57.5%	58.9%
Precision	62.2%	63.4%	Precision	54.8%	55.8%
Recall	64.9%	67.1%	Recall	63.3%	64.7%

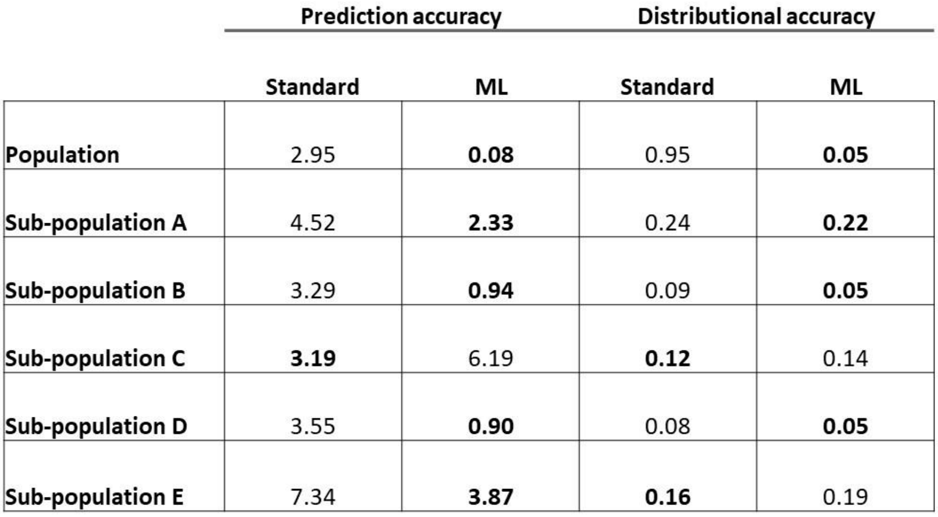


Occupation - SINCO

	TF IDF			Fasttext	
	Preprocessing			Preprocessing	
	No	Yes		No	Yes
Accuracy	81.6%	82.0%	Accuracy	71.4%	72.6%
F1	54.4%	55.7%	F1	45.4%	46.5%
Precision	52.5%	53.8%	Precision	42.3%	42.7%
Recall	58.5%	59.9%	Recall	53.9%	56.0%

Appendix 4

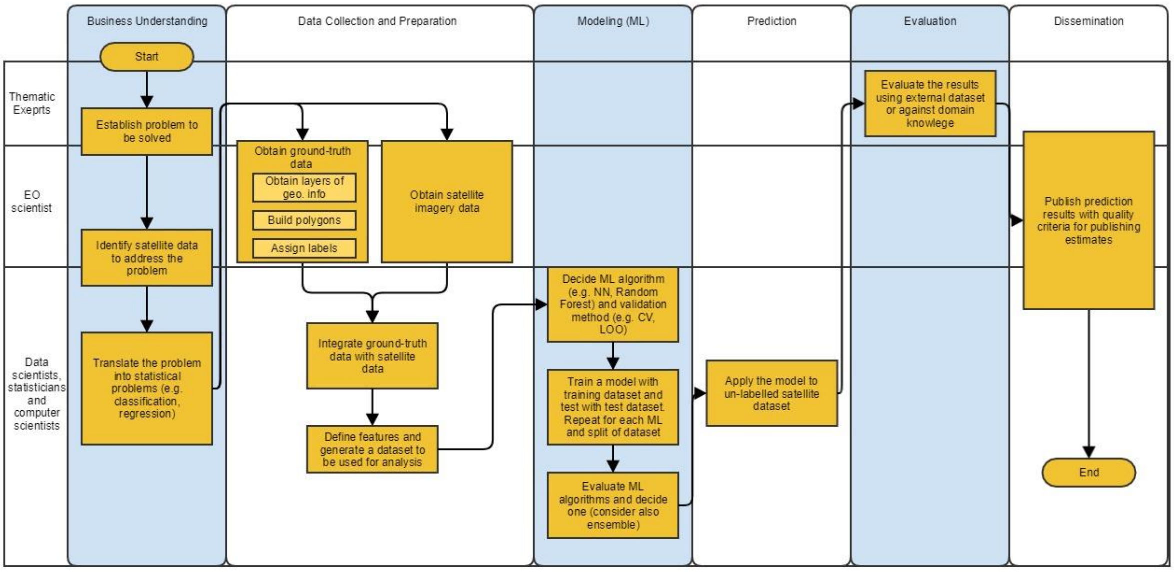
Accuracy of imputation of Attained Level of Education (Italy)



	Prediction accuracy		Distributional accuracy	
	Standard	ML	Standard	ML
Population	2.95	**0.08**	0.95	**0.05**
Sub-population A	4.52	**2.33**	0.24	**0.22**
Sub-population B	3.29	**0.94**	0.09	**0.05**
Sub-population C	**3.19**	6.19	**0.12**	0.14
Sub-population D	3.55	**0.90**	0.08	**0.05**
Sub-population E	7.34	**3.87**	**0.16**	0.19

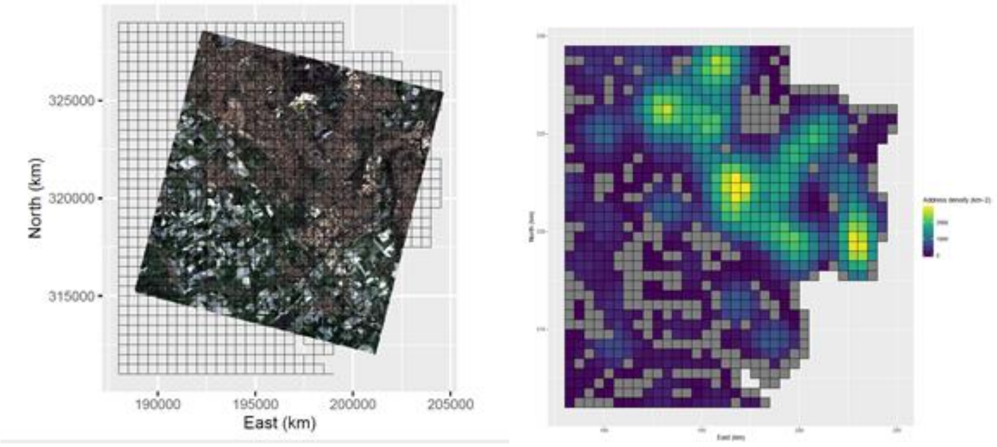
Appendix 5

Satellite Imagery Process Pipeline (UNECE and Mexico)


Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Evaluation
Dissemination
Thematic Exeprts
EO scientist
Data scientists, statisticians and computer scientists
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
End

Appendix 6

Example of satellite imagery and statistical information (Netherlands)



Relationship to be learned between satellite image (left) and statistical information (right). In this example, the image is taken by Superview on 20190401 of the city of Heerlen and surroundings and the statistical information is urbanicity per 500m × 500m square.