UNECE – HLG-MOS Machine Learning Project
Executive Summary – Work Package 1

Organisation: ONS (Office for National Statistics) - UK
Author(s): Claus Sthamer, Eric Deeben
Date: 08/12/2020
Version: 3.0

Introduction and background

This executive summary is intended for Heads of Statistical organisation, Chief statistician, Chief Methodologist, Chief Data Scientist, CTO/CIO and Senior leadership management.

The Machine Learning (ML) Project was proposed and approved during the HLG-MOS November 2018 workshop. Based on a paper by the Blue Skies Thinking Network (BTSN) the project aims to inform policymakers about the possibilities to use ML in the production of official statistics and to demystify official statisticians unfamiliar with it.

The interest in the use of ML for official statistics is rapidly growing. For the processing of some secondary data sources (incl. administrative sources, big data and Internet of Things) it seems essential that Statistical organisations consider opportunities offered by modern ML techniques. Although promising, so far there is only limited experience with concrete applications within the UNECE statistical community, and some issues have yet to be solved. The aim of this project is to therefore develop a proof of concept and to unearth any issues and challenges prior to full-scale development of any statistical outputs. The varied contexts of the NSOs are therefore hugely helpful to develop a full understanding of the challenges and opportunities inherent in the use of ML in official statistics.

The business proposal concludes with:
"ML is a key modern technology that the worldwide statistical community should consider and the methods, IT solutions and other related issues can be dealt with in a universal; manner. Since, at this moment in time, basically all NSOs are in the same pioneering phase this is an excellent opportunity for shared development and mutual collaboration. The ML proposal seamlessly fits the HLG-MOS mission, all four elements of its vision are covered, and all five HLG-MOS values are addressed."[1]

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.[2]

[1] https://statswiki.unece.org/download/attachments/261818141/BSTN_2019%20project_Machine%20learning.pdf?version=1&modificationDate=1581420211156&api=v2

[2] https://en.wikipedia.org/wiki/Machine_learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.[3]

Table of contents

1. National Statistical Organisation, their role and strengths

National Statistical Organisations (NSO) play a role of ever increasing importance to inform government policy makers of a true picture in their respective country. They produce trusted statistics based on the trust placed into them by the wider public, the respondence and the user of the statistical output. To stay relevant, NSOs have to adapt to and embrace new technologies and data sources to shorten the time between data collection and statistical output. The Covid-19 pandemic and NSOs response is testimony to this.

2. Project Objectives

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**. To achieve this objective the ML project will aim in year two, to:

- Report on the various Pilot Studies to demonstrate the value-added of ML.
- Identify and share best practices in the implementation of ML techniques.
- Share knowledge, tools and best practices on implementing the ML techniques, and how National Statistical Organisations (NSOs) are organized to move them quickly to the production processes.
- Propose a quality framework components for evaluating ML processes and statistics produced using them, as well as to bridge the gap between these components and those in existing frameworks.[4]

2.1 Work Packages and Themes

During its first project sprint in May 2019 at the ONS in Newport, UK, the group consulted the Generic Statistical Business Process Model (GSBPM). It was decided to choose 3 processes from the model's 'Process Phase' as the themes to conduct ML pilot studies to fulfil the objectives of this project:

- Classification and Coding (C&C)
- Edit and Imputation (E&I)
- Integrate Data: This GSBPM process will be represented in the project by pilot studies on Imagery and Twitter data. The latter will be attached to the C&C theme.

[3] https://www.sas.com/en_gb/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human%20intervention.

[4] https://statswiki.unece.org/display/hlgbas/Modernisation+Projects

The pilot study investigations form Work Package 1 (WP1). The other two WPs are:

- WP2 – Quality: ML evaluation framework, practices and techniques
- WP3 – Integration: Investigates issues to be considered for operationalising ML in the production pipeline of official statistics

3. Pilot studies and Theme Reports

The three themes agreed by the project: C&C, E&I and Imagery are all based on classification tasks with the exception of Imputation. For Editing it is a classification of records into two classes, the 'Change' class where the data are inconsistent, missing or suspicious looking and the 'No-Change' class, where the data do not need any further attention and are deemed to be correct or consistent. The sentiment analysis of web based data is included in the C&C theme as it is a classification task to classify the data into the chosen sentiment categories. Imagery, an example of big data and alternative data sources classifies satellite or aerial images or their components into classes like 'Urban' and 'Non-Urban'.

All pilot studies used supervised ML, this is where the algorithm 'learns' from training data that have been labelled, e.g. where the correct code has been assigned manually. This can be an occupation code assigned to a data record with an occupational text description, or the type of object visible on a satellite image. These labelled training data allow the algorithm during the learning phase to recognise rules or patterns in the data without having to explicitly formulating those rules. New data that have not been labelled can then be fed into the algorithm for it to categorise and recognise these data.

The participants of this project have submitted reports on their respective pilot studies. These were then summarised for each of the three themes into Theme Report. Given references to these reports are to the statswiki.unece.org web site. All reports will be accessible to the public. Further information can be obtained by contacting unece.org or the authors of the reports. This summary report uses information provided in the pilot study reports as well as the 3 theme reports.[5]

4. Classification and Coding

*"**Classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.*

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance."[6]

The given set of data referred to in the above quote is typically a text narrative provided by the respondent to describe, their occupation, the industrial activity of a company, injury information of an injured worker, product descriptions scraped from the internet or sentiment text obtained from Twitter.

There was only one pilot study with binary classification, the twitter sentiment classification into positive and negative sentiments. The target classes for this are ['Positive', 'Negative']

[5] https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies
[6] https://en.wikipedia.org/wiki/Statistical_classification

4.1 Summary of C&C pilot studies

Most of these pilot studies fall into the group of Multi-Class classification tasks. These aim to classify text descriptions into internationally accepted coding frames, e.g. SIC, SOC, NAICS, NOC or ECOICOP that

Table 1 – Coding Frames

Coding Frame	Description
SCIAN NAICS	Spanish Version: North American Industrial Classification (Sistema de Clasificación Industrial de América) English version: **N**orth **A**merican **I**ndustry **C**lassification
NOC	**N**ational **O**ccupational **C**lassification is Canada's national system for describing occupations
SINCO	National Classification System for Occupations (Sistema Nacional de Clasificación de Ocupaciones)
NACE	European Classification of Economic Activities (Nomenclature statistique des Activités économiques dans la Communauté Européenne)
SIC	**S**tandard **I**ndustrial **C**lassification – Established by the USA in 1937, replaced by NAICS in 1997
SOC	**S**tandard **O**ccupational **C**lassification
OIICS	Occupational Injury and Illness Classification System – Developed by the BLS
ECOICOP:	European Classification of Individual Consumption by Purpose
CTS:	**C**atalogue of **T**ime **S**eries by the IMF

offer many target classes and to assign an appropriate code from the coding frame. See Table 1 for a list of the coding frames used in the C&C pilot studies. Even though the aim of these studies appears to be mostly the same, their approach and software solutions used differ as their results do too. The sentiment analysis of twitter data is a binary classification task where each twitter message is classified as either Positive or Negative.

There were 9 pilot studies submitted, of these, 3 are in production with one of them used as a supporting tool for human coders to make a faster decision:

- Canada – StatsCanada[7]: Industry and Occupation Coding (NAICS & NOC), in production, 13.3% of cases are auto-coded, error rate < 5%
- USA – BLS[8]: Workplace Injury & Illness (SOC & OIICS), this study has shown that ML with Support Vector Machines (SVM) or Logistic Regression can outperform human coders. The use of a Neural Network has improved on this, it makes an estimated 39% fewer errors than the manual coding process. In production, predicted codes above a set threshold are auto coded, the remaining ones are manually assigned >85% of codes auto-coded
- Norway – Statistics Norway[9]: Standard Industrial Code (SIC), used ML as a supporting tool; up to 5 ML predicted SIC codes are presented to coders in decreasing order of their prediction score, in 22% of cases the highest prediction score is higher than 95%

The other 6 studies have advanced considerable during the 18 months since the project started.
Future plans for all of these include investigation into the use of other ML algorithms, increase in prediction accuracy or IT infrastructure before imbedding this technology into production pipelines.

The three operationalised projects have shown that building a "Golden Data Set" or ground truth is essential. For this, all labels are manually assigned and are deemed to be correct. To achieve this is very labour intensive and Subject Matter Experts (SME) have to check and re-check the assigned codes. This serves two purposes:

1. The ground truth is used to train the ML algorithm. The better this data set is, the better the ML algorithm can establish rules and find patterns in the data.
2. As importantly, this allows to compare ML prediction results with traditional work by establishing how accurate the traditional manual coding process is.

To commit this resource is a challenge for most NSOs and can be a serious blocker in the development of ML solutions. However, such resources should already be committed to assess the current processes.

[7] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_Canada.pdf?version=1&modificationDate=1605171571083&api=v2
[8] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_USA-BLS.pdf?version=2&modificationDate=1605171512748&api=v2
[9] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_Norway.pdf?version=1&modificationDate=1605171509316&api=v2

Even though the success criteria for most of these studies was that ML has to perform at least as good as human coders, their motivation in conducting these was mainly gains in efficiency and timeliness. Accuracy as the objective was only stated in two studies. To achieve this, ML is used either to only auto code predictions made above a set threshold and/or predictions for coding classes that have a high prediction threshold. Predictions for minority classes, the ones that are rarely seen in the data, are then excluded from the ML solution and are still made by human coders. The BLS and Stats Canada solutions use a mix of ML auto coding and human coding to ensure that the overall accuracy is higher than human or machine coding alone. Instead of just showing the overall performance, this approach assessed and demonstrates if, how and where ML was at least as good or even better than human coders. Setting a prediction threshold for the groups of records where ML works better, where it works good enough to assist humans and where it does not perform good enough (e.g. minority classes) are important areas to analyse. The ultimate objective is that NSOs end up with a better overall process. Please see [10] for individual pilot study results.

The BLS solution, as the highest advanced in this group, has the highest proportion of auto coding, > 85%. This has been made possible by using Neural Networks for the ML algorithm that run on 4 Graphical Processing Units (GPUs) with 3584 cores.

The Norwegian solution was the only one using Cloud Computing, but it is anticipated that this technology will be utilised more often as it offers the provision of very powerful IT solutions without capital expenditure, but national privacy and data protection laws have to be considered first. All other studies relied on standard desktop or laptop hardware.

5. Edit and Imputation

The E&I theme report gives this definition for the E&I task:

"To make clear what the two *parts (editing on the one hand side, imputation on the other) are of, it is necessary to introduce the following differentiation: for the machine learning project, we treated*

- *editing as the task to identify missing and problematic data (i.e. implausible values, contradictions in records, …) in data sets and*
- *imputation as altering incorrect values and inserting missing values.*

Note that other definitions (which are not used here) treat process of altering incorrect values as part of the editing."[11]

There were 7 pilot studies in the E&I theme:

- Imputation:
    1. Italy: Imputation of Attained level of Education in base Register of individuals
    2. Poland: Imputation in the sample survey on participation of Polish residents in trips
    3. Germany: Machine Learning methods for Imputation
    4. Belgium - VITO: Early estimates of energy balance statistics using Machine Learning
- Editing
    1. Italy: Machine learning for Data Editing Cleaning in NSI, Some ideas and hints
    2. Italy: machine learning tool for editing in the Italian Register of the Public Administration
    3. UK: Editing of social survey data with ML

[10] https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies
[11] https://statswiki.unece.org/display/ML/WP1+-+Theme+2+Edit+and+Imputation+Report

Editing

Possible value added aspects of ML for Editing as discussed by the group:

- Discovery of hidden rules in the data only known by intuition to conserve this knowledge in the team
- To allow the human editors to concentrate on validation of the important records
- Classify records or data cells as "plausible" or "not plausible"
- ML may offer an efficient mechanism for a not rule based perspective on editing

Imputation

All for pilot studies for Imputation have shown that ML can add value to the data processing pipeline:

- Better and faster Imputation
- Building the Imputation model might be time consuming
- More timely statistics with less human intervention

5.1 Summary of E&I pilot studies

The E&I Theme report concludes on Editing:
"According to the study so far, with machine learning the editing process can be completed much faster, more consistent, possibly even lead to higher quality data and allow for much sooner publication, but the effort required to maintain training data, the machine learning model and the analysis of the results in a short term might not prove to be a cost saver. I. e., the gain until now seems to be not so much in efficiency of the results but in the efficiency of the statistical process: machine learning methods allow using huge amount of data with much less a priori knowledge, hypotheses and data preparation (general underlying structure of the data, stratification, etc.)."[12]

The E&I Theme Report points out: *"Note that parametric models are always the best, from every point of view, if the hypothesis is good! Unfortunately, we often make mistakes in specifying the underlying hypothesis, i.e. in modelling the phenomena; hence the parametric model is not able to provide good predictions. Non-parametric models run less risk from this point of view but fit (in the finite data situation) less well than the "true" parametric model."*[13]

All the pilot studies in the E&I theme used non-parametric models. The ML algorithms were left to find the structure and relationships in the data to build their own rules and parameters to predict values for imputation and to predict if a record needs human attention. The paper from Italy on Editing is not a pilot study but it is a framework on which to develop ML for Editing. This investigation together with the other pilot studies have shown that a great deal of progress has been made by the E&I group, things look generally promising, but there is still a way to go.

6. Imagery

Satellite or aerial images are becoming more and more available from a host of providers with increasing image resolution and frequency of updates. This does not just open up opportunities to estimate the fulfilment of the sustainable development goals as expressed in the United Nations document: "Transforming our World: the 2030 Agenda for Sustainable Development", but it also allows for many other applications. The imagery theme falls into the alternative and big data source category and explored how ML can be used to extract information from satellite and aerial images for statistical purposes. Pilot studies in the imagery theme show the application of classification algorithms that seek to relate patterns found in labelled satellite images to unlabelled images.
There were 5 pilot studies in the imagery theme, covering tasks like the growth of urban space in Mexico, land usage in Switzerland, Poverty detection in the Netherlands, Address Register maintenance in Australia and a Generic Pipeline for Production of Official Statistics Using Satellite Data. And Machine Learning by the UNECE.

[12] https://statswiki.unece.org/display/ML/WP1+-+Theme+2+Edit+and+Imputation+Report
[13] https://statswiki.unece.org/display/ML/WP1+-+Theme+2+Edit+and+Imputation+Report

The Imagery theme report states: *"The expectations of the participants involve the need to create a new process that complements the activities of the NSOs or simply to improve existing processes. Either way, progress will be based on the application of Machine Learning techniques to satellite images."*[14]

Pixel resolution range from ~23cm for aerial images to 30m for Landsat 5,7 and 8 images. In addition, complementary information such as ESRI shape files or the open GeoPackage format which contain statistical or geographic information can aide the imagery ML processes.

6.1 Summary of Imagery pilot studies

The main motivation for the participating pilot studies in this group was to reduce cost and time required by either manual inspection of all images or the respective existing methods of collecting the data. Just like the pilot studies from the other themes, Imagery needs labelled data for the algorithms to learn and establish rules to recognise the image features as required. This can be down to pixel level or areas of an image. Even though this labelling process is very time consuming for Imagery, the long term benefit can be significant.

All 4 pilot studies in this group used Convolutional Neural Networks (CNN) as the ML algorithm. And as the pilot study from Switzerland on land use and land coverage has shown, this can be augmented with other algorithms and data sources, e.g. Random Forest to increase the prediction capability of the pipeline. The skill level needed for imagery ML solutions is therefore higher compared to other ML classification solutions.

7. What is needed to do ML?

Even though specialist knowledge of ML is needed to get the best possible results, a Proof of Concept (PoC) can be fairly quickly built if suitable training data that have already been labelled are available.

Collaboration within this UNECE ML project has shown that knowledge sharing and mentoring can set a NSO very quickly, in a matter of hours, onto its path of building a first ML PoC. Further steps utilising more complex ML algorithms, data pre-processing and evaluation of results are then much easier to achieve. With expanding ML skills and experience results can certainly be improved, as can be by applying other ML algorithms or even Neural Networks running in a cloud solution or on a complex and expensive IT infrastructure.

On-line tutorials, literature with example source code and data sets are also readily available. To translate this onto a PoC with survey data is another step. Data pre-processing to take care of the many data issues inherent with collected data becomes important and can be a major part of the project.
Most pilot studies contributions of this UNECE ML project were carried out on standard office laptops or desktop PCs, to stay within reasonable ML algorithm training times, only the more complex applications of Neural Networks demand much higher processing power.

However, to advance the initial idea via a promising looking PoC towards implementation poses a lot more challenges. These can range from ethical, legal and technical questions to overcoming process hurdles such like:

- How to monitor the ML model performance?
- How to keep the ML model relevant?
- How to create new labelled training data for the ML algorithm.
- How to integrate the ML process into the existing data production pipeline?

But the most challenging aspect has shown to be getting the acceptance of all stake holders for this change. ML seems to promise a much faster data production pipeline at reduced costs. This coupled with the expectation that data consistence and error rates are lower compared to the traditional manual approach, makes ML a good proposition to be pursued.

[14] https://statswiki.unece.org/display/ML/WP1+-+Theme+3+Imagery+Analysis+Report

8. What worked well and what worked even better

This UNECE HLG-MOS ML project and the contributed pilot study reports that have been operationalised have clearly demonstrated the added value of ML in the production of official statistics. The other reports are about ML investigations still in the proof of concept stage. They have shown a varying degree of added value ranging from “some potential” to “significant” in all three themes, but specially for C&C and Imagery.

Good results can be achieved quickly, but better results need a good “Ground Truth” data set, advanced ML skills and possibly very powerful IT hardware. Once the ML model has been trained with labelled training data, ML can classify data much faster than humans can. A direct financial benefit has not been shown by any of the pilot studies as it would be difficult to ascertain this. But participants with operationalised ML reported that a possible financial gain was not the primary driver. Being able to produce statistical output faster with at least the same or even higher level of accuracy has been mainly the motivation. For imagery pilot studies, ML offers a solution that would have otherwise not been possible. Some pilot studies have shown that ML used as a verification tool can be of great benefit in supporting and accelerating manual classification tasks.

ML works well when it sits alongside with humans and is not perceived to replace them.

ML can be deployed to help NSOs to stay relevant by delivering accurate, consistent and economically viable official statistics faster.

9. Organisational and Skill requirements

ML projects have a lot in common with traditional software development projects, but these two are also different in many ways. Were the traditional project can follow a straight line, mapped out during the project planning phase, a ML project will most likely have to go through many iterations of experiments with various ML algorithms their tuning as well as a discovery phase on the data. This might not just highlight blockers or new opportunities, but also possibly a different project, a solution to a different problem.

As all software solutions, ML solutions will certainly also need support and bug fixing. But operationalised ML solutions need more than that. They need monitoring for model drift, that is when the trained ML model, used to predict the classes new data belong to, does not match the data anymore. This happens over time when the relationship between target variables and the independent variables change. An example of this is when the socio-economic behaviour or situation of the survey respondents change.

A new model has to be trained. This requires new labelled training data to form a new Gold Standard data set.

ML skills need to be available for monitoring and retraining of a model. This requires that ML skills are available throughout the lifecycle of the ML pipeline. And for this to happen, the traditional departmental and skill silos have to be replaced with cross departmental teams to provide and maintain the right skill levels when they are required.

10. Conclusion and what is next

Digitisation has driven the growth of big data, and machine learning makes big data more usable. It's easy to see why NSOs are exploring different ways of augmenting their existing decision-making and business operations with alternative data sources and machine learning. For example, the Office for National Statistics (ONS) has invested in infrastructure capable of supporting frictionless access to and harnessing of alternative data sources (specifically, administrative data and satellite data). It has also sought to grow its in-house data science capacity through the establishment of the Data Science Campus.

The value of traditional approaches to data gathering (such as surveys) and analysis, however, is still high, especially in circumstances where big data isn't (yet) readily available. Applied machine learning can add value to these traditional approaches by making them more operationally efficient.

The UNECE ML project has been successful by demonstrating that participating NSOs have made advances in the use of ML. It has shown the added value ML can bring to the production of accurate, timely and cost-efficient official statistics.

Summary of outputs

- 21 reports, mostly demonstrations of added value of ML (pilot studies)
- 3 summary analysis reports on the use of ML in Classification & Coding, Edit & Imputation and use of Imagery
- An initial quality framework for statistical algorithms (WP2)
- Integration challenges and practices (WP3)
- Shared code from numerous studies and shared product description data to practice some ML
- Links to learning and training material; references

In summary, our findings on the use of ML, that ML should be used for Coding & Classification, we have numerous positive demonstrations on a variety of data sources and contexts. Currently we have a few applications in production or near production.

ML shows great promise for Edit & Imputation and Demonstrations with a varying degree of positivity, pilots are in advanced stages and plans are in place to move these to production.

ML is essential in the use of imagery, especially in the context of increasing access to large amounts of imagery data. There are more advanced developments in this theme group and the Generic Pipeline for Production of Official Statistics Using Satellite Data and Machine Learning[15] shows the business processes required to develop ML solutions for Imagery.

This project concludes in November 2020, but it will carry on as ML 2021 under the guidance of the ONS' data Science Campus to continue with the aim to develop and integrate ML solutions in statistical processes at NSOs. It will also follow and support the development of the ML applications not yet operationalised.

[15] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_Imagery_UNECE.pdf?version=1&modificationDate=1605171593842&api=v2