Data Science Campus

UNECE HLG-MOS ML Project Work Package 1 Summary Report

Eric Deeben

Eric.Deeben@ONS.gov.uk

Work Package 1 (WP1)

- WP1 consists of 3 themes taken from the Generic Statistical Business Model (GSBPM):
  - Classification & Coding (C&C)
  - Editing and & Imputation (E&I)
  - Integrate Data → Imagery

- WP1 objective:
  - Conduct Pilot Studies to demonstrate the value-added of Machine Learning (ML)
  - Have the Pilot Studies advanced NSOs ML capabilities?

- WP2 & 3 will be covered by other presentations

Classification and Coding

1.	**BLS – USA**	Survey of Occupational Injuries and Illnesses	Workplace Injury – SOC, OIICS, 6 codes
2.	**Stats Canada**	Canadian Community Household Survey	Occupation & Industry – NAICS, NOC
3.	**Statistics Norway**	New Companies for the Central Coordination Register	Standard Industrial Code – SIC
4.	INEGI – Mexico	Household Income and Expenditure	Occupation & Economic activity - SCIAN, SINCO
5.	Statistiek Vlaaderen – Belgium	Sentiment of Twitter Data	Positive/Negative
6.	SORS – Serbia	Labour Force Survey	Economic Activity – NACE
7.	Statistics Poland	Web scraped food products	Food description - ECOICOP
8.	IMF	Catalogue of Time Series - CTS	

Value added by ML for C&C

- 3 pilot studies are in operation
- The others have advanced considerable
- All of them plan to experiment/research other ML algorithms
- Good results can be achieved with little IT resources
- ML can make official statistics:
  - more consistent
  - Accurate
  - Faster
- Combining human and ML coding gives best results

Editing and Imputation

Imputation:

1. Istat Italy: Imputation of Attained level of Education in base Register of individuals
2. Statistics Poland: Imputation in the sample survey on participation of Polish residents in trips
3. DESTATIS Germany: Machine Learning methods for Imputation
4. Belgium - VITO: Early estimates of energy balance statistics using Machine Learning

Editing:

1. Istat Italy: Machine learning for Data Editing Cleaning in NSI, Some ideas and hints
2. Istat Italy: machine learning tool for editing in the Italian Register of the Public Administration
3. ONS UK: Editing of social survey data with ML

Value added by ML for E&I

Pilot study results for Editing suggest:

- Much faster, more consistent, higher quality
- ML builds the rules for Editing, human expertise can be utilised to build training data
- Might not be a cost saver

Imputation:

- ML delivers comparable to traditional methods results in a more automated way
- Often plausible predictions, but in some cases implausible
- Much faster – some data pre-treatment can be skipped

Imagery

1.	**ABS Australia**	Reducing manual intervention for Address Register maintenance
2.	CBS Netherlands	ML for detecting poverty and population distribution from aerial/satellite imagery
3.	FSO Switzerland	ML for classification of land use - Arealstatistik Deep Learning (ADELE)
4.	INEGI Mexico	Use of Landsat satellite data for the mapping of urban areas in non-census years
5.	UNECE	Generic Pipeline for Production of Official Statistics Using Satellite data and Machine Learning

Value added by ML for Imagery

- Satellite/aerial images becoming more available – sometimes free
- Resolution is increasing as is frequency of updates
- Labelling of images as training data is very time consuming
- Convolutional Neural Networks performed best for 3 Pilot Studies
- Strengthened collaboration between methodologists and data scientists

WP1 - Lessons Learned

- Solid business case for the ML project – What is good enough?
- Golden Data Set or Ground Truth
- Some ML applications require powerful specialised IT hardware
- But good results can be achieved with less resource hungry algorithms

Conclusions

- ML can find rules/relationships between data features
- ML can add value to the production of official statistics

  Speed, Accuracy, Consistency
- This has been shown for C&C and Imagery
- The potential for E&I is high, but more work is needed here
- A financial gain is possible but difficult to determine
- Good training data – Golden Data set is needed
- ML works best alongside human coders

Data Science Campus

Machine Learning Project 2021 – Future Directions

“Discussion today 16:45 CET”

Data Science Campus

Thank you