## Data Science in Official Statistics (Introductory Level)

Kelly McConville and Wesley Yung International Conference on Establishment Statistics

June 17, 2024



Delivering insight through data for a better Canada



# Plan for the afternoon

- 1:30 1:45pm: Introductions
- 1:45 3:00pm: Data science in official statistics: Opportunities and challenges (Wesley)
- 3:00 3:30pm: Break
- 3:30 4:15pm: Model-assisted estimation via machine learning models (Kelly)
- 4:15 4:45pm: Survey estimation in R (Kelly)
- 4:45 5:00pm: Recap (Both)



# Data science in official statistics: Opportunities and challenges



# Outline

- Context of Official Statistics
- Production of Official Statistics
- Data Science
- Data Science use cases at NSOs





Many National Statistics Organizations (NSOs) follow the UN's Fundamental Principles of Official Statistics (FPOS)

- 1. Relevance, Impartiality, and Equal Access
  - Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

#### 2. Professional Standards, Scientific Principles, and Professional Ethics

 To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.



- 3. Accountability and Transparency
  - To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics. Principle 4: Prevention of Misuse The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.
- 4. Prevention of Misuse
  - The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.



- 5. Sources of Official Statistics
  - Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.
- 6. Confidentiality
  - Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

### 7. Legislation

• The laws, regulations and measures under which the statistical systems operate are to be made public.

Canadä



- 8. National Coordination
  - Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.
- 9. Use of International Standards
  - The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

### **10.International Cooperation**

• Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.





- Principles which are pertinent to this talk include

- 1. Relevance, impartiality and equal access (Serving the **government**, **the economy and the public** with data about the **economic**, **demographic**, **social and environmental situation**)
- 2. Professional standards, scientific principles and ethics (Scientific principles and professional ethics, on the methods and procedures for the **collection**, **processing**, storage and presentation of statistical data)
- 9. Use of international standards (Use by statistical agencies in each country of international concepts, classifications and **methods**)



- Statistics Canada (and other NSOs) has two main roles:
  - To provide statistical information and analysis about Canada's economic and social structure to develop and evaluate public policies and programs
  - Improve public and private decision making for the benefit of all Canadians



#### -ONS Purpose

- 'The Office for National Statistics (ONS) collects, analyses and disseminates statistics about the UK's economy, society and population.'
- We are independent of ministers and instead report through the UK Statistics Authority to Parliament and the devolved administrations of Scotland, Wales and Northern Ireland. We are the executive office of the UK Statistics Authority, who are responsible for promoting and safeguarding the production and publication of official statistics which serve the public good.
- Key Point: Information about the population for the public good





- Many NSOs have frameworks which help them meet these goals
  - Examples include quality guidelines, data ethics and sampling theory to allow **inference to the population**
- At Statistics Canada, our quality guidelines define six elements of quality
  - Relevance, Timeliness and punctuality, Accuracy and Reliability, Accessibility and clarity, Interpretability and Coherence and Comparability
- European Statistics Code of Practice
  - Statistical Output: Relevance, Accuracy and Reliability, Timeliness and Punctuality, Coherence and Comparability, and Accessibility and Clarity

Statistics Statistique Canada Canada Canadä

12

- Inference to the population
  - Traditional survey sampling relies on well-developed sampling theory
  - Design based sampling theory
    - Accounts for samples not selected
    - Provides framework for unbiased point and variance estimation
  - Censuses and registers
    - Significant amount of time and resources spent on ensuring coverage and concepts
- For Official Statistics to be for the public good, they must allow inference about the population



- Many NSOs are now modernizing
  - Response rates are declining
  - Alternative data sources are now available
  - Competition in the production of statistics/information space
    - Mis-information
    - Dis-information
- Leverage data science methods to remain relevant/timely
- Where can these methods be used in the production of official statistics?



- Research question/data need is known
  - A policy need
  - Data required for evaluation of a policy
  - Description of the situation
    - Eg. GDP, Retail trade statistics, Census of Population, etc.
- Design stage
  - Questions/content, planning, etc.
    - What information is needed?
    - How will it be collected?
    - What level of quality is needed?



- Data collection/gathering
  - Sample survey
    - Probability sample leads to sampling weights which allows inference based on sampling theory
    - Non-probability sample
      - Hard to make inference to the population
      - Some techniques exist to estimate weights which allows for inference to population
  - Census
    - Weights equal to 1
    - Other challenges such as coverage, concepts and timeliness



- Administrative data are a census which means weights of 1 and inference to the population
  - Challenges similar to those of a census
- Alternative data (sensor, images, social media, Google trends, etc.)
  - Challenges
    - Is it a census (coverage)?
    - Do we know the probabilities of selection? Can we assign weights?
  - Can we do inference?



- Coding of information to classifications
  - Industry, product, occupation, etc.
- Edit and imputation
  - Editing outliers, validity checks, checks for coherence, etc.
  - Imputation replace missing values or improbable observations
    - Framework exists to estimate uncertainty from imputation



- Estimation inference
  - Use weights to produce estimates
  - Variance related to sources of error or random processes
  - If a sample, then there is a sampling variance
    - Framework takes into account one sample out of all possible samples
  - If a census, no sampling variance but variance due to imputation
  - Total survey error concept
- Dissemination
  - Visualization, tables, analysis, confidentiality, etc.



# **Data Science**

- Data science has many definitions but no one definitive definition
- To keep things simple, for this presentation, data science = machine learning
- Common dichotomy supervised vs unsupervised
- For this presentation, we will consider algorithmic vs inferential data science (Estep, 2023)



# **Data Science**

- Algorithmic data science
  - Produces good predictions or finds underlying patterns in the data
  - Not necessarily concerned about inferring characteristics to a population
- Inferential data science
  - Extract knowledge and drawing conclusions to the population of interest
  - Quantification of uncertainties
- Both have their places in Official Statistics



# Data Science Use Cases in NSOs



# Financial information from balance sheet and income statements



- Information Extraction from Cash Flows Statement
  - Extract cashflow table from Financial Statements
- Traditionally, manual review was timely and could cover only a portion of documents
- Objective
  - Use machine learning to correctly **identify** correct Cash Flows Statement page and **extract** key variables from that page





#### - Process

- Random Forest Classifier to find Balance Sheet, Income Statement or Cash Flows statement
- Identify rows and columns in IS, BS or CF page
- Extract financial variable in digital format using Spatial Layout based Information and Content Extraction (SLICE)
- In production since Oct 2022
- Automated financial variable extraction for close to 70,000
  PDFs per year
- Led to improved quality for a larger number of enterprises



# Health and safety information



- Example from US Bureau of Labour Statistics (borrowed with permission)
- Produce information on Workplace Health and Safety



## **Survey of Occupational Injuries and Illnesses**

#### **Example Narrative**

Job title: sanitation worker

What was the employee doing just before the incident? mopping floor in gym

What happened? slipped on water on floor and fell

What part of the body was affected? fractured right arm

What object directly harmed the employee? wet floor

29 — U.S. BUREAU OF LABOR STATISTICS • bis.gov

## Occupation

Occupation: 37-2011 (Janitor) Nature: 111 (Fracture) Part: 420 (Arm) Event: 422 (Fall, slipping)

**Codes Assigned** 

Source: 6620 (Floor) Secondary: 9521(Water)



## **Survey of Occupational Injuries and Illnesses**

- 300,000 written descriptions each year
- Information on occupation, activity, what happened, etc. need to be coded
- Historically done by hand
  - Very time consuming 10 FT employees
  - More importantly very difficult and subjective
- Automatically code using machine learning



30 — U.S. BUREAU OF LABOR STATISTICS • bis.gov

# **Survey of Occupational Injuries and Illnesses**

- Initial foray into machine learning, so logistic regression was used
  - Produced probabilities for different codes select the most probable
  - Accept codes with probability above threshold and human code those below
  - Balance between quality and efficiency
- Implemented in phases starting in 2014
- Logistic model replaced by neural network in 2018



31 — U.S. BUREAU OF LABOR STATISTICS • bis.gov

## Percentage of codes automatically assigned



# Capture of receipt information



# **Collection – Image Processing**

- Indigenous Communities Food receipts Optical Character Recognition (OCR)
- Fill data gaps in food prices in northern and isolated communities
  - Information on food diversity, availability and subsidy benefits (Nutrition North Canada Subsidy Program)
  - Insight into the cost of living in northern and isolated communities
  - Insight into purchasing patterns and preferences, which could inform the subsidy list and levels
- Web app to allow users to upload receipts and OCR will parse food prices and descriptions



## **Process flow**



Canadä

# **Collection – Image Processing**

- Text on receipts detected via CRAFT (Character Region Awareness for Text Detection)
  - CRAFT outperformed other models
    - DBNet, FCENet, Tesseract, TextSnake
- Text recognition
  - Tesseract-OCR outperformed other models
- Unfortunately, funding for the project was not continued but StatCan is looking at using the technology for household spending survey



# **Environmental statistics**



# **Collection – Image Processing**

- Machine learning and satellite imagery to detect change in woody biomass
- Objective was to identify and analyse woody vegetation using automated remote sensing for woody biomass mapping and change tracking
- Data sources include LIDAR and high-resolution satellite images
- Training data is needed
  - Ground truthing expensive, inaccurate and time consuming
- Evaluate use of LIDAR and satellite optical images for ground truth

38

Canadä





Canadä

# **Collection – Image Processing**

- Combination of LIDAR and satellite images use to create 'ground truth'
- Convolutional Neural Net used for detection model
- Different land use types (cropland, forest, settlement, other) seen during both training and testing
- Good to very good model quality (based on F1-score) ranging from 80 % to 97 % in most areas depending mostly on the complexity of the area
- Model applied to estimate change in Woody Biomass



#### Example 1: New woody biomass growth 2013 and 2020



2013 image 2020 image

Difference between the model predictions on the 2020 image compared to the 2013 image



#### Example 2: Woody biomass that has disappeared 2013 and 2020



2013 image



2020 image



Difference between the model predictions on the 2013 image compared to the 2020 image



#### Example 3: Woody biomass that has disappeared 2013 and 2020



2013 image



2020 image



Difference between the model predictions on the 2013 image compared to the 2020 image 43



Statistics Statistique Canada Canada

# **Classifying comments**



# **Coding - Census Comments Classification**

- Respondents have the opportunity to provide comments during Census of Population
- Comments provide information on Canadians' perceptions of the census and possible burden and quality issues; they also inform content decisions for the next census
- Millions of comments received at each Census
- -Objective
  - Classify Census respondents' comments into classes (subject matter area, general comments, burden, experience with EQ, etc.)



## **Coding - Census Comments Classification**

#### **Class Distribution of labels** Comments in each category 40 38.73 35 14.05 11.72 10.57 8.88 6.54 6.27 5.24 4.94 4.36 5 2.04 2.57 2.52 0.4 13 0 Old questions No longer relevant Not included in study General – unrelated to the census Payments and housing Labour and Journey to work Rights-holders and Education Veterans General – negative census experience Burden Geo Demo Coverage ğ Language ADL Indigenous Sex and Gender Ethnocultural General – positive census experience Class

Cod e	General Categories				
0	Old questions/ no longer relevant/ not include in study				
1	General comments - unrelated to the census				
2	General comments - positive census experience				
3	General comments - negative census experience				
4	Burden of response				
5	Experience with the electronic questionnaire				
6	Payment and Housing (SMAs)				
7	Labour and Journey to work (SMAs)				
8	Geography, Demography and Coverage (SMAs)				
9	Ethnocultural (SMAs)				
10	Language (SMA)				
11	Rights-holders and Education (SMAs)				
12	Activities of daily living (SMA)				
13	Indigenous (SMA)				
14	Sex and gender (SMA)				
15	Veterans (SMA)				

#### Canadä

46

# **Coding - Census Comments Classification**

- Single comment can address multiple subjects
  - Multi-label, multi-class classifier needed
- ML models investigated
  - SVM, Semi-supervised Bi-LSTM (bi-directional Long Short-Term Memory – a recurrent neural network), CNN and transformers
- The transformer model was put into production for Census 2021
  - Classified 1.8 M comments
- Use of the model provided higher quality classification in a timelier manner



# Validation of information (editing)



## Aerial Image Recognition (AIR)



- ABS is using aerial images to update their Address Register
- Aerial Image Recognition is combined with Admin to determine address use.
- Main benefits of AIR
  - 12% of addresses which have images but no admin data
  - Admin data knows a residential building is planned at an address, but it isn't finished yet.
- Significant reduction in cost and time over manual

17/06/2024

## Aerial Image Recognition (AIR)

- CNN aerial image classifier
- Residential / Under Construction / Vacant High Density / Commercial / Roads
- Addresses associated with lat/long point and property boundary polygon where available
- 150x150 pixels, covering ~30x30m
- Production Deployment August 2019, updated 2023
- Canvasses ~800,000 images per quarter



# Outlier detection for International Trade data



# **Editing – Error Detection and Correction**

- Unit Value (UV) Error Detection and Correction
- Merchandise import data
  - Administrative data consisting of over 12 M records with 40 variables and over 12K Harmonized System (HS) codes
  - Value and quantity are requested
  - Unit Value is Value/quantity
- -Quantity not well reported (low value to customs)
- Existing treatment of UV not sophisticated enough
- Significant resources spent on low value manual validation due to large volume





# **Editing – Error Detection and Correction**

- Objective was to detect and correct errors in quantity variable
- Use XGBoost regression to determine if quantity should be changed and to what
- Defined target variable to be log of reported quantity and true quantity
- For training data, subject matter knowledge was applied, when necessary, to define the 'true' value for records defined as incorrect



# **Editing – Error Detection and Correction**

- -XGBoost regression was used using 21 features from the following general classes
  - Date, Unit of measure, Quantity and value, Weight, HS10 code, Country and Business
- Predicted quantity computed from predicted Y (reported over predicted) and is used if 'far' from the original quantity
- Monthly QA Dashboard which allows analysts to target their manual verification using filtering on relevant information
- Outcome was higher quality data by manually verifying records with larger impact



# **Estimation**



# **Estimation**

- After the break, Kelly will cover ML methods in a model assisted framework
- As of now, few NSOs have produced estimates directly from ML models
  - For example, crops from satellite images
- Policy of informing users of data quality
  - Measures of uncertainty are required
  - Kelly's model assisted framework provides these
- Research is ongoing but more is needed





# Dissemination



## **Dissemination – Large Language Models (LLM)**

Statistics

Canada

### StatCanChat

- LLM laid over StatCan website
- Enhance access to information on website in a conversational manner
- 20,000+ documents indexed for retrieval through LLM interface
- Uses Retrieval Augmented Generation (RAG)
- Includes citations for direct access to the article the text is based on

can you provide birth rate information, or fertility information after 2010 The total fertility rate in Canada is a common indicator used for analyzing fertility trends, and for the year 2011, 3.6% of all births were to mothers under age 20<sup>1</sup>. The estimated completed fertility rate for women born in 1970, who were aged 41 in 2011, was calculated based on their fertility rates having already peaked <sup>1</sup>. The fertility rate by age for selected cohorts shows that at age 40, the fertility level of the cohort born in 1970 was 20.9 births per 1,000 women <sup>1</sup>. From 2012 to 2016, there has been an average of 997 births per day in Canada, with the highest daily average recorded in September and the lowest in December and January<sup>2</sup>. The total fertility rate by birth order in Canada from 2001 to 2011 is detailed in Table A1 of the retrieved document<sup>1</sup>. 2 references Al-generated content may be incorrect 1 Fertility: Overview, 2009 to 2011 - Part 5 2 Report on the Demogr...ility (2012 to 2016) - Part 1 Enter your question: D

Statistique

Canada

Reference × 11-001-X20181564950 Report on the Demographic Situa Canada: Internal migration (2015/ fertility (2012 to 2016) Released a a.m..Eastern time in The Daily, Tue 5.2018 Today, Statistics Canada is updatii chapters on internal migration an fertility in the Report on the Demi Situation in Canada..The first char analyzes internal migration trends Canada in 2015/2016 (July 1 to Ju based on the most recent final es from Statistics Canada's Demogra Estimates Program..The second cl presents the most recent fertility Canada from 2012 to 2016...It prov detailed analysis of the evolution total fertility rate (number of chilc woman), including provincial and variations, average age at childbir including by birth order, complete rates, the annual and daily numbe births, biological childlessness, an births..Interprovincial migration ir lessens in 2015/2016

🖻 Share

Show chat history

In 2015/2016, there were 277,029 interprovincial migr 58

Canadä

# **Dissemination - Table-to-Text (LLM)**

- Create draft articles from data tables for the website
- Possibilities
  - Generic prompts without details
  - Add titles of the draft and table
  - Add examples of articles as style and content guide
- Input from SMEs incorporated into prompts
- Example output





Population Growth: Insights from the 2021 Census Data	Rank	Municipality	Province	Census metropol area in which the municipality is lo
The latest census data underscores the trend of high population growth, not only at a national level but also within Canada's largest municipalities. In 2021, the five municipalities with the highest population were Toronto (2,794,356). Montréal (1,762,949). Calgary				
(1,306,784), Ottawa (1,017,449), and Edmonton (1,010,899).		Toronto	Ontario	Toronto
	2	Montréal	Quebec	Montréal
Between 2016 and 2021, the municipalities that experienced the highest population growth rates were Brampton, Ontario (10.6%),		Calgary	Alberta	Calgary
Oakville, Ofitario (10.5%), Kitchener, Ofitario (10.1%), London, Ofitario (10.0%), and Surrey, British Columbia (7.7%).	4	Ottawa	Ontario	Ottawa-Gatineau
In each province, the municipality with the highest growth rate was Edmonton in Alberta (8.3%), Surrey in British Columbia (9.7%),	5	Edmonton	Alberta	Edmonton
Winnipeg in Manitoba (6.3%), Halifax in Nova Scotia (9.1%), Brampton in Ontario (10.6%), Longueuil in Quebec (6.1%), and Saskatoon in	6	Winnipeg	Manitoba	Winnipeg
Saskatchewan (7.7%).			Ontario	Toronto
Among the top 25 municipalities. Ontario was home to 11. Quebec to 5. British Columbia to 3. Alberta and Saskatchewan each to 2. and		Vancouver	British Columbia	Vancouver
Manitoba and Nova Scotia each to 1.	9	Brampton	Ontario	Toronto
	10	Hamilton	Ontario	Hamilton
The average population growth rate of these top 25 municipalities from 2016 to 2021 was 6.3%. The total population of these municipalities in 2021 was 16,315,480, an increase from 15,482,080 in 2016, indicating a total population growth of 833,400.		Surrey	British Columbia	Vancouver
		Québec	Quebec	Québec
When comparing the population growth rates between Western and Eastern Canada, municipalities in Western Canada had a slightly			Nova Scotia	Halifax
higher average population growth rate (6.8%) than those in Eastern Canada (6.0%).	14	Laval	Quebec	Montréal
The largest municipality in the top 25. Toronto, had a nonulation of 2.794.356 in 2021, which is approximately 13 times larger than the	15	London	Ontario	London
smallest municipality in the top 25, which had a population of 213,759 in 2021.	16	Markham	Ontario	Toronto
			Ontario	Toronto
The range of population growth rates among the top 25 municipalities was 11.1%. Notably, one municipality, Mississauga in Ontario,			Quebec	Ottawa-Gatineau
experienced a decrease in population from 2016 to 2021, with a decrease of 0.5%, from 721,599 in 2016 to 717,961 in 2021.		Saskatoon	Saskatchewan	Saskatoon
In conclusion, the 2021 Census data provides valuable insights into the population growth patterns in Canada. It highlights the population	20	Kitchener	Ontario	Kitchener-Cambri
disparities among municipalities, the varying growth rates across provinces, and the overall steady increase in population. Understanding			Quebec	Montréal
these trends is crucial for policymakers, urban planners, and researchers to make informed decisions and plan for the future. As Canada continues to evolve it is essential to monitor and apply a nonvertice dynamics to ensure sustainable growth and development access the	22	Burnaby	British Columbia	Vancouver
country.	23	Windsor	Ontario	Windsor
	24	Regina	Saskatchewan	Regina
	25	Oakville	Ontario	Toronto
Statistics Statistique Canada Canada	Source	e: Statistics Canal	da, Census of Popula	tion.

Rank	Municipality	Province	Census metropolitan area in which the municipality is located	Population			
				2016 20		21 Growth from 2016 to 2021	
				nun	ber	percent	
1	Toronto	Ontario	Toronto	2,731,571	2,794,356	2.3	
2	Montréal	Quebec	Montréal	1,704,694	1,762,949	3.4	
3	Calgary	Alberta	Calgary	1,239,220	1,306,784	5.5	
4	Ottawa	Ontario	Ottawa-Gatineau	934,243	1,017,449	8.9	
5	Edmonton	Alberta	Edmonton	933,088	1,010,899	8.3	
6	Winnipeg	Manitoba	Winnipeg	705,244	749,607	6.3	
7	Mississauga	Ontario	Toronto	721,599	717,961	-0.5	
8	Vancouver	British Columbia	Vancouver	631,486	662,248	4.9	
9	Brampton	Ontario	Toronto	593,638	656,480	10.6	
10	Hamilton	Ontario	Hamilton	536,917	569,353	6.0	
11	Surrey	British Columbia	Vancouver	517,887	568,322	9.7	
12	Québec	Quebec	Québec	531,902	549,459	3.3	
13	Halifax	Nova Scotia	Halifax	403,131	439,819	9.1	
14	Laval	Quebec	Montréal	422,993	438,366	3.6	
15	London	Ontario	London	383,822	422,324	10.0	
16	Markham	Ontario	Toronto	328,966	338,503	2.9	
17	Vaughan	Ontario	Toronto	306,233	323,103	5.5	
18	Gatineau	Quebec	Ottawa-Gatineau	276,245	291,041	5.4	
19	Saskatoon	Saskatchewan	Saskatoon	247,201	266,141	7.7	
20	Kitchener	Ontario	Kitchener-Cambridge-Waterloo	233,222	256,885	10.1	
21	Longueuil	Quebec	Montréal	239,897	254,483	6.1	
22	Burnaby	British Columbia	Vancouver	232,755	249,125	7.0	
23	Windsor	Ontario	Windsor	217,188	229,660	5.7	
24	Regina	Saskatchewan	Regina	215,106	226,404	5.3	
25	Oakville	Ontario	Toronto	193,832	213,759	10.3	

# **Dissemination - Table-to-Text**

- -SME comments
  - Not sophisticated enough to disseminate
  - Requires more manual work to be edited and verified
  - Texts were repetitive, not insightful and used terminologies incorrectly
- Not a failure as we did learn something
- Refocus
  - Analysts pose questions to LLM for answers
  - Help analysts access relevant data efficiently



# Summary

- Clearly data science has its place in the production of Offical Statistics
- However, the *majority* of uses cases that I have presented would be considered as algorithmic data science
- After the break, Kelly is going to show us how data science methods can be used to make inference



