Data science for official statistics

Sixth Statistical Commission for Africa, 1 to 4 October 2018
Karen Gask, Office for National Statistics, UK

1. Introduction

The amount and variety of data that is available is growing rapidly and at a quicker pace. There is a wider range of data available in many formats, including audio, video, computer logs, text, satellite, purchase transactions, sensors and social networking sites. This has created large, often unstructured datasets that are available, potentially in real time. At the same time, new data science techniques for maximising the value of these newer types of data and other data sources are constantly being developed.

The tools and techniques for analysing these data sets, which are often very large, are also being developed. Data can no longer be managed using spreadsheets, and open source programming languages such as R and Python are becoming widely used by national statistics offices. Distributed processing is required for large datasets, and sophisticated techniques, such as machine learning and natural language processing are required to extract value from these new data sets.

National Statistics Institutes (NSIs) in countries around the world have been working to understand the opportunities and challenges that big data and data science offer to enhance and supplement more traditional statistical processes and outputs.

This paper outlines:

- What data science is
- Examples of using data science from the UK, Canada and Rwanda
- Common challenges and how NSIs can get started.

2. What is data science?

Data science, combined with new data sources, can be used to improve and enhance official statistics, especially improving timeliness and granularity, to develop new insights and richer research, and to improve the efficiency of production processes.

Data science is commonly described as the intersection between mathematical and statistical skills, computer programming, and subject matter expertise. Mathematical and statistical skills are required for building models and understanding concepts such as bias and representativity, whereas computer science techniques are required for storing and processing often very large, diverse types of data (including text or images). Subject matter expertise (for example health, if that is the area of data science focus) is required to understand and interpret real-world problems and solutions.

New data sources and techniques offer opportunities for NSIs in:

- Combining different sources of data to enrich our understanding and improve how we inform policies

- Improving the timeliness of statistics
- Making operational efficiencies (for example, in survey design)
- Automating some processes which are currently done manually
- Filling in known gaps in statistics
- Offering new, often more granular insights.

3. Examples from the UK

3.1. Automated coding of crimes from survey responses

The Crime Survey for England and Wales asks the public whether they have been a victim of crime over the past year. This provides estimates of crime rates for the country which inform the public and policy makers. If a respondent has been a victim of crime, they are asked for a description of what happened and hundreds of further questions. An example of a response is shown in Figure 1. Once the survey data arrives at the office, these responses are manually coded into crime types (such as burglary), taking around 11 weeks of a statistician's time per year.

Figure 1: Example response to the Crime Survey for England and Wales

Before I ask you a number of detailed questions, can you tell me, very briefly, what happened?	Respondent came home and noticed that car was scratched along the side and the wing mirror was broken and hanging off the side of the car
What did they damage?	Car / van
What did they do to the vehicle?	Damaged wing mirrors

Natural language processing is a branch of data science that helps interpret and manipulate human language. Machine learning is the study and construction of algorithms that can learn from and make predictions on data. These techniques were used to research whether some of the manual coding could be automated. The result was that around 40% of cases could be automated, with 97% accuracy that the correct crime code had been used.

The crime team needs to fully test the implementation of this algorithm but it is expected that this will reduce the time taken to manually code crimes from the survey by over 4 weeks per year, and will allow the statistician to focus on more complex coding cases. The project has taken two people ten months to complete, working one day per week.

In this example, the accuracy of the algorithm is only improved by three percentage points by including information from the description of the crime, so most of the accuracy came from the respondent's answers to the closed questions. Therefore, a similar method could be used in an African survey where there is some manual coding to a classification, including those surveys where there is no free text.

Other examples of possible use of natural language processing in an African context could include:

- Improving the matching of names collected in a census to that in a Post-Enumeration Survey, thereby improving methods to assess coverage
- Automating the coding of causes of death from written death certificates.

3.2. Aerial imagery to predict the location of caravan homes

A census is run in the UK every ten years and it is essential to have an accurate list of addresses to count where people are living. However, caravan homes are recorded inconsistently in different sources of data and they are often clustered in caravan parks in holiday and coastal areas of the country. This means than any errors in counting the number of people living in caravans can be concentrated in certain areas.

During the last census, census officers spent a significant amount of time establishing the facts of the number of caravans in a park and whether they were occupied by residents or holiday makers. Using satellite imagery can provide timelier and more cost-effective insight than field intelligence.

A research project was undertaken to assess the feasibility of using Google aerial images to predict whether an image contained a caravan or not. A model was built using machine learning and free software which predicted a caravan with 97% accuracy. However, a 3% error rate would still predict hundreds of false positives across the country, so the algorithm was enhanced to predict clusters of caravans. The project took one person working full-time around four months to complete.

The results from the algorithm were compared against the current address list with very favourable results, highlighting areas where the current address list contains anomalies. The work will be taken forward as part of the next census in 2021. It highlights how data science and aerial imagery can be used to help fill gaps in knowledge.

A method such as this might be used to help support a household level geo-referencing exercise in a census in the Africa. For example, the Pakistan Bureau of Statistics used satellite imagery to assess how much change and dwelling development had occurred before their census in 2017 as their previous census had taken place in 1998. This informed planning of enumeration areas and workloads, reducing the risk of insufficient resource to enumerate areas and thus improving enumeration coverage.

3.3. Applying computer science techniques to regular statistical production

Estimates of expenditure on research and development are required annually by the UK government. These estimates were previously calculated using Microsoft Excel, often by ‘copying and pasting’ complex formulae which ran the risk of introducing errors.

Good software development practices which can be used in statistical production include:

- Using open source and free software such as R or Python
- Using software to help with version control
- Implementing automated testing to help with quality assuring outputs
- Producing automatically generated documentation about what the code is doing.

These practices were used in a project to move the existing system for calculating estimates of expenditure on research and development from Excel to Python. This took two people eight months, working full-time. Automated testing was coded using Python (to check that totals added up correctly for example), and version control software (Git) was used for code development, to ensure that code changes were tracked and reversible if needed.

This project has led to a large efficiency saving. Previously it took two or three people seven weeks to produce the estimates for annual publication, whereas now the code runs in minutes. There is also a significant reduction in the risk of errors being introduced, since it is easy to see who has made changes and when. This means that statisticians are spending less time on routine work and can concentrate on more complex tasks.

This method can be used where statistics are being produced regularly, particularly where they are being produced using Excel. The National Institute of Statistics of Rwanda will be experimenting with this method for their statistics about prices shortly.

4. Example from Canada

In June 2018 the Canadian parliament passed a law to legalise the recreational use of cannabis. Data users indicated that they require economic, health and public safety information to both understand the implications of legalisation and to ensure that the proper policy and regulations are put in place.

Statistics Canada will be using data from existing and new surveys to understand cannabis consumption. However, given the difficulty in obtaining this information, they will supplement this with waste water analysis to measure drug consumption levels in the general population.

Daily samples will be combined over time, from various waste water treatment plants, to understand geographic trends in consumption and trends over time. It is expected that by combining data from multiple sources (traditional and non-traditional), a fuller picture of the impact of cannabis consumption will be provided.

5. Examples from Rwanda

5.1. Reducing the vulnerability of farmers to weather

Risk from weather is a direct cause of chronic poverty for farmers in developing countries. One way to address climate vulnerability is through the provision of agricultural insurance products. However, traditional agricultural insurance products developed for large industrial farms rely on loss verification, which has so far proven too expensive to provide on a large scale to small isolated farmers.

To address this shortcoming, insurance companies in Rwanda have begun to offer insurance products, which pay out when an index that is correlated with crop performance (such as rainfall) exceeds a set threshold level. Despite the high theoretical promise from index insurance, the product only works if the data and model can accurately predict outcomes at a farm level. The accuracy of these indices is still an open question, especially in the context of poor data quality and the microclimates present in Rwanda.

In a recent study, Prof McSharry of Carnegie Mellon University in Kigali, Rwanda et al constructed a forecasting model for production at Rwandan tea estates to test the feasibility of using weather data from satellites to construct an index based insurance product.

They found that it is possible to build a weather based index insurance model with relatively high predictive power (64%), which outperforms local ground weather stations by 33%. However, a large

amount of the variation in production remains unexplained and microclimates are still likely to be a concern.

5.2. Predicting malaria outbreaks in Kigali

Epidemiological studies traditionally focus on medical records from hospitals and clinics to monitor disease. However, not all patients go to a clinic when suffering from an illness; many go directly to a pharmacist where they obtain drugs based on the symptoms they present.

Using satellite data about rainfall and point-of-sale data from pharmacies on anti-malarial drugs, it is observed that monthly sales of anti-malaria drugs are best predicted by cumulative rainfall from two months previously. This is because it takes around one month for eggs in stagnant water to grow into an adult mosquito and a further month for malaria symptoms to arise, causing the infected person to visit a pharmacy.

The model facilitates real-time monitoring and probabilistic forecasts that can both inform policymakers and help individuals that may want to take preventative medication during the months where the risk of infection is highest.

6. How to build data science capability

Building data science capability requires not only building coding and data science skills, but also the right IT infrastructure, the right legal, policy and good practice environment, especially for data sharing, and the development of real-world data science projects to demonstrate value.

This presentation draws on the experience of the Data Science Campus of the UK's Office for National Statistics to talk through the requirements, challenges and opportunities for developing and implementing data science and big data strategies. This is based on the Campus' experience of leading on the development of data science capability across UK Government, and internationally, especially in Rwanda.

There is no one-size fits all data science solution for all countries and applications, so we will present a step-by-step guide to the areas to consider, and some examples of how the challenges in those areas have been successfully addressed.