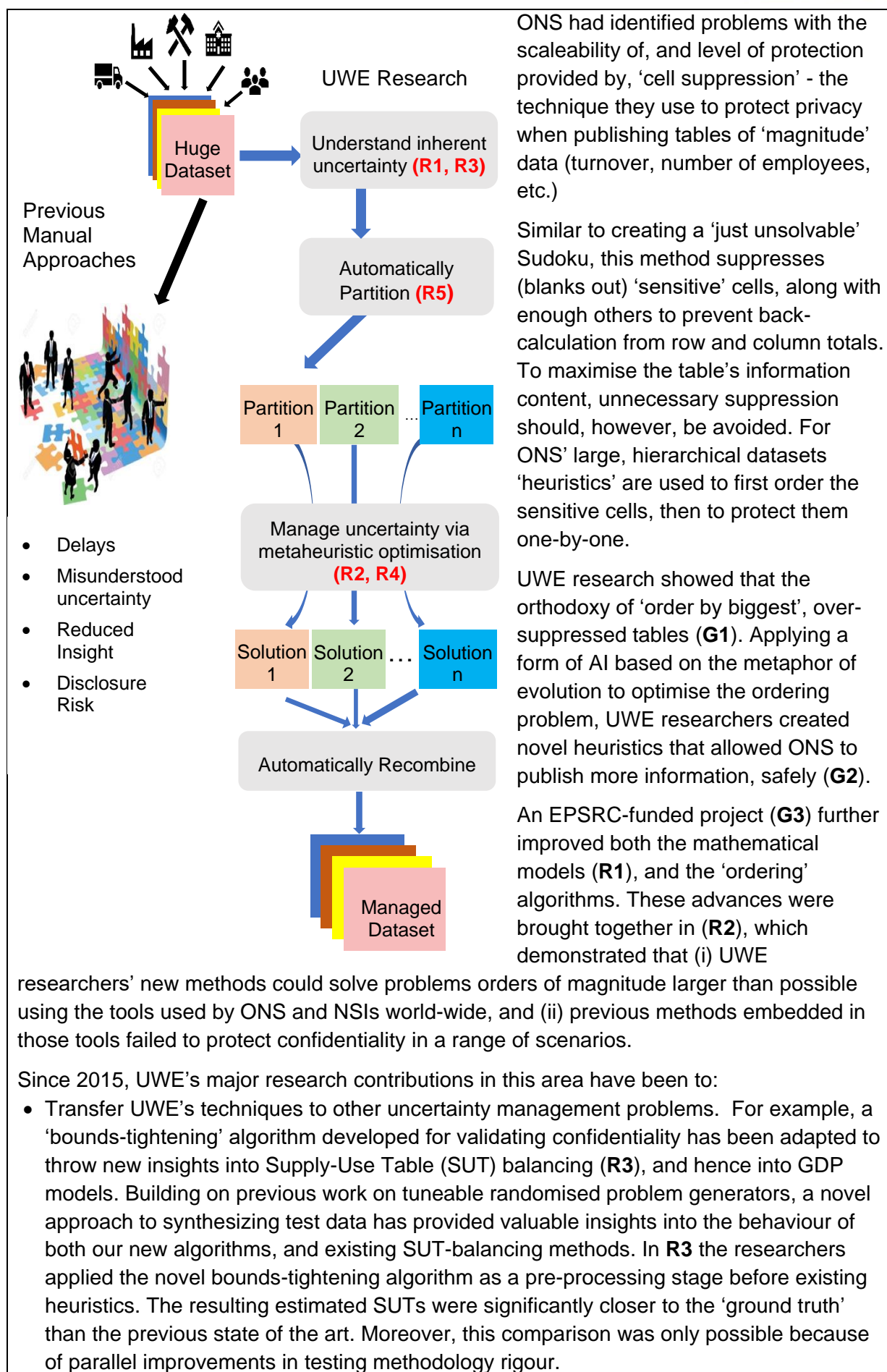


Impact case study (REF3)

Institution: University of the West of England, Bristol		
Unit of Assessment: 11		
Title of case study: Using AI to improve government management of uncertainty in published statistical data		
Period when the underpinning research was undertaken: 2007 – 2019		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Professor Jim Smith	Professor of Interactive AI	1996 – present
Dr Martin Serpell	Senior Lecturer	2011 – present
Dr Alistair Clark	Associate Professor	1998 – present
Period when the claimed impact occurred: 01.08.2013 - 2020		
Is this case study continued from a case study submitted in 2014? No		
1. Summary of the impact		
<p>The UK's Office for National Statistics (ONS) have adopted techniques developed at UWE into their toolchain for creating publishable information from large complex datasets. Replacing a time-consuming and potentially risky combination of manual effort and less effective heuristics with AI-based methods, has increased the speed and accuracy for tackling a range of core tasks related to managing confidentiality and uncertainty. The ONS has benefitted from these changes in processing methods through: improved protection of confidential data for the National Census and key headline economic indicators such as employment statistics; and more accurate estimates of Gross Domestic Product, which are used by decision makers at the Treasury and Bank of England.</p>		
2. Underpinning research		
<p>A key role of National Statistical Institutes (NSIs) is to collect data from a range of sources and summarise these in published tables to inform their nation's decision-makers. The collected data is first merged into huge datasets, and then processed to manage uncertainty prior to publication, either as periodic reports (e.g. employment, trade figures, Gross Domestic Product: GDP), or in response to specific queries. In the UK, ONS balances several conflicting demands: (i) providing information that is detailed and timely; (ii) understanding potential limits on accuracy; and (iii) the duty to protect privacy. Demands (ii) and (iii) can be modelled as problems in 'uncertainty management', but the large size of data sets make 'exact' mathematical techniques unusable. Thus, typically, they are manually decomposed into smaller 'partitions' that are then processed and rebuilt. However, this process is lengthy and risks introducing errors which can be extremely damaging. ONS' statutory duty to protect individual respondents' privacy also far exceeds GDPR, and loss of goodwill negatively impacts the quality and timeliness of survey data collected in the future.</p> <p>UWE research has created new ways for automatically reasoning with, and managing, uncertainty in huge data sets – whether that uncertainty arises from the way data is collated and linked, or is deliberately introduced in order to preserve confidentiality.</p>		

Impact case study (REF3)



Impact case study (REF3)

- Produce an open-source algorithmic framework for uncertainty management in statistical data in general, focussing on performance enhancements to improve the ability to tackle large-scale problems. It includes novel developments of ‘data-driven models’ for optimising the ordering problem (R4), and automatic algorithm adaption to dynamically exploit available hardware resources (G4).
- Automate the process of decomposing large datasets into ‘partitions’ small enough to be processed. Here UWE researchers have developed adaptive coarsening thresholds and built on our experience of hybrid evolutionary algorithms to improve on the state of the art in multi-level hypergraph partitioning (R5).
- Increase algorithmic stability and reduced run-times for cell-suppression algorithms from weeks to a few days (R3, R5).

3. References to the research

R1 Serpell, M., Smith, J., Clark, A., Staggemeier, A. (2013) A preprocessing optimization applied to the cell suppression problem in statistical disclosure control. *Information Sciences*, vol 238, pp 22–32. <https://doi.org/10.1016/j.ins.2013.02.006>

R2 Smith, J., Clark, A., Staggemeier, A. and Serpell, M. (2012) A Genetic Approach to Statistical Disclosure Control. *IEEE Transactions on Evolutionary Computation*, vol 16(3), pp 431-441. <https://doi.org/10.1109/TEVC.2011.2159271>

R3 Serpell, M. (2017) Incorporating Data Quality Improvement Into Supply-Use Table Balancing. *Economic Systems Research*, vol 30(2), pp 271-288. <https://doi.org/10.1080/09535314.2017.1396962>

R4 Smith, J., Stone, C. and Serpell, M. (2016) Exploiting diverse distance metrics for surrogate-based optimisation of ordering problems. *Proc. of the ACM-SIGEVO Genetic and Evolutionary Computation Conference*, pp 701-708. <http://dx.doi.org/10.1145/2908812.2908854>

R5 Preen, R., and Smith, J. (2019) Evolutionary n-level Hypergraph Partitioning with Adaptive Coarsening. *IEEE Transactions on Evolutionary Computation*, vol 23(6), pp 962-97. <https://doi.org/10.1109/TEVC.2019.2896951>

Evidence of the quality of the underpinning research

G1 Clark, A. *Improvements to Cell Suppression in Statistical Disclosure Control*, Office for National Statistics, 2005 – 2006, £20,100.

G2 Smith, J. *Evaluation of Heuristic Approaches to Statistical Disclosure Control*, Office for National Statistics, 2007, £4,935.

G3 Smith, J. *Mathematical Modelling of Statistical Disclosure Control*, EPSRC, 2007 – 2011, £63,000.

G4 Smith, J. *Application of Meta-heuristics for Statistical Disclosure Control of Business Data*, Office for National Statistics, 2013 – 2014, £7,500.

4. Details of the impact

UWE research and tools developed from this have significantly changed ONS data processing practices. This has resulted in improved confidentiality reducing the risk of data breaches, reduced processing times, and improvements in data quality. This in turn has benefitted users of ONS data across government, policy-makers and businesses.

Improved data confidentiality and reduced processing times

ONS Business Related Employment Statistics (BRES) is a huge complex dataset, detailing employment by area, gender, full/part time, and business type. Producing a set of

publishable tables from this highly complex dataset is time-consuming and takes staff away from other tasks that could be improving the accuracy of the data (S1). UWE researchers conducted confidential work for the ONS and showed they were able to completely expose every protected cell in the 2010 BRES publication (S2). This was evidence that ONS were publishing extremely vulnerable datasets and risking serious breaches of confidentiality. ONS' Head of the Labour Demand Branch commented that this research had raised awareness *'of the deficiencies of existing toolkits (even on small datasets) and hence our working practices around them'* (S1). In response, ONS commissioned UWE (G4) to investigate integrating UWE algorithms into ONS workflows for protecting BRES publications. ONS subsequently licensed and deployed UWE's 'unpicker' algorithms (S3, S4), providing ONS with: *'tools that automatically validate the protection applied to datasets pre-publication. These have been used as part of our tool chain for the BRES data'* (S1).

UWE research has, for the first time, made it possible for ONS to automatically protect large tables. In September 2019, ONS took the decision to *'deploy the UWE tools for protecting the next and subsequent releases of BRES data prior to publication'* (S1). As S1 notes, the impacts of this work have included:

'decreased risk of the exposure of confidential information, decreased risk of inadvertently introduced errors during manual decomposition-reassembly of data sets, and the opportunity to devote more time to other activities (reconsidering sampling methodologies etc.) that might improve data quality' (S1).

UWE research has also helped to improve the confidentiality of the National Census, one of ONS' 'flagship' activities. In 2018, the team responsible for designing the methodology of the 2021 census approached UWE to test whether their existing tools sufficiently protected the data in the census (S5). UWE applied their tools to subsets of protected data provided from the 2011 census, exposing between 92% and 99% of cells to within +/- 2 of their true values (S6). This significantly influenced the ONS team's approach to the 2021 census. A Principal Methodologist at ONS commented:

'As a result, we have decided to adapt our planning for the protection of the 2021 census... we have also decided to incorporate UWE's algorithms into our development tool chains, to provide us with valuable extra rigour in testing our methods as they evolve' (S5).

Improvement in data quality and user-benefit

Economic indicators such as employment and trade figures, and GDP, are very widely used by government departments and agencies including the Treasury and Bank of England, the finance sector more generally, investors, businesses and other analysts. UWE's research has contributed to ONS developing models that account for uncertainty in ways that are more rigorous than before, and which better facilitate targeted effort in acquiring evidence to refine the models.

Published national GDP is derived from the balanced Supply and Use Tables (SUTs), core data for the ONS and other national statistics institutes, and provides a detailed picture of the economy, split into various industrial sectors. External users base their analyses on these data (S7). Collecting the data and constructing the SUTs is both expensive and time consuming. Consequently, deriving algorithms to automate this is *'a key part of ONS Economic Statistics and Analysis Strategy in order to deliver efficiencies and improve the quality of the SU solutions and, therefore, the estimate of GDP'* (S7).

Impact case study (REF3)

UWE researchers integrated boundary-tightening techniques as a pre-processing step of the SUT balancing problem, and developed more rigorous methods for quantifying uncertainty (R3). According to ONS' Deputy Director of the Economic Statistics Change division, the research: *'helped inform ONS understanding and parameter setting with respect to the boundary tightening approach and its application to large sets of data'* (S7). These techniques also enable greater mathematical rigour, leading to better data, reduced processing times and therefore cost-savings. UWE research:

'contributed to the distinctive (compared to other National Statistics Institutes) functional form of the current ONS model, specifically its use of upper and lower bound parameters rather than the standard subjective variance parameters' (S7).

Cross-governmental impacts

The 2018 National Statistician's Quality Review *'researched current methods, understood the challenges and set out a plan for future investment in the field of privacy and confidentiality methods'* (S8). The Head of the Government Statistical Service at the time *'welcomed the findings'* of this Review and stated it would:

'help the Government Statistical Service (GSS) to take full advantage of the cutting-edge developments and research conducted by world leading experts from across academia and the private sector' (S9).

UWE researchers were invited to contribute a chapter to this Review (S10) and present its recommendations to a meeting of the GSS Methodology Advisory Committee (this encompasses ONS, NHS, HMRC, DWP, Ministry of Justice, and other national statistic institutes). The GSS established a cross-department task force to implement the recommendations:

'One of the recommendations influenced by Professor Smith's research was "Explore the practical applications of machine learning and artificial intelligence protection of privacy and confidentiality and options for avoiding the replication of human biases"... A GSS task force was established [2019] to implement these recommendations consisting of experts across government and heads of statistical professions from some government departments' (S8).

5. Sources to corroborate the impact

S1 Testimonial from Head of ONS Labour Demand Branch

S2 Confidential report for ONS *'The UWE Cell Suppression algorithm and its application at ONS'*

S3 Licence contract for ONS to use UWE software

S4 Memorandum of understanding between ONS and UWE

S5 Testimonial from Principal Methodologist, Methodology: Census and Population Statistics Hub, ONS

S6 Confidential report for ONS *'Census Protection examination'*

S7 Testimonial from Deputy Director of Economics Statistics Change Division, ONS

S8 Testimonial from Principal Research Officer, ONS

S9 Government Statistical Service webpage quoting the National Statistician (2014-2019), forward to NSQR, available [online](#)

S10 Ritchie, F. & Smith, J. (2018) [Confidentiality and linked data](#). Paper published as part of *The National Statistician's Quality Review*. London, December 2018. Also available as <https://arxiv.org/abs/1907.06465>