Institution: University of Southampton		
Unit of Assessment: 10 Mathematical Sciences		
Title of case study: 10-08 Improving population censuses around the world		
Period when the underpinning research was undertaken: 2006 – 2020		
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:**
Paul Smith Peter van der Heijden Li-Chun Zhang James Brown	Professor of Official Statistics Professor of Social Statistics Professor of Social Statistics Reader in Sampling Statistics	September 2014 – present October 2012 – present September 2012 – present September 1999 – January 2008; October 2010 – August 2013
Period when the claimed impact occurred: August 2013 – December 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

Statistical techniques developed at the University of Southampton (UoS) have informed the design of recent and upcoming population censuses in the UK, Netherlands, Ireland and New Zealand. Key impacts since August 2013 include:

- In the UK, the design of the 2011 censuses has contributed to an estimated £500m of census benefits in a range of sectors in the decade following publication in 2012/13. UoS's enhanced coverage survey design for the 2021 census is contributing to savings of £750,000 in survey costs.
- In the Netherlands, improved methods following up the 2011 "virtual census" have had both national benefit and an improvement in the quality of their population estimates 2011-16 as reported to the EU, leading to an increase in Dutch seats in the European Parliament.
- In Ireland the trimmed dual system estimator developed at UoS has been used to evaluate the quality of administrative sources, particularly their overcoverage, in the development of an administrative data-based population census in the absence of a person register.
- In New Zealand, UoS has worked to counter deficiencies in estimating the Māori population, which is critical to meet crown obligations under the Treaty of Waitangi and to ensure policies such as those in health are inclusive of this disadvantaged group.

2. Underpinning research

The periodic population census is a crucial component of a nation's statistical system, providing an opportunity to produce detailed statistical estimates nationally and for small domains (both geographical areas and subpopulations). These statistics have many uses, including for local planning and monitoring, as a component of population estimates and social statistics, and for formation and evaluation of government policy. They have a long shelf-life because the census is expensive (e.g. the England & Wales census in 2021 will cost around £900m) and only undertaken periodically. Traditional census enumeration involves asking every household to complete a questionnaire, but some countries, including the Netherlands, use a census based on administrative data. This reduces field costs and potentially improves the frequency of estimates, and many more countries are considering similar approaches.

The research portfolio encompasses a suite of methods for estimating population sizes (overall and in subgroups) using imperfect data. Traditional census enumeration suffers from selective non-response, and it is often groups of policy interest which are under-represented. Appropriate designs to collect relevant data and estimation approaches [**3.1**, **3.5**] are therefore needed to generate high quality estimates. Administrative data sources are prone to both under- and over-

coverage, and the research also presents statistical approaches for dealing with these deficiencies [**3.2**, **3.3**, **3.4**, **3.6**].

The first component of Southampton's research is on the efficiency of the methods for sampling and estimation. In 2006-7 Professor James Brown (with Professor Paul Smith who was at the ONS at the time) designed a coverage survey to give both the best overall accuracy and minimum accuracy standards in defined domains for the 2011 population census in England & Wales, and developed a cost model which was used to investigate the costs of different approaches to clustering of the fieldwork commensurate with practical field implementation [**3.1**]. The main elements of the coverage survey design and the cost model were developed while Brown was at Southampton, with detailed work on specific cost options and write-up taking place after Brown moved to IoE. Brown worked on census estimation while at Southampton 2010-13, and after Smith's arrival at UoS in 2014, he and Brown (now at the University of Technology Sydney) expounded a framework for census estimation using the coverage survey information and adjusted for the situation in the 2011 UK censuses [**3.5**]. The framework is based on capture-recapture methods, with robust ratio estimation to extrapolate from the sample areas to the population.

The second research component explores the robustness of a population census based on administrative data sources, as used in the Netherlands and elsewhere, to estimate population totals. Immigrant populations with specific nationalities are investigated in [**3.2**] and [**3.3**], which uses police records as a third source (alongside the person register and an employment register) to estimate the size of the non-registered population. This gives primary evidence of the size of some hard to count populations, which are typically missing in administrative data sources, and how these can be accounted for in population census estimates. Professor Peter van der Heijden used data from the Netherlands census to explore alternative approaches to estimating the population size in an administrative census, focusing particularly on the imputation of missing usual residence information [**3.2**]. Usual residence is missing not at random in the linked registers, so standard non-response assumptions are not effective; therefore there are advantages to taking a restricted range of donor records in estimating the usually resident population total. Van der Heijden explored several approaches based on capture-recapture estimation and multiple imputation using predictive mean matching.

Administrative data sources are prone to overcoverage (often because there is no requirement for someone to deregister when they move away), which violates the standard assumptions of dual system estimation (DSE) and is challenging to deal with. Professor Li-Chun Zhang proposed an innovative method [**3.4**] to work with dual system estimation in this context through progressive trimming of the datasets, leading to the *trimmed DSE*.

Smith and van der Heijden extended the use of multiple system estimation, with additional covariates observed in only some sources, to linked administrative sources [**3.6**]. They investigated estimation of the size of the Māori population in New Zealand, where ethnicity is available in each of four sources, but with inconsistent definitions. They derived a general result to deal with the partial coverage of the target population of some of the sources, and extended the methods to impute ethnicity data which were not provided, as well as data that were structurally missing through the linkage of different sources. They also developed an extended form of latent class analysis to deal with the missing information. This was used to generate a single consensus estimate of (in this case) the Māori population in New Zealand, and estimates of the error with which the different administrative sources measured this concept.

Smith, van der Heijden and Zhang all have international reputations from this and previous census-related research, and have been commissioned to assist in the methodological development of the population census in the Office for National Statistics (UK) and Statistics Netherlands (Netherlands) during 2013-2020.

3. References to the research

3.1 Brown, J., Abbott, O. and Smith, P.A. (2011) Design of the 2001 and 2011 Census Coverage Surveys for England and Wales. *Journal of the Royal Statistical Society: Series A* 174 881–906. https://doi.org/10.1111/j.1467-985X.2011.00697.x

3.2 Gerritse, S.C., Bakker, B.F.M. & **van der Heijden, P.G.M.** (2015) Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the Netherlands. *Statistical Journal of the IAOS* 31 613–627. https://doi.org/10.3233/SJI-150938.

3.3 Bakker, B.F.M., **van der Heijden, P.G.M.** and Gerritse, S.C. (2018) Estimation of non-registered usual residents in the Netherlands. In D. Bohning, P.G.M. van der Heijden & J. Bunge (Eds.), *Capture Recapture Methods for the Social and Medical Sciences* (pp. 259-273). CRC Press.

3.4 Zhang, L.-C. & Dunne, J. (2017) Trimmed dual system estimation. In D. Bohning, P.G.M. van der Heijden & J. Bunge (Eds.), *Capture Recapture Methods for the Social and Medical Sciences* (pp. 239-259). (Chapman & Hall/CRC Interdisciplinary Statistics). CRC Press. https://doi.org/10.4324/9781315151939 .

3.5 Brown, J.J., Sexton, C., Abbott, O. & **Smith, P.A.** (2018) The framework for estimating coverage in the 2011 Census of England and Wales: combining dual-system estimation with ratio estimation. *Statistical Journal of the IAOS*. https://doi.org/10.3233/SJI-180426.

3.6 Van der Heijden, P.G.M., Cruyff, M., **Smith, P.A.**, et al. (under review) Multiple system estimation using covariates having missing values and measurement error: estimating the size of the Māori population in New Zealand. Available from https://arxiv.org/abs/2007.00929.

4. Details of the impact

The University of Southampton (UoS) has a long history of working with census agencies, informing best practices in conducting censuses, as well as designing coverage surveys and estimation procedures that adjust for undercoverage and improve the quality of the census outputs.

UK Censuses

Before the 2011 population census in England and Wales, UoS worked with the Office of National Statistics to design the census coverage survey and the methods for estimating the population size; these methods were also applied in censuses in Scotland and Northern Ireland. With the next census not undertaken until 2021, the improved accuracy has benefits across the entire period 2013-2020. As noted by the UK Statistics Authority in their 2015 Special Assessment of the census: “The UK’s decennial population census is central to decisions in all areas of society – whether by businesses, councils, the health service or charities. It is the basis of population estimates; it underpins funding formulae; it provides insight into the wellbeing and needs of communities throughout UK society” [**5.1**].

Use of these approaches led to high quality census outputs published July 2012 – March 2013. The quality assessment of the census [**5.2**] said “An Independent Review of Coverage Assessment, Adjustment and Quality Assurance methodology took place in 2011. The review team stated that ‘the further procedures for Quality Assurance and [coverage assessment and] adjustment significantly strengthen ONS's strategy for successful population estimation’. Outputs cover statistics relating to the number of people, where they live, their characteristics and their housing. Beneficiaries throughout the decade include [**5.3**]:

- **Government**, whose uses include the calculation of local government funding, and planning by local authorities themselves for vital services such as schools, transport, energy, waste and emergencies.
- **The public sector**, whose uses include diversity monitoring, and data for The National Archives, who consider the census “one of the most important datasets we have”, for genealogists or individuals researching family history.

- **Businesses**, whose uses include market research and decisions on location and size of local branches.
- **Third sector and community groups**, whose uses include allocation of funding for parishes and faith-based services.

In their census benefits evaluation report [**5.4**] the ONS identified £500m of benefit from the census outputs, among which central government benefits are "significantly underestimated".

Looking to the next UK censuses, UoS research has been central to the development of the coverage estimation and adjustment methods for 2021. The coverage survey design methods and cost model have been developed with the expert input of Smith to investigate less clustered and therefore more efficient designs, and coupled with a geographic tool to optimise the use of enumerator resources. The Head of Census Methodology at the ONS stated: "The survey design has also been adapted in cooperation with Paul Smith from Southampton so that it is optimised for the situation expected in 2021, which has included revising the hard-to-count index used in stratification to account for expected changes from an on-line-first approach to data collection and the challenges of digital exclusion in some subpopulations." This contributes to expected savings of £750,000 in the costs of the Census Coverage Survey. [**5.5**]

The coverage adjustment methods have also been extended, with expert input from Smith and van der Heijden, to take advantage of the faster availability of data that online response provides. These methods are also applied in the 2022 Scotland and 2021 Northern Ireland censuses by National Records of Scotland and Northern Ireland Statistics and Research Agency respectively.

Netherlands virtual census

Statistics Netherlands undertakes a "virtual census" – it uses only administrative data (principally from a person register) and does not directly contact individuals. In 2016, Statistics Netherlands implemented the methods developed by van der Heijden to estimate the *unregistered* usually resident population, based on the 2011 virtual census and then updated for 2013, 14, 15 and 16 [**5.6**]. This demonstrates the quality of estimates from the virtual census, which increases confidence among users of the census outputs. It is also a legal requirement to report the census quality to the European Union (EU), and this affects representation in the EU's institutions – a Senior Methodologist and former Head of Team Methodology at Statistics Netherlands said: "It is important for the Netherlands that the number of usual residents reported to EUROSTAT is correct. It is used for the distribution of seats in European Parliament over the different member states. As the number of usual residents is underestimated using the population register only, the applied methodology of Van der Heijden has led to more seats in European Parliament and therefore political influence in Europe for the Netherlands." [**5.7**]

Transition to an administrative data based population census

A range of countries including the UK, Ireland, Italy and New Zealand are all seeking to move from periodic large-scale censuses to a virtual census similar to the Netherlands model, based on administrative datasets. This has significant advantages in reduced and (crucially) more evenly spread costs, and also in the potential to update census estimates much more frequently. It is a complicated transition even in countries with a person register; UK, Ireland and New Zealand do not have one, and rely on linked administrative data sources only, which suffer from overcoverage. UoS's trimmed dual system estimator (TDSE) has already been implemented in Ireland as part of the evaluation of the quality of administrative sources for use in an administrative data census. The Assistant Director General at the Central Statistics Office (CSO) Ireland stated: "The research at Southampton has acted, and is continuing to act, as a significant catalyst in the transformation of population estimates at CSO, Ireland. In particular, the research at UoS has made realistic the compilation of Census like population estimates from administrative data in the absence of a Central Population Register." [**5.8**]

The CSO also note that "the TDSE methodology has attracted significant interest among NSIs [national statistics institutes]". This method has already made an impact in New Zealand, where Statistics New Zealand stated: "Trimmed DSE techniques were used to investigate potential over-coverage in our administrative NZ resident population. We applied the technique in multiple ways and this enabled us to rule out some possible mechanisms as leading to over-coverage,

and provided insight into the over-coverage influences of various data sources". New Zealand ran a population census in 2017 which had significant fieldwork challenges and did not produce good quality population estimates. Methods based on administrative data and the census were used as a fallback, so TDSE was an important tool in the recovery from the poor census operation.

The ONS are also using TDSE, which "has been implemented in the research outputs" which are being produced in the expectation of moving to an administrative data based census after 2021 [**5.5**]. Smith has been working with ONS on the design of a coverage survey for an administrative data census in England & Wales, building on similar design work for the Census Coverage Survey.

New Zealand ethnic population estimates

The operational deficiencies in the census operation in 2017 in New Zealand meant that the census did not provide robust estimates of the Māori population. These are critical to meet crown obligations under the Treaty of Waitangi, and because the Māori are a disadvantaged group and therefore of key policy interest (for example in health outcomes). Van der Heijden and Smith have been working with Statistics New Zealand, exploring the use of linked census and administrative data to estimate the size of the Māori population. A Principal Statistician at Statistics NZ stated that this "offers a new approach to estimating the Māori population which makes use of census information in combination with additional information from administrative sources. Two features of the method make it of considerable interest: (a) it should be robust to the coverage challenges in the census and (b) the results provide useful insights into the mis-classification errors of the contributing (census and administrative) sources. … Further exploration of this approach is expected in the future." [**5.9**] The approach also facilitates exploration of the effect of using only the administrative data without a census, which provides valuable evidence in the transition to an administrative data census.

5. Sources to corroborate the impact

5.1 UK Statistics Authority, Special Assessment of the 2011 Censuses in the UK: Phase 3 https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-assessmentreport3182011censusphase_tcm97-45033.pdf

5.2 Office for National Statistics, UK: Quality and Methodology Information, 2011 Census Statistics for England and Wales: March 2011 http://www.ons.gov.uk/ons/guide-method/method-quality/quality/quality-information/population/population-and-household-estimates.doc

5.3 Office for National Statistics, UK: How others use census data https://www.ons.gov.uk/census/2011census/2011censusbenefits/howothersusecensusdata

5.4 Office for National Statistics, UK: 2011 Census Benefits Evaluation Report (last updated October 2017) https://www.ons.gov.uk/census/2011census/2011censusbenefits/2011censusbenefitsevaluationreport

5.5 Letter from Head of Census Methodology, Office for National Statistics, UK 26/10/20.

5.6 Statistics Netherlands, Usual Residence Population Definition: Feasibility Study, The Netherlands 30/12/16. https://www.cbs.nl/-/media/_pdf/2017/08/statistics-netherlands-feasibility-study.pdf

5.7 Letter from Senior Methodologist and former Head of Team Methodology at Statistics Netherlands 09/10/19 .

5.8 Letter from Assistant Director General, Central Statistics Office, Ireland 19/10/20.

5.9 Letter from Principal Statistician, Statistics New Zealand 16/10/20.