UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Title : Intruder testing – an empirical measure of the quality of Census 2021 England and Wales Disclosure Control methods

Author(s) Samantha Trace (Office For National Statistics) Dominic Nelson (Office For National Statistics

e-mail samantha.trace@ons.gov.uk

Abstract

By law, the Office for National Statistics (ONS) must protect the confidentiality of respondents to Census 2021. We protected the confidentiality of individuals' data in three ways: swapping records between areas, applying a cell key method to each table, and applying disclosure rules in deciding which tables could be published. To assess the effectiveness of these methods and provide assurance, an intruder test was performed on Census 2021 data using a secure version of the outputs system. 51 intruders were recruited to attempt to identify individuals in the planned data outputs. 30 Intruders took part, 81 claims were made, and more than half of these claims (41/81) were incorrect. Further steps were taken reduce the risks identified by the test, making the data the majority of these claims were made from no longer possible to access through the Create a Custom Dataset system. This gave the Office for National Statistics evidence there was sufficient uncertainty in the data to meet the standard required by legal guidance and we would meet our ethical duty to protect confidentiality.

1 Introduction

The Office for National Statistics (ONS) has legal obligations under the Statistics and Registration Service Act (SRSA, 2007) Section 39 and the Data Protection Act (2018) that require the ONS not to reveal the identity or private information about an individual or organisation.

We have a pledge to respondents that the information will only be used for statistical purposes, so we must look after and protect the information that is provided to us. Moreover, a breach of disclosure could lead to criminal proceedings against an individual who has released or authorised the release of personal information, as defined under Section 39 of the SRSA.

The SRSA defines "personal information" as information that identifies a particular person if the identity of that person:

- is specified in the information
- can be deduced from the information
- can be deduced from the information taken together with any other published information

Therefore, in order for data to be released, the risk of identifying individuals from it, potentially with additional publicly available information, must be minimal.

Intruder testing is an empirical test to check that the measures applied to make data sufficiently difficult to identify individuals within have been successful. This involves recruiting 'friendly intruders' who emulate the actions of potential 'real intruders' upon the data.

The standard that needs to be met is suggested by the National Statistician's Guidance, "the design and selection of intruder scenarios should be informed by the means likely reasonably to be used to identify an individual in the statistic".

So, intruder tests are designed to measure what could be done with the means likely to be available to an opportunistic attacker, it does not have to cover every imaginable scenario, just the most probable.

The 2011 Census outputs were tested in this way, and the findings were useful in providing assurance that the disclosure controls measures used on the data were adequate, and provided evidence to what further steps should be taken to further reduce disclosure risk. Other ad-hoc exercises have been undertaken by the ONS as required since, with the same purpose – to determine the level of identification risk in a dataset.

For Census 2021, new disclosure control methods were required for a new output system. On top of the imputation of missing records done to make the Census as representative as it can be, which also adds doubt as to whether a particular record is 'real' or not, there were new measures in place to protect the data:

- Targeted Record Swapping – swapping households that are marked as unique in the data with a similar record in the local area. The geographies were changed for between 7% and 10% of households, and for between 2% and 5% of individuals in communal establishments.
- Cell Key Perturbation - this adds noise to the figures, making slight changes to cell counts including zero cell counts, by a method which means that where the same records are presented in a cell, the number should remain consistent. A typical dataset would have around 14% of cell counts perturbed by a small amount, and small counts were more likely to have been perturbed than large counts.

- Disclosure rules (in the Create a Custom Dataset system) – automated rules including measures of how many small counts are in the table, that can stop data being given for an area.

These methods were intended to combine as a 'lighter touch' approach, allowing some detail to be possible at low level geography, whilst maintaining the usefulness of the data within the new Create a custom dataset (CACD) system, and other census outputs. The CACD system allows users to create their own multivariate datasets, so the rules are set to prevent the possibility of identifying a single record and building up a list of potential attributes. The level of identification risk should still be minimal, using information public or private.

2 The Intruder Test

2.1 Method

51 intruders, all ONS employees, were recruited. All had appropriate security clearances and consented to an enhanced non-disclosure agreement. They were given training on how to use the output system, and possible methods of working against our statistical disclosure controls. A safe area of an approved file management system was set up, and they were given access to individualised folders to record their findings and keep notes.

A version of the planned outputs system was created on a secure internal-access platform and loaded with the usual resident database. This is the main basis for Census outputs as it includes all people who are 'usually resident' at the enumeration address at the time of the census. This was also programmed with all the current planned variables and classifications for those variables. A version of the planned statistical disclosure rules was placed in this system, to auto-control outputs requested by intruders, and deny access if the output does not pass these rules. The system had built in perturbation so automatically created outputs with some values slightly changed.

The data placed in the system had targeted swapping already applied and imputed records present, just as it would be when published. The main census 2021 geographies were available in this system, the smallest geography used was output area (OA), an area with at least 100 persons in it, though more typically 400 persons.

Intruders were given individual access to the system, encouraged to collaborate on a private Teams channel, and to share resources, such as web pages, hints and tips. An errors log was set up to record system issues, and the details of the claim, including geography, variables and classifications used, as well as the name and address of the individual being claimed as found, and the confidence level in the identification as a percentage.

Claims were transcribed from the individual file folders to a single sheet that the checkers had access to. These checkers were from a different team to ensure the data was fully firewalled from the intruders, and no actual disclosure would result from the exercise.

The checkers had access to record level data, so could determine whether a claim was correct, partial, or incorrect. A correct claim would match on name and approximate address. Inaccurate address matches were counted as correct so long as they would have been within the geographical area used to make the claim.

Inaccurate name matching was counted as incorrect. A partial match would be where a claim was made on a 1 in a cell, where more records would have been in that cell but were perturbed down to 1.

.

2.2 Limitations

We had considered engaging a third party to take part in the test, however we could not be sure of start time, and there are few companies engaged in exactly this sort of testing that could have gained security clearances in time, so it was deemed impractical to engage a third party in this exercise. Therefore, there may be some organisational biases in our exercise.

Although attempts were made to recruit people from more sparsely populated areas of England and Wales, most people were still clustered geographically around ONS offices and reflect the socio-demographic mix of ONS staff rather than the general population.

Intruders also had to use their spare time around their regular work, and the exercise ran in August when many took leave, although it took place over three weeks to allow more people to participate.

The dataset looked at was not the full range of planned Census outputs. The final system includes not just Usual Resident, but also Usual Residents in Households and Communal Establishments, Households and Household reference Persons. The Usual Resident dataset used was taken to be a sufficient test of the general level of risk in the data.

2.3 Results

2.31 Claims

81 identification claims were made, excluding duplicates. These claims are where an intruder highlighted a '1' cell count in a dataset, and gave the details of this, and claimed they knew which person it related it to. Some (2) claims listed various methods to approach the same identification, in these cases this was still counted as one claim and measures such as cell count were taken from the first tables stated.

40/81 or 49% of identification claims were correct (the intruder correctly named an individual in a cell)

8/81 or 10% of identification claims were *partially* correct (the intruder correctly names an individual in a cell of apparent size 1, but the cell count is greater than 1 – due to cell key perturbation – the cell could have been representing any of the people in it)

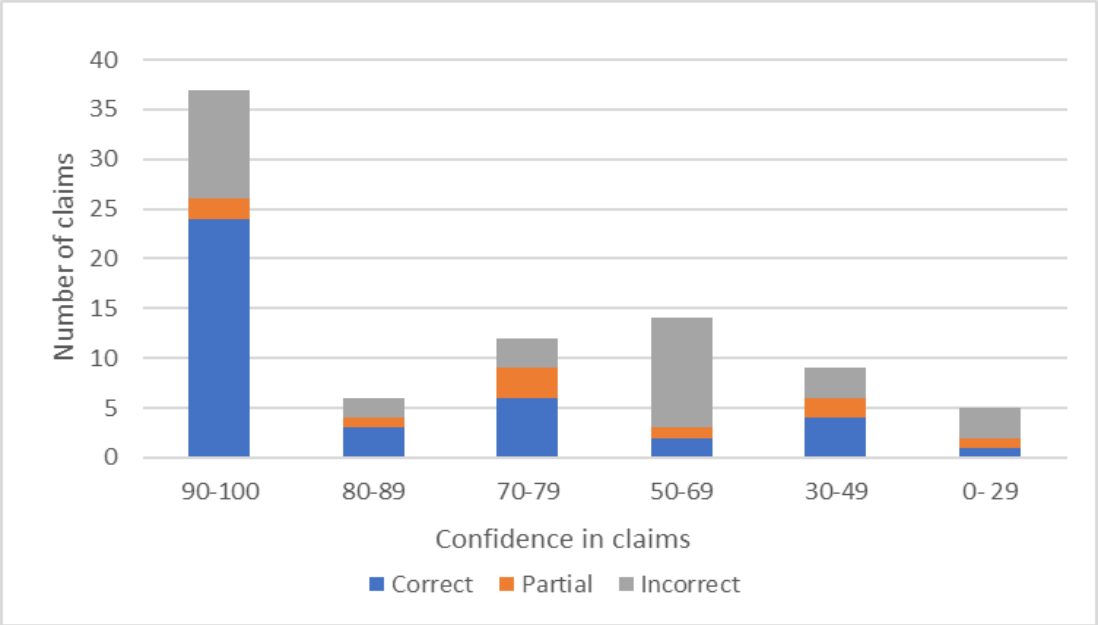
33/81 or 41% of identification claims were incorrect, the record marked in the cell did not relate to the individual named.

No attribute claims were made, an attribute claim is where an intruder claims to have found something new about a person through the data presented.

Of the initial 51, 12 dropped out, citing workload or holiday as reasons, and a further 9 filed no notes and made no claims. Of the 30 intruders that took part, 6 (20%) did not make any claims. Reasons cited included not being able to claim anything with certainty, some may also have lacked time to spend on the project.

2.32 Confidence

Figure1: Confidence, correctness and number of claims



This histogram shows numbers of claims by the percentage confidence the intruder reported in the claim, banded by whether they were correct, partially correct or incorrect.

There was a range of 7.5-100% confidence in claims

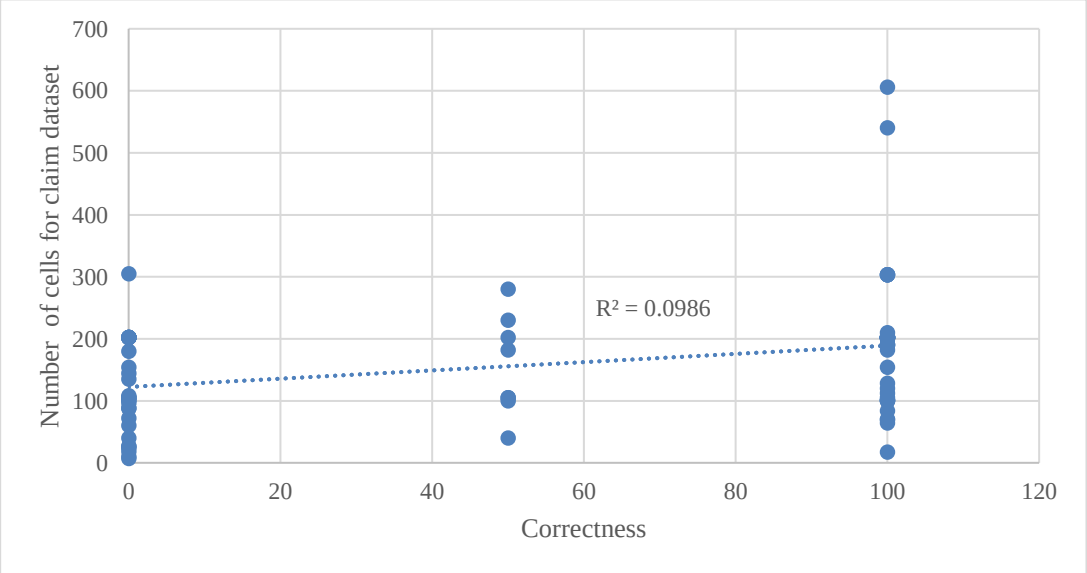
The mean confidence placed in a claim was 73.6%, the median was 80%.

2.32 Cell Counts and correctness

The cell count is the number of cells (row * columns) present in the table used to make the claim.

A wide range of table sizes were used to inform claims, (range 7 – 2100, mean 183, median 182).

Figure 2: Cell counts and correctness



The scatter plot shows claims rated by percentage correctness. Partially correct claims are 50% correct, fully correct are 100% correct. One outlier (cell count 2100) was removed. This shows a

positive correlation (R^2 = 0.0986), but with outlier, this relationship was zero. This could suggest that higher cell counts may increase possibility of identification within limits – very high cell counts may not.

2.34 Variables Used

To assess which variables were most likely to result in a claim, and which in a correct claim, the claims were coded to variable type. Any table constructed with a single classification making the bulk of the cells would be coded to that variable, e.g., any claim using single year of age, or single year of age plus another less detailed classification such as sex, was coded to 'age', any claim using a 3-part country of birth classification, 10-part age, and sex would be coded 'multivariate'. A few variables with only a few claims each were coded to 'other', such as country_of_birth.

Table 1: Number of claims by variables used in the datasets those claims came from

Variable	Number of Claims	Number of Correct Claims	% Claims that were correct
Age	35	21	60%
Multivariate	28	12	43%
Occupation	9	2	22%
Other	8	5	63%

The table shows claims where age was the main component had the highest number of claims, and highest number of correct claims. Multivariate tables were less than 50% likely to yield a correct claim, and occupation was unlikely to result in a correct claim. The main cause of correct claims from the 'Other' category were claims using country_of_birth.

2.35 Geography

Table 2: Number and correctness of claims by Geography used in datasets

Geography of the table used for the claim	Number of Claims	Number of Correct Claims	% of Claims made that were correct	Mean % confidence	Mean cell count
OA	67	34	51%	75	142
LSOA	9	5	56%	73	248
MSOA	5	1	20%	47	610

The largest geography used for any claim was Middle Super Output Area (MSOA), Output area (OA) was the main area of risk with the bulk of claims (65/81 or 80%) being made using OA datasets. It was also the focus of correct claims (34/40 or 85%). There were few claims at MSOA, and only one

correct claim. Lower Super Output Area (LSOA) again was used in few claims, and though these were majority correct, with such a small sample it cannot be concluded that this would always be more likely to be correct or not.

Table 3: Subject of the Disclosure Claim

	Number of Claims	Correct claims	Percentage of Claims correct
Family and friends	59	25	42%
People from news/ web	16	11	68%
Self-identification	6	4	67%

Many those known about through news or online articles were centenarians, identified through age and location.

Other

Though intruders were given access to 'fixed' tables as csv files, at least 7 intruders used them, there were no correct claims from these.

Qualitative evidence suggested the intruders found the new flexible outputs system was very easy to use (rated 4.3 out of 5 by the 15 intruders surveyed), and low amounts of time were recorded as typical to arrive at a claim (5-30 minutes) though it is hard to calculate total time taken per claim accurately as time spent logged in could not be taken as an indication of time spent on this project. Intruder feedback suggested that the disclosure rules built into the system were working as intended and when they tried to obtain a cell value of 1 at lower geography, the rules prevented this by denying the data.

3 Discussion

The overall results show that over half of identification claims were incorrect. However, unlike other intruder testing exercises carried out previously by ONS, intruders were fairly unlikely to make claims where they had low confidence. Almost all claims were made with a confidence of 60% or greater.

Generally, the higher percentage of confidence the intruder rated a claim, the more likely they were to be correct. Although this was statistically significant, the relationship was not so strong, and a significant portion of those who were over 90% confident were still incorrect or partially correct (35% or 13/37).

The exercise on 2011 census data saw a drop off in percentage correctness at very high confidence claims which was not seen here. Possibly, the ease of using the system may have made all intruders more confident, and meant intruders went for easier identifications, rather than putting forwards ones they were less sure of.

The method used for this exercise did not allow us to know whether an identification was wrong due to swapping, or other reasons – only if it was perturbed and therefore a 'partial'. Therefore, it is hard to evaluate the success of swapping as a single method from this evidence.

Cell counts of tables present an unclear picture, as no correlation was found with table size in cell count and correctness. Smaller tables may be easier to be sure where a person might be represented, where a larger table makes it more likely to get a small count to base an identification claim. It seems more detailed classifications may offer additional risk in some circumstances, but dependent on geography.

There were no claims at any geography higher than MSOA. It is likely that an intruder would have far more confidence over a claim at lower geographies since they may have considerable knowledge as to who lived in an OA with which they are familiar, but far more uncertainty as the geography level increases. Observing a cell count of 1 in an OA may convince them that the person they know is the only one with that combination of attributes. They might have less certainty at MSOA that the 1 corresponds to the subject of the claim given the lower likelihood of familiarity with the individuals in the population, as well as 'noise' introduced by error, imputation, record swapping and the cell key method.

The high level of claims and correct claims at OA make this the main area of risk to address in planned outputs. Claims made at OA also had the highest level of confidence with an average of 75% confidence expressed in the claims. The variables used for these claims were consistent with the general picture, that is, age was a main variable used for identifications, followed by other detailed classifications such as occupation and country_of_birth. Multivariate tables made the basis for 22 of the OA claims, of which most were incorrect or partially correct (13/22 or 59%), which demonstrates that the protections did well at protecting multivariate data as they were designed to do.

Whilst most of the claims were correct at LSOA (5/9 or 56%) this was a small sample and could equally have been majority incorrect with one fewer correct claim. However, some of the claims made at OA could equally have been made at LSOA, as they are small enough to make small counts prevalent, and intruders might have a moderate level of familiarity with most residents within a typical sized LSOA (1600 people). The level of confidence in LSOA claims was not much less than that shown in claims made from OA level tables (73% confidence in LSOA, 75% in OA claims). A majority of LSOA claims (5/9 or 56%) were based in multivariate tables, though a minority of these were correct (2/5 or 40%). The mean cell count of tables used for claims at LSOA was consequently much higher.

There was little risk of a correct claim (only 1/6 or 17%) from an MSOA table, so this supported earlier evaluations of the data that looked only at the sparsity of the likely tables, and restricted fixed-table outputs of detailed univariates to MSOA geography. The cell counts used for MSOA tables were higher on average, which is unsurprising given the higher population (typically 7000) that would have to be divided in the classifications to obtain a cell count of '1' to base an identification upon. The level of confidence was also significantly lower at average 47%.

That age was shown as a specific risk should be noted; however, some of these claims were claims made using already publicly available information on Centenarians so arguably the disclosure came from these sources, not the output. That said, many claims were also identifying people who happened to be the only one of that age in their area, so single year of age at Output Area geography has been shown as a specific risk to mitigate.

The variables used for correct claims supports current thinking that more 'definite' variables are more disclosive, that is age and country_of_birth are both variables that are likely to be reported consistently by the person filling in the Census.

Claims based upon occupation were very unlikely to be correct on the other hand, which may be due to uncertainty about how the question may have been interpreted by the person answering, and how their answer would have been coded by the automated processing system.

Multivariate claims are also less likely to be correct, possibly because increasing the number of variables increases the chances an answer would not have been given or been recorded the way the intruder guessed. The level of risk in these detailed univariates was still limited to smaller sized geography, so there is no evidence from this test to restrict the use of these variables at MSOA or higher geography.

In terms of the variables that relate to special category data there was no evidence that variables such as health, disability, ethnicity, religion, sexual_orientation and gender_identity, all of which were included in the test, were at significant risk of correct identification claims. This may be due to the protections put in place for these, and the less definite nature of these variables. Though we know 7 intruders tried to use the sexual_orientation and gender_identity datasets, these were made available separately through .csv files which may have made them harder to access. In the final outputs they would not be available below MSOA, so this intruder testing exercise seems to support that decision in terms of sufficient protection for that data.

The test was conducted pragmatically, and therefore recruited people with more statistical awareness and knowledge of the data than would be found in the general population, as they were ONS employees. This may be taken a slightly over-stringent test, as it may over-estimate the risks from intruder attempts made by the public.

4. Conclusion

The standard to be met to fulfil legal requirements is that claims should not be both made with confidence and correctness. The level of risk found in the current planned outputs found by this exercise would meet these legal definitions of safety, and additional steps were taken to decrease this risk further.

In response to the findings, the rules in the table builder were altered to restrict the availability of detailed classifications at lower geography, and one more detailed topic summary was replaced with a classification with fewer categories that consequently posed less risk. The majority of claims made here would not be possible to make using the actual output system.

Perturbation, swapping, the disclosure rules and general level of doubt in the data together were shown to be effective at preventing correct identifications.

Awareness of perturbation and swapping did not appear to result in lower levels of intruder's confidence in making claims, so this alone cannot be relied upon to meet the legal standards. Further steps were also taken to ensure LSOA level data was protected by restriction of the level of detail available at this geography.

The evidence seen here, with lower risk at MSOA, supports the decision to limit the geography of usual residents in communal establishments and households to MSOA, even though those datasets were not included in the test

The CACD system has been launched since this test took place, and sees some 900,000 interactions per month (ONS data), demonstrating the usefulness of Census data delivered in a flexible and immediate format. If this system is to be employed for a wider range of statistical products, further intruder testing should be considered as a means of measuring and mitigating disclosure risk in those datasets.

Intruder testing is a highly useful exercise for data providers to employ, where the level of risk presented by a dataset is in doubt. It gives evidence on the likely level of risk, where that risk lies, and can inform appropriate action to mitigate those risks.