Stats NZ
Tatauranga Aotearoa

Microdata output guide

Fifth edition

New Zealand Government

Citation
Stats NZ (2020). *Microdata output guide (fifth edition)*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-003201-1

Published in August 2020 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

Contact
Stats NZ Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600

www.stats.govt.nz

Contents

1 Purpose of this guide

Microdata output guide describes the methods and rules that researchers must use to confidentialise output produced from Stats NZ's microdata.

If you are a researcher using our microdata, following the rules in this guide ensures your output has a high chance of being released.

These rules do not cover every eventuality and should be applied as 'rules-of-thumb'.

If you produce an output that violates a confidentiality rule, or is not covered by any of the rules in this guide, you will need to explain why the output contains no disclosure risks. Please submit your explanation using the output submission form on the virtual machine desktop.

Note that it is against Data Lab security rules to include numbers, pictures, and output from within the Data Lab environment in any email correspondence, even with Stats NZ staff.

See chapters 5 and 6 for instructions on how to submit your output for release.

All final output and publications must include the appropriate disclaimer – see chapter 5 for more detail. Additionally, chapter 7 contains all disclaimers for final microdata output.

This guide contains output rules for all microdata datasets except the Census of Population and Dwellings and Longitudinal Census.

See 2013 Census confidentiality rules and how they are applied for output rules relating to the Census of Population and Dwellings and the Longitudinal Census.

If you have any questions about your output, email the microdata access team at access2microdata@stats.govt.nz or phone +64 4 931 4253.

What's new in the fifth edition

This section provides the key changes we made in the fifth edition of the *Microdata output guide*. Much of the content and layout have been revised as we aim to make the guide more user friendly.

- Stats NZ survey section and IDI/LBD section have been combined as many of their rules are similar.
- Deleted repetitions throughout the document.
- Added and updated examples, especially in the Appendix.
- Most rules have been modified slightly to make them easier to interpret for both researchers and the output checkers. The modifications include adding and updating the 2016 rules with new dataset applicable rules.
- 'Phase 1/Phase 2' are now both referred to as 'final output'.
- More guidance and steps have been added for the output checking process.

2 Principles of confidentiality for microdata access

Researchers who access our microdata must comply with output rules designed to maintain the confidentiality of information entrusted to us.

This chapter covers:

- Microdata access for researchers
- Why confidentiality is important
- Confidentiality, privacy, and security
- Goals for confidentiality
- Legislative requirements for protecting information.

Microdata access for researchers

We provide approved researchers with access to anonymised unit record datasets, which are known as microdata. Microdata contains information about specific people, households, and businesses.

These datasets are rich sources of information, which allow you to perform advanced statistical analysis and answer complex questions. We treat microdata datasets with extreme care, and only allow access to the data under specific conditions that meet the requirements of section 37C of the Statistics Act 1975.

You can access data from social and business surveys collected by Stats NZ, as well as administrative data. These datasets contain information about people, households, businesses, education providers, and other entities.

Examples of datasets available to you include:

- Census of Population and Dwellings, and Longitudinal Census
- Integrated Data Infrastructure
- Household Economic Survey
- Survey of Working Life
- New Zealand General Social Survey
- Disability Survey
- Te Kupenga.

The Integrated Data Infrastructure (IDI) is one of the more commonly used microdata datasets. The IDI is a linked longitudinal dataset comprising a series of datasets from different source agencies. You can produce statistical outputs on the pathways, transitions, and outcomes of people.

See Integrated Data Infrastructure for more information.

Our microdata access service also includes confidentialised unit record files (CURFs).

See confidentialised unit record files for more information, as CURFS are not covered in this guide.

Why confidentiality is important

We collect a diverse range of information to produce official statistics. Much of the data we collect and use is about individuals, households, and businesses, which contain personal and commercially sensitive information.

We rely on people's trust and goodwill to continue supplying us with high-quality information, so we can produce the data that New Zealand needs to grow and prosper. Therefore, maintaining privacy, confidentiality, and data security is one of our core values as leader of the Official Statistics System.

Confidentiality, privacy, and security

We are committed to ensuring the privacy, security, and confidentiality of all our information. This includes the collection, use, storage, and distribution of the information we collect from and about individuals, households, businesses, and administrative sources.

The terms privacy, security, and confidentiality are often used interchangeably, but each term has a different meaning:

- **privacy** refers to the ability of a person to control the availability of information about themselves
- **security** refers to how the agency stores and controls access to the data it holds
- **confidentiality** refers to the protection of information from, and about, individuals and organisations, and ensuring that the information is not made available or disclosed to unauthorised individuals or entities.

To protect confidential information, we have policies and protocols to control statistical disclosure. This protection is applied when we process data and publishing outputs. It also extends to our microdata access service, and the output researchers produce from microdata.

Goals for confidentiality

We operate a risk management framework for the microdata access service, which protects against the disclosure of confidential information. The output rules and checking processes covered in this guide are part of the risk management framework.

A disclosure may occur when a person recognises or learns something they did not already know about an individual or organisation, through microdata or output produced from microdata. For a disclosure to occur, this information must enter the public domain. The Statistics Act 1975 states that this type of disclosure must be prevented.

Microdata output rules are designed to meet the following four goals:

- **utility** – we want the research output to be as rich, detailed, and unmodified as possible
- **safety** – we manage the risk of disclosure of particulars about data subjects, down to the level required by the Statistics Act 1975, our ethical obligations, and the preservation of trust
- **simplicity** – we want the rules to be as simple to apply and check as possible
- **consistency** – we aim to maximise consistency across output produced by different channels (microdata access, Stats NZ production), and across similar output from different source collections.

The first two goals (utility and safety) lead to rules that aim to release as much detail as possible, and to protect the entities that need to be protected.

We aim to maximise the potential utility of research outputs by ensuring that the unit record data provided through the Stats NZ microdata access service is as rich and detailed as possible, and by not adding confidentiality perturbation to this data.

Legislative requirements for protecting information

We are required by law to protect the information we collect. These requirements are outlined in the Statistics Act 1975 and the Privacy Act 1993.

Sections 37 and 37C of the Statistics Act 1975 govern confidentiality protection of personal information and access to microdata.

See the Statistics Act 1975, section 37: Security of information on www.legislation.govt.nz.

See the Statistics Act 1975, section 37C: Disclosure of individual schedules for bona fide research or statistical purposes on www.legislation.govt.nz.

3 Summary of microdata output rules

The tables below show the microdata output rules that apply to different types of output produced from Stats NZ surveys and the Integrated Data Infrastructure. The next two chapters describe the output rules you need to apply to each type of output.

Breaking down types of output

Type of statistic	Type of output	Output rule/s	
		Survey (4.0.1)	IDI/LBD (4.0.2)
Descriptive statistics	Unweighted counts	4.1.1–4.1.3	4.1
	Weighted counts	4.2.1–4.2.4	4.2
	Count magnitudes		4.12
	Totals and means (value magnitudes)	4.3.1	4.3.2–4.3.4
	Medians and other quantiles	4.4	4.4
	Percentages, proportions, and ratios	4.5.1–4.5.2	4.5.4
	Maximum/minimum values	4.6	4.6
	Aggregation	4.10	4.10
	Suppression	4.11	4.11
	Underlying entities (for example, businesses)		4.13
	Simulated output		4.15
Analytical output	Regression models	4.7	4.7
Output from specific datasets	Suppression under 6 and 3		4.11.4
	Census data		4.14
	Annual Enterprise Survey data		4.16
	Overseas Merchandise Trade data		4.17
	Agricultural production data		4.18
	IDI population explorer		4.19
Graphical output	Graphs	4.8	4.8
Programming code	Programming code and logs	4.9	4.9

4 Microdata output rules

4.0 Related datasets

4.0.1 Stats NZ surveys

Apply the all-survey data rules to output produced from social surveys, apart from those in the Integrated Data Infrastructure (IDI): **These rules do not apply to outputs produced from the IDI.**

The output rules apply to the following stand-alone surveys:

- New Zealand General Social Survey
- Survey of Working Life
- Survey of Dynamics and Motivation for Migration
- Childcare Survey
- Disability Survey
- Time Use Survey
- Te Kupenga.

4.0.2 Integrated Data Infrastructure (IDI) and Longitudinal Business Database (LBD)

The Integrated Data Infrastructure (IDI) and the Longitudinal Business Database (LBD) are linked longitudinal datasets composed of administrative datasets and surveys from Stats NZ and other agencies.

See Data in the IDI for the **current list** of datasets in the IDI which can be found on the Stats NZ website.

Datasets that are both stand alone and IDI linked:

- New Zealand General Social Survey
- Household Economic Survey
- Household Labour Force Survey.

The following historic surveys are included in the IDI:

- Survey of Family, Income and Employment
- Longitudinal Immigration Survey
- New Zealand Income Survey
- 2013 Census.

For any of the following datasets, please contact access2microdata@stats.govt.nz as project-specific rules may apply:

- Survey of Family, Income, and Employment
- Intellectual Property Office of New Zealand Patent data
- Annual Enterprise Survey
- Business Operations Survey.

Note: Stats NZ datasets are also available as 'stand-alone' datasets outside the IDI. You must apply the IDI output rules to output produced from these datasets.

4.1 Unweighted counts

Unweighted counts refer to the number of observations that possess certain characteristics before any weighting has been applied to the data. You can produce unweighted counts from full-coverage datasets and sample surveys. Unweighted counts from survey data are often requested to assess data quality and the reliability of results.

4.1.1 Show all empty cells as zero, unless suppression is required for that cell.

4.1.2 When producing the same count in the same cell, apply the rounding to the count in the same direction, even if it is the same cell but in a different set of output. This check must be done manually.

4.1.3 Randomly round all counts to base 3.

The following 4.1 rule only applies to IDI/LBD (4.0.2) data:

4.1.4 Suppression of value magnitudes and other output may lead to the calculation of unweighted counts. If this occurs, apply the rounding to the count in the different direction.

See Appendix: Random rounding to base three for more details and examples.

4.2 Weighted counts

Weighted counts refer to the number of observations that possess certain characteristics after weighting has been applied to the data. Stats NZ gives weights to survey respondents to represent the population they characterise, and to allow publication of population estimates.

See Appendix: Weighted counts for an example.

The usual procedure for weighted counts is as follows:

- Suppress below a specified threshold and round to a specified base.
- Suppress all zeros.
- Secondary suppression is not required.

See Appendix: Suppression for more details.

4.2.1 For output produced from the following datasets, suppress weighted counts that are below the described threshold, and conventionally round using the described rounding base.

Survey	Threshold	Base	Additional rules
Household Economic Survey (HES)*	3,000	1,000	4.2.4
Disability Survey	1,000	1,000	4.2.3
Time Use Survey	1,000	1,000	
New Zealand General Social Survey	1,000	1,000	
Survey of Working Life	1,000	100	
Survey of Dynamics and Motivation for Migration	1,000	100	

Childcare Survey	1,000	100	
Household Labour Force Survey	1,000	100	
New Zealand Income Survey	1,000	100	
Survey of Family, Income and Employment	1,000	100	
Te Kupenga	500	500	4.2.2
Longitudinal Immigration Survey	20	20	

4.2.2 For output produced from **Te Kupenga**, apply the following rules:

- Suppress estimates with a relative sampling error (RSE) of 100 percent or greater.
- Identify estimates with an RSE between 30 percent and less than 50 percent with one asterisk (*).
- Identify estimates with an RSE between 50 percent and less than 100 percent with two asterisks (**).

4.2.3 For output produced from the **Disability Survey**, apply the following rules:

- Estimates with very few contributors are deemed a risk to respondents' confidentiality.
- Estimates based on an estimated population of less than 1,000 are suppressed. This is indicated in tables by an 'S'.
- Estimates with high RSE are suspect in quality. Therefore, all estimates with an RSE of 50 percent or greater are suppressed.
- Estimates with an RSE of 30 percent to 49.9 percent should be viewed with caution (indicated in tables by an asterisk *), and an error of 50 percent or greater will be indicated by an 'S'.

4.2.4 For output produced from the **Household Economic Survey**, also apply the following quality rules (due to the small sample size and under-reporting). Apply these rules only if relative sample errors are provided:

***Survey: HES18-19 only**	**Threshold**	**Base**
HES Income Component	1,000	100
HES Expenditure Component	3,000	1,000
HES Net Worth Component	3,000	1,000

HES18-19 specific rounding differs to previous Household Economic Surveys as the sample size significantly increased.

- Estimates with a relative sample error in the range 21–50 percent are flagged with a warning that they are unreliable.
- Estimates with a relative sample error of over 50 percent are suppressed.
- Cross tabulation – only estimates with a relative sample error of 20 percent and under are cross-tabulated with another variable.
- Weighted counts with a corresponding unweighted count of less than 6 are suppressed.

The following 4.2 rule only applies to IDI/LBD (4.0.2) data:

4.2.5 For output produced with the LBD datasets, treat weighted firm counts by randomly rounding all counts to base 3 (as per rule 4.1).

4.3 Value magnitudes (cell totals and means)

Value magnitudes refer to measures (cell totals and means) from a numerical variable, which is usually a financial variable. For example, average income by age group and sex in Wellington for 2019, total hours spent attending a cinema by sex, and labour force status in Christchurch between June and August.

4.3.1 For 4.0.1 data, suppress cell totals and means if the unrounded count is less than 5.

The following 4.3 rules only apply to IDI/LBD (4.0.2) data:

For value magnitudes produced from the IDI or LBD, apply any applicable below rules. Which rule to apply depends on whether the tabular output contains information about businesses, or information about individuals or households.

If the tabular output contains information about **businesses**, then the businesses need protection to ensure a business's contribution to a value magnitude cannot be estimated with accuracy. To protect businesses, apply the p% rule.

4.3.2 For tabular output containing business information, apply the p% rule to value magnitudes. If a cell fails this test (ie is deemed to be sensitive), protect the cell by either aggregation or suppression. For means, apply the p% rule to cell totals and calculate means from rounded counts.

See Appendix: The p% rule for more details and examples.

If the tabular output contains information about **individuals or households**, the individuals or households need protection to ensure an individual or household's contribution to a value magnitude cannot be estimated with accuracy. To protect individuals and households, you must calculate value magnitudes from at least 20 observations. For example, average income of PhD graduates in Auckland for 2018, average number of hours worked by migrants in Otago between September and February.

4.3.3 For tabular output containing social information, suppress cell totals and means if the unrounded count is less than 20. Calculate means from rounded counts.

4.3.4 **For** output containing means of log transformed variables and means of growth rates, suppress means if the unrounded count is less than 10.

4.4 Medians, quantiles, and percentiles

4.4.1 Suppress medians if the unrounded cell count is less than 10. For other quantiles and percentiles, use the table below to find how many observations are needed for each quantile or percentile. Where a median, quantile, or percentile is equal to the minimum or maximum value, apply rule 4.6.

Quantile or percentile	Number of observations needed overall
0.01	500
0.05	100
0.10	50
0.25	20
0.50	10
0.75	20
0.90	50
0.95	100
0.99	500

.

4.5 Percentages, proportions, and ratios

4.5.1 For 4.0.1 data, you may calculate percentages, proportions, and ratios using the unrounded counts.

4.5.2 For 4.0.1 data, round percentages calculated from unweighted counts to 1 decimal place.

4.5.3 Suppress percentages, proportions, or ratios where either, or both, of the counts used to calculate the percentage, proportion, or ratio have been suppressed.

The following rules 4.5 only applies to IDI/LBD (4.0.2) data:

4.5.4 For 4.0.2 data, derive all percentages, proportions, and ratios (including odds ratios) from the rounded counts.

4.5.5 Please see rules regarding crude and standardised rates in the Appendix in age-standardisation of data. These rules are especially relevant to **health-related data**.

4.6 Maximum and minimum values

4.6.1 Suppress maximum and minimum values. Where a maximum or minimum value is not identifying, it may be considered for release.

4.6.2 If you produce a maximum or minimum value that you believe is not identifying, please provide an explanation in your submission form.

See Appendix: Maximum and minimum values for more details and examples.

4.7 Regression models

Regression output does not usually have confidentiality issues, except in the circumstances listed below. However, you must ensure that pieces of output are not equivalent to statistics subject to other confidentiality rules, particularly small counts or statistics based on small counts. Please see section 4.15 for more information regarding simulated output if this is relevant to your output.

4.7.1 Regression output may contain counts or lead to the calculation of counts. Suppress regression output if the underlying unweighted observation count is smaller than 5 (including 0).

4.7.2 Classification and regression tree models may produce the equivalent of detailed count tables. If this occurs, suppress unweighted counts smaller than 5 (including 0).

4.7.3 Regression outputs equivalent to other forms of output need to have the relevant rules applied. For example, coefficients produced by ordinary least squares regressions with binary (0/1) right-hand-side variables are equivalent to cell means and, therefore, need to comply with the means rule. In contrast, the inclusion of continuous independent variables in such models negates this requirement, as the coefficients on the binary variables are no longer raw means.

4.8 Graphs

There are four main types of graph:

- Type A: Graphs produced from aggregated data, or tables that have been confidentialised (for example, frequency histograms, bar charts of magnitudes).
- Type B: Graphs produced directly from the unit record data, but aggregated in the process by the software (for example, frequency histograms, kernel density plots).
- Type C: Graphs produced directly from the unit record data, and displaying unit record values (for example, scatterplots, residual plots).
- Type D: Graphs produced from the results of modelling or derivation that use the unit record data (for example, regression curves).

You can format graphs in the following ways:

- Static – the graph is simply a picture with no data attached.[1]
- Interactive – can be modified by the software that contains the data.

4.8.1 Release type A graphs – either static or interactive format.

4.8.2 Release type B graphs – static format, only if the graph provides a high level of uncertainty.[2]

4.8.3 Release type C graphs – static format after further processing, at the discretion of the output checker. For this type of graph to be released, you need to ensure that individuals cannot be recognised and that values can only be estimated with a high level of uncertainty.[2] Further processing can include but is not restricted to: cutting off the tails of a distribution, removing outliers, jittering the actual values, and removing or modifying axis values.

4.8.4 Release type D graphs – either static or interactive format, but only if the values shown in the graph cannot be used to find the original data values (ie where the modelling or derivation cannot be reversed to find the original data values for each individual).

[1] When graphs are released in this format, you need to ensure that the points on the graph cannot be recalculated in some way (eg by counting the pixels).

[2] The level of uncertainty is high if the level of uncertainty about the data values is equal to or larger than that in the confidentialised tables. For graph type B you do not need to provide the underlying data, but you do need to include a justification in your output submission form explaining why the graph has a high enough level of uncertainty to be released. For graph type C, you need to include the underlying data and a justification in your output submission form.

4.9 Programming code and logs

Programming code and logs are subject to the normal output checking processes.

4.9.1 Apply rules as you would to any other type of output to programming code and logs. Do not include unit record data or counts in comments; however, it is acceptable to state the general size of a count. For example, the count is 'too small' or 'sufficient for analysis'.

4.10 Aggregation

Aggregation is a method for protecting sensitive cells by collapsing categories.

4.10.1 If a table contains sensitive cells, a method you can use to protect these cells is to aggregate (collapse) those categories. If the produced tables are the same as other tables released by Stats NZ (eg information release tables, NZ.Stat tables) then the same aggregation must be applied. These published tables are available on the Stats NZ website.

See Appendix: Aggregation for more details and examples.

4.11 Suppression

Suppression (or primary suppression) is the removal of a cell's value when it has been deemed sensitive.

4.11.1 When suppressing output use consistent notation within your files for consistency. If using non alphabetic notation such as 'S' and 'C' please clearly note your suppression notation.

4.11.2 When sensitive cells still occur and no further grouping is appropriate, suppress the cell (remove its value), then suppress other cells to stop the first cell from being determined. This later stage is called secondary suppression.

4.11.3 Secondary suppression is the suppression of other cells or marginal totals in the table so that the suppressed cell cannot be recalculated. There are no universal guidelines for applying secondary suppression, except there must be enough secondary suppression to ensure primary suppressed values cannot be derived. If the produced tables are the same as other tables released by Stats NZ (for example, information release tables, NZ.Stat tables), then the same secondary suppression must be applied. These published tables are available from the Stats NZ website.

See Appendix: Suppression for more details and examples.

The following 4.11 rules only apply to IDI/LBD (4.0.2) data:

4.11.4 Suppression under 6 and 3 for 4.0.2 data

Suppress counts of fewer than 6 **before** applying random rounding to base three. Raw counts of 4 and 5 must not be rounded to 6.

4.11.4.1 For output produced from the following datasets, suppress output if the **underlying unrounded count is fewer than 6** (including 0):

- Auckland City Mission

- Births, Deaths, Marriages, and Civil Unions (Department of Internal Affairs (DIA))
- Census (see rule 4.14 for additional census rules)
- Child, Youth, and Family (Ministry of Social Development (MSD))
- Children's Action Plan (CAP)
- Department of Corrections
- Family Start
- Housing New Zealand (HNZ)
- Immigration data (Ministry of Business, Innovation and Employment (MBIE))
- Industry training education data (Ministry of Education), 2015 dataset only
- Inland Revenue (IRD)
- International travel and migration data (New Zealand Customs Service)
- Meshblock level (eg counts of individuals, families, and households. Any geographical unit lower than meshblock should be suppressed and/or aggregated to at least meshblock level) (Census).
- Ministry of Health (including Mental Health) (MOH)
- Ministry of Justice (MOJ)
- New Zealand Police (eg Recorded Crime Victims Statistics (RCVS) and Recorded Crime Offender Statistics (RCOS))
- NZ Rugby
- Student loans and allowances data from StudyLink (Ministry of Social Development (MSD))
- Youth Services.

4.11.4.2 For output produced from the following datasets, suppress output if the **underlying unrounded count is fewer than 3** (including 0):

- Accident Compensation Corporation (ACC)
- Transport data (New Zealand Transport Agency (NZTA)).

ALL following rules (4.12 – 4.19) only apply to IDI/LBD (4.0.2) data:

4.12 Count magnitudes

Count magnitudes are cell totals of the contributed values of the businesses in the cells. The contributed values come from a numerical variable which is a count of individuals. For example, number of employees in the retail industry in Auckland for 2019, number of sheep in Wellington in 2020.

4.12.1 Apply graduated random rounding to all count magnitudes.

See Appendix: Graduated random rounding for more details and examples

4.13 Output relating to business, education, and other underlying entities

4.13.1 For output containing business information, suppress data if the underlying count of entities is fewer than 3. This rule applies to businesses. Use both the permanent business number (PBN) and the enterprise (ENT) variable as employer variables.

4.13.2 For output containing education information, suppress data if the underlying count of entities is fewer than 2. This rule applies to education providers and industry training organisations.

4.13.3 For output containing mental health information, suppress data if the underlying count of entities is fewer than 2.

4.13.4 For output containing youth services information, suppress data if the underlying count of entities is fewer than 2.

4.13.5 For output containing children's action plan (CAP) information, suppress data if the underlying count of entities is fewer than 2.

See Appendix: Underlying entities for more details.

4.14 Census data

4.14.1 All measures have simple conventional rounding applied. Different variables require different level of rounding:

- Measures from annual income are rounded to the nearest $100
- Measures from weekly rent paid are rounded to the nearest $10
- Measures for age are rounded to one decimal place
- Measures from whole number count variables are rounded to one decimal place.

4.15 Simulated output*

4.15.1 Suppress simulated output if the underlying count is smaller than 6 or has been suppressed because of any other confidentiality rule.

4.15.2 Provide formulae and information of how you created your simulated output.

4.15.3 Clearly identify and caveat all simulated output so it is not mistaken for real data. This is for before and after output is released from Stats NZ.

*Simulated output includes synthetic datasets created with datasets from within the Data Lab environment.

4.16 Annual Enterprise Survey (AES) data

If an output is equivalent to any published AES output please contact access2microdata@stats.govt.nz to ensure that the AES specific confidentiality thresholds are applied.

4.16.1 When producing industry-level output using AES data alone, ensure that the aggregation and secondary suppression rules have been followed. In particular, if the produced tables are the same as other tables released by Stats NZ then the same aggregation or secondary suppression must be applied.

4.16.2 Suppress output for the following ANZSIC96 industries:

- G5110 – Supermarkets and grocery stores
- G511010 – Supermarkets

Outputs can be released for one (but not both) of the following ANZSIC96 and ANZSIC06 industries subject to the general IDI rules:

- G51 – Food retailing
- G511020 – Groceries and dairies

Suppress output for the following ANZSIC06 industries:

- G411 – Supermarkets and grocery stores
- G411000 – Supermarkets and grocery stores

Outputs can be released for one (but not both) of the following ANZSIC06 industries subject to the general IDI rules:

- G41 – Food retailing
- G412 – Specialised food retailing

This requirement is due to a long-standing confidentiality agreement between Stats NZ and organisations within the industry.

4.17 Overseas merchandise trade data

The overseas merchandise trade dataset contains confidential items. For these items, an exporter or importer has requested suppression and Stats NZ has accepted their request. Outputs produced from the IDI that contain overseas merchandise trade data must not lead to the disclosure of these confidential items.

Generally, aggregated totals and counts based on Harmonised System (HS) groupings that include confidential items will not be released, due to the need to protect these items. Outputs of this nature will only be considered for release on a case-by-case, discretionary basis. To be considered for release, you must clearly demonstrate that the output does not risk disclosure of the confidential item(s).

See Trade confidentiality for the list of confidential export and import items.

4.18 Agricultural production data

Additional restrictions relating to the release of 2017 agricultural production statistics

Agriculture production statistics for 2017 are confidentialised using the noised counts and magnitudes (NCM) method. Stats NZ is currently consulting with data suppliers on wider perceptions associated with making this data available publicly at a low level (for example, at meshblock level).

This consultation is still taking place, so outputs produced at geography level lower than territorial authority area will be considered for release on a case by case, discretionary basis.

Note that while this consultation phase is occurring, requests to output statistics at an area unit level or lower are unlikely to be approved without suppression of statistics associated with low farm counts.

4.19 IDI population explorer data

Information from Datamart can be used for your published research, however, you still need to go through the normal output checking process.

4.19.1 For output produced from Datamart, apply rules as you would to any other type of output, except section 4.13 (business, education, and other underlying entities).

5 Guidance for sharing microdata output

This chapter provides guidance on situations when it is acceptable, or not, to share your microdata output.

Final output

Any output that includes data produced from microdata access must be reviewed by Stats NZ before it is released. The review will ensure all confidentiality measures have been applied.

Final output is confidentialised output that has been checked and released from the Data Lab by Stats NZ staff. You can publicly share final output but you must include the appropriate Stats NZ disclaimer. Please include the appropriate disclaimer(s) when submitting output for checking. Disclaimers need to be included in each microdata output submission as well as in final publications (see final publications section below). Refer to section 7 for disclaimers for final microdata output. There are different disclaimers provided for different types of data used, and varied disclaimers for publishing output in different formats (that is, ministerial briefings and journal articles).

If you change the format (‘repackage’) of your released output then you do not need to resubmit to Stats NZ for checking unless you have added new microdata to your output. For example, you have submitted a table of data, which has been checked and released by Stats NZ, then you prepare a presentation (for example, PowerPoint) using this data (and no new information). In this situation, the presentation material does not need to be checked by Stats NZ.

However, please be cautious about adding additional context and data that may reveal other information. If you are unsure about your released output after adding additional information, please email access2microdata@stats.govt.nz.

Any material produced using the 'phase 1' output and was checked before January 2018 must be submitted to Stats NZ for checking.

Final publications

You are required to submit your publications to Stats NZ for all projects and research outputs using Stats NZ microdata. You will also need to provide a record of what has been made publicly available. Ensure that you include the appropriate Stats NZ disclaimer/s to your output for any publications (see chapter 7). These include briefings, ministerial papers, journal articles, conference posters, or presentations.

Stats NZ microdata includes the Integrated Data Infrastructure (IDI) and Longitudinal Business Database (LBD), other datasets available in the Data Lab, and Confidentialised Unit Record Files (CURFs).

Submitting your research publications: If you have published research using Stats NZ microdata, please email a link to access2microdata@stats.govt.nz so we can add it to the database. This includes anything you’ve produced, such as journal articles, reports, presentations, and websites.

Letting us know of your published research ensures we are able to keep an accurate record of what microdata has been used for. To find current records of research using Stats NZ microdata, see Research using Stats NZ microdata.

6 Submitting microdata output for release

Requirements for checking microdata output for release

Our preferred file types are Microsoft Excel or comma separated value (CSV) files. Microsoft Excel compatible formats are ideal and are usually quickest to process. Other types of output such as programming code and MS Word are also preferred. Submit your output in tabular format wherever possible.

Difficult file types are those that cannot easily be opened or edited. In some cases, the checker may require you to reformat your output.

Supporting evidence

You must include supporting evidence when requesting output for release. Supporting evidence may include:

- a raw copy of your output (before confidentiality was applied)
- underlying entity counts
- definitions for all new variables
- code for analysis.

Raw output and entity counts should be presented in tabular format at the individual cell level.

Please provide your raw output and supporting evidence in a separate document to your confidentialised output.

Five steps for the submission and checking process

Follow these steps to get your microdata output checked and released by Stats NZ.

Step 1

Carefully apply the appropriate confidentiality rules to your output intended for release. This includes complying with output and dataset specific rules from data dictionaries.

Step 2

Place your output files in the assigned folder for checking, along with a copy of your raw output and any supporting evidence needed by the output checker (for example, underlying entity counts). Please provide your raw output and supporting evidence in a separate document to your confidentialised one unless previously arranged with the output checking team. Do not make additional subfolders. For large checks it is helpful to provide the code used to create the output (for example, SAS, R, SQL, and Stata).

Step 3

Select the output submission application from the virtual machine desktop and complete the necessary details. Please provide detailed explanations and include as much supporting evidence as possible so we can complete the checking process quickly and efficiently.

Note that large quantities of output, and/or more complex output, may require more than five working days to check. The checker will inform you if this is the case. You may wish to indicate which output files are highest priority, so the checker can release the more urgent files first.

If you need the output checked urgently, let us know on the submission form and a checker will review your output as quickly as possible.

Step 4

The output checker reviews your final output, which normally takes up to five working days to complete.

The output checker may need to email or phone you to clarify the output rules that you used, if this is not clear on the output submission form.

If the output needs modification before we can release it, the output will be withheld, and you will be notified. You will then need to modify the output and contact the output checker to inform them that you have made changes. You will only need to submit a completely new request with your modified output if you are asked to by a Stats NZ staff.

Step 5

If there are no confidentiality issues, the checker will release the output to you using the email address you provided.

If there are confidentiality issues in the output submitted that are concerning, the individual request will be declined at the checker's discretion.

7 Disclaimers for final microdata output

Final microdata output must include the appropriate disclaimer, depending on whether the output is produced from:

- Stats NZ surveys
- Integrated Data Infrastructure and/ or Longitudinal Business Database

If you have any questions about these disclaimers, email the microdata access team at access2microdata@stats.govt.nz

Disclaimer for output produced from Stats NZ surveys

All final output produced from Stats NZ surveys must include an acknowledgement and disclaimer.

Acknowledgement – stating that Stats NZ is the source for any tables, graphs, or data (supplied by Stats NZ) that are quoted in the paper or presentation.

Disclaimer – stating that the researcher takes full responsibility for the paper, that Stats NZ will not be held accountable for any error or inaccurate findings within the paper or presentation, and acknowledgement that access to data is in accordance with the Statistics Act 1975.

For outputs produced from **Stats NZ surveys or Census data**, use the following wording:

> Access to the data used in this study was provided by Stats NZ under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975. The results presented in this study are the work of the author, not Stats NZ or individual data suppliers.

Disclaimer for output produced from the IDI and/or LBD

Final output produced from the Integrated Data Infrastructure (IDI) must include the following disclaimer. The first four paragraphs must always be used:

> These results are not official statistics. They have been created for research purposes from the [Integrated Data Infrastructure (IDI) and/or Longitudinal Business Database (LBD)] which [is/are] carefully managed by Stats NZ. For more information about the [IDI and/or LBD] please visit https://www.stats.govt.nz/integrated-data/.

The additional paragraphs on the next page can be ignored if they do not apply to the data you used. These paragraphs are for:

- Inland Revenue (IR) tax data
- Publishing on the Stats NZ website.

Disclaimer for Inland Revenue tax data

You must also include the following paragraphs in your disclaimer if your final output uses Inland Revenue tax data:

> The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Disclaimer for publishing on the Stats NZ website

You must also add the following paragraphs if your final output will be published on the Stats NZ website (aligned with standard Stats NZ disclaimers):

> **Copyright:** This work is licensed under the Creative Commons Attribution 4.0 International licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to [*insert name of research owner*] and abide by the other licence terms.
>
> **Liability:** While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, [Stats NZ, *insert name of research owner*] gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

8 Glossary

aggregation: method for protecting sensitive cells is to collapse (aggregate) categories.

anonymised data: data with direct identifiers (for example, name, address, ID number, phone number) removed.

attribute disclosure: occurs when confidential information is revealed about an individual or organisation.

confidential data: data to be protected from disclosure.

confidentialised data: data modified or with suppressions in order to protect individuals' and organisations' information.

confidentiality: protection of individuals' and organisations' information, and ensuring the information is not made available or disclosed to unauthorised individuals or entities.

conventional rounding to base x: method that rounds numbers to the nearest multiple of x.

count magnitudes: cell totals of the contributed values of the businesses in the cells; the contributed values are counts, for example, number of employees in the food service industry.

Data Lab: secure facility on Stats NZ premises; a place where external researchers can be permitted access to microdata under contractual agreements that cover the maintenance of confidentiality, and that placed strict controls on the uses of the data.

data utility: property of data products that enables them to meet the information needs of users.

disclosure: inappropriate attribution of information to an individual or organisation.

family: two or more people living in the same household who are either a couple, with or without children, or one parent and their children; a child in a family can be of any age.

graduated random rounding (GRR): random rounding where the size of the added uncertainty (that is, the rounding base) increases with the value being rounded; this method can be used for count magnitudes.

household: either one person, or one or more families, or a family plus other people, or a group of people living together who are not a family.

identity disclosure: occurs when an individual or organisation is revealed as a respondent of a data collection.

microdata: unit-record level data, or data corresponding to information at the respondent level.

output checker: Stats NZ staff member who checks output for confidentiality issues.

p% rule: determines whether a cell is sensitive; a cell is considered sensitive if the value for any contributor can be calculated to within p percent.

perturbation: disclosure control methods that add uncertainty to data by changing some values.

privacy: ability of a person to control the availability of information about themselves.

random rounding to base x: method that randomly, and in an unbiased way, rounds values either up or down to the nearest multiple of x; for example, random rounding to base 3 (RR3) rounds values to multiples of 3.

raw data: data that has not been confidentialised, for example, unweighted counts without random rounding to base 3 (RR3).

secure microdata access environment: physical environment where researchers can be permitted access to microdata under contractual agreements that cover the maintenance of confidentiality, and that place strict controls on the uses of the data; this environment is specified in the Microdata Access Agreement signed between the research institution and Stats NZ.

security: refers to how the agency stores and controls access to the data it holds.

sensitive cell: cell for which knowledge of the value would permit an unduly accurate estimate of the contribution of an individual or organisation, or that reveals a small count.

suppression: removal of a cell's value when it has been deemed sensitive (also referred to as primary suppression); secondary suppression is the suppression of other non-sensitive cells or marginal totals in the table so that the (primary) suppressed cell cannot be recalculated.

threshold rule: rule that defines a cell as sensitive, based on the number of observations contributing to the cell.

unique: individual or organisation that can be distinguished from all other members in the sample (sample unique) or population (population unique), by using a set of identifying variables.

unrounded counts: also known as **raw counts** that have not been confidentialised, for example unweighted counts without random rounding to base 3 (RR3).

unweighted counts: refer to the number of observations that possess certain characteristics before any weighting has been applied to the data; unweighted counts can be produced from full-coverage datasets and sample surveys; those from survey data are often requested to assess data quality and the reliability of results.

value magnitudes: cell totals or means from a (non-count) numerical variable, which is usually a financial variable; for example, personal income, household expenditure, business revenue (or income), or hours worked.

weighted counts: refer to the number of observations that possess certain characteristics after weighting has been applied to the data; Stats NZ gives weights to survey respondents to represent the population they characterise, and to allow publication of population estimates.

9 References and further reading

References

International Statistical Institute. (2010, July 22 & 23). Declaration on Professional Ethics. Retrieved 2020, from https://www.scb.se/contentassets/db09cdb81aae41dd8a153bb366b00a36/isi-declaration-on-professional-ethics.pdf

Stats NZ. (1997, June). New Zealand Standard Institution Sector Classification 1996. Retrieved 2020, from Stats NZ: http://archive.stats.govt.nz/~/media/Statistics/surveys-and-methods/methods/class-stnd/institutional-sector/NZ%20Institutional%20Sector%201996%20Manual.pdf.

Stats NZ. (2007, May). Principles and Protocols for Producers of Tier 1 Statistics. Retrieved 2020, from Stats NZ: https://www.stats.govt.nz/assets/Uploads/Principles-and-protocols-for-producers-of-tier-1-stats/principles-and-protocols-for-producers-of-tier-1-stats.pdf

Stats NZ. (n.d.). Methodological standard for confidentiality in business collections. Stats NZ. Unpublished.

Stats NZ. (n.d.). Methodological standard for confidentiality in social collections. Stats NZ. Unpublished.

Further reading

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). *Statistical Confidentiality: Principles and Practice.* Springer Science+Business Media. doi:10.1007/978-1-4419-7802-8_1

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., . . . Wolf, P.-P. D. (2010, Jan). *Handbook on Statistical Disclosure Control .* Retrieved 2020, from ESSNet SDC: https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

Laerd Statistics. (2018). *Measures of Spread*. Retrieved 2020, from Laerd Statistics: https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php

Ministry of Health. (2014, December 16). *Mortality and Demographic Data 2011.* Retrieved 2020, from Ministry of Health: https://www.health.govt.nz/publication/mortality-and-demographic-data-2011

Office for National Statistics. (2018, November 12). *The Office for National Statistics guide to social and economic research.* (C. Deeley, Ed.) Retrieved 2020, from Welsh Baccalaureate: http://resource.download.wjec.co.uk.s3-eu-west-1.amazonaws.com/vtc/2018-19/bacRes/1048%20ONS%20Welsh%20Bacc%20Handbook-English_P.pdf

Stats NZ. (2006). *Confidentiality best practice manual (First edition).* Wellington: Unpublished.

Stats NZ. (2009a). *Methodological standard for confidentiality in census.* Wellington: Unpublished.

Stats NZ. (2009b). *Methodological standard for Confidentiality Standard for Microdata Access.* Stats NZ. Wellington: Unpublished.

Stats NZ. (2015, May 8). *New Zealand Period Life Tables: 2012–14.* Retrieved 2020, from Stats NZ: http://archive.stats.govt.nz/browse_for_stats/health/life_expectancy/NZLifeTables_HOTP12-14/Commentary.aspx

Stats NZ. (2016, January 6). *Period life tables.* Retrieved 2020, from Stats NZ: http://archive.stats.govt.nz/browse_for_stats/health/life_expectancy/period-life-tables.aspx

Stats NZ. (2019, October 3). *Serious injury outcome indicators: 2000–18.* Retrieved 2020, from Stats NZ: https://www.stats.govt.nz/information-releases/serious-injury-outcome-indicators-200018

Appendix: Output rules – extra details and examples

Refer to this Appendix for additional details and examples to supplement the microdata output rules in chapters 4.

Suppression

Suppression (or primary suppression) is the removal of a cell's value when it has been deemed sensitive. Secondary suppression is the suppression of other cells or marginal totals in the table so that the suppressed cell cannot be recalculated.

Justification:

If secondary suppression is not applied when appropriate, a user can recalculate the suppressed data.

Example:

Before confidentialising:

Total turnover in the retail industry ($million)				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	11	47	58	116
Fuel retailing	2	32	33	66
Other	1*	31	20	53
Total	14	110	111	235
Note: * This cell is deemed sensitive (by the p% rule) and needs protection.				

After suppression has been applied: **Total turnover in the retail industry ($million)**				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	11	47	58	116
Fuel retailing	S	S	33	66
Other	S*	S	20	53
Total	14	110	111	235

Symbols: S suppressed
Note: * This cell is primary suppressed. The other three suppressed cells are secondary suppressed.

Aggregation

A method for protecting sensitive cells, or cells containing small counts, is to aggregate (collapse) those categories.

Justification:

Tables may contain counts, magnitudes, or measures. Small counts need protection. Magnitudes may allow a contributor to be estimated with insufficient uncertainty. Aggregation is a way of avoiding these cells.

Example:

Before confidentialising: **Total turnover in the retail industry ($million)**

Industry	Invercargill	Queenstown	Dunedin	Christchurch
Food retailing	11	47	58	116
Fuel retailing	2	32	33	66
Other	1*	31	20	53
Total	14	110	111	235
Note: * This cell is deemed sensitive (by the p% rule) and needs protection.				

After aggregation: **Total turnover in the retail industry ($million)**

Industry	Invercargill	Queenstown	Dunedin	Christchurch
Food retailing	11	47	58	116
Other	3**	63	53	119
Total	14	110	111	235
Note: **This cell is now safe (passes the p% rule).				

Random rounding to base 3 (RR3)

Unweighted counts are randomly rounded to base 3. Marginal totals of these counts can be independently and randomly rounded to base 3. Alternatively, you can calculate marginal totals by summing the rounded counts, but this introduces avoidable noise.

Random rounding to base 3 (RR3) involves randomly changing each count in a table to a multiple of 3. Apply RR3 by rounding values to:

- the nearest multiple of 3 with a probability of 2/3
- the second nearest multiple of 3 with a probability of 1/3.
- Values that are already multiples of 3 are left unchanged.

Justification

Small counts are sensitive. This rule protects counts of 0, 1, and 2. It also protects small counts from being revealed when differencing occurs, as all counts (large and small) are rounded.

Macros

Macros that perform RR3 are available from Stats NZ.

Examples

An original (unrounded) count of 17 would be rounded to 15 with a probability of 1/3, and rounded to 18 with a probability of 2/3. Since 15 x 1/3 + 18 x 2/3 = 17, the expected value is unchanged and over a table, bias is avoided.

A researcher performs some analysis and produces two tables of counts, which are disaggregations (by different demographics) of the same population. Suppose one cell in each of the tables

represents the same count and this unrounded count is 11. This count could be rounded to 9 or 12. The researcher must ensure this count is rounded in the same direction for the two tables. So if the count is rounded to 9 in the first table, then it must also be 9 in the second table.

A researcher performs some analysis on Monday (for example) and produces an unrounded count of 8, which could be randomly rounded to 6 or 9. Suppose it is rounded to 9. If the researcher re-runs this analysis on Tuesday and produces the same unrounded count of 8, then this count must be rounded to 9.

The following tables are a made-up scenario, where we assume that no further confidentiality measures are needed.

Unrounded tables

Age	15–19	20–24	25–29	30–34	35–39	40–44	**Total**
Yes	35	23	65	85	63	26	297
No	34	75	25	13	36	47	230
Total	69	98	90	98	99	73	527
Age	15–19	20–24	25–29	30–34	35–39	40–44	**Total**
Yes	16	57	64	83	35	23	278
No	53	41	26	15	64	50	249
Total	69	98	90	98	99	73	527

Rounded table

Age	15–19	20–24	25–29	30–34	35–39	40–44	**Total**
Yes	33	21	66	84	63	27	297
No	33	75	24	15	36	48	231
Total	69	96	90	99	99	72	528
Age	15–19	20–24	25–29	30–34	35–39	40–44	**Total**
Yes	15	57	66	81	33	24	279
No	54	39	24	15	63	51	249
Total	69	99	90	96	99	75	525

Rounded table adjusted for consistency issues

Age	15–19	20–24	25–29	30–34	35–39	40–44	**Total**
Yes	33	21	66	84	63	27	297
No	33	75	24	15	36	48	231
Total	69	99	90	99	99	72	528
Age	15–19	20–24	25–29	30–34	35–39	40–44	**Total**
Yes	15	57	66	81	33	24	279
No	54	39	24	15	63	51	249
Total	69	99	90	99	99	72	528

The ‘Rounded table adjusted for consistency issues’ table has its final totals fixed to be rounded in the same direction. This fix needs to be checked manually.

Graduated random rounding (GRR)

Graduated random rounding (GRR) rounds the total number of individuals (for example, employees, cows, and sheep) within a cell by adding protection that depends on the size of the number. In this way, the noise added forms a proportion that increases with the size of the number. Apply the following GRR procedure to tables containing count magnitudes:

Count magnitude	Rounded to base
0–18	3
19	2
20–99	5
100–999	10
1,000+	100

Justification

GRR prevents the derivation of the exact value of a contributor, even if the number of contributors is small.

Macro

A macro that performs GRR is available from Stats NZ.

Example

Before confidentialising

Number of employees in the retail industry

Industry	Invercargill	Queenstown	Dunedin	Christchurch
Food retailing	384	992	1,226	3,156
Fuel retailing	77	24	71	98
Other	2	34	284	555

After rounding

Number of employees in the retail industry

Industry	Invercargill	Queenstown	Dunedin	Christchurch
Food retailing	390	990	1,200	3,100
Fuel retailing	75	20	75	100
Other	0	35	280	550

Weighted counts

The usual procedure for weighted counts is as follows:

- suppress below a specified threshold, which is usually three times the mean weight
- conventionally round all other weighted counts to a specified base, which is usually three times the mean weight

- suppress all zeros
- secondary suppression is not required.

Example

The following table contains unrounded weighted counts produced from the Household Labour Force Survey (HLFS):

Number of people in part-time employment in Wellington

Unrounded table

Age	Male	Female
15–19	7,707	5,408
20–24	13,310	15,601
25–29	24,548	25,123
30–34	32,353	34,021
35–39	21,134	11,346
40–44	5,603	3,017
45–49	2,450	874
50+	1,789	902

Rounded table with suppressions

Age	Male	Female
15–19	7,700	5,400
20–24	13,300	15,600
25–29	24,600	25,100
30–34	32,400	34,000
35–39	21,100	11,300
40–44	5,600	3,000
45–49	2,400	S
50+	1,800	S

The p% rule

A table of value magnitudes usually contains totals of the contributed values of the businesses in the cells of the table. The contributed values usually come from a financial variable, measured in dollars. Stats NZ's confidentiality standards state that any estimate for the contribution of a business to a cell total needs to have a sufficient level of uncertainty attached. The p% rule provides a measure of this level of uncertainty, and therefore the sensitivity of the cell.

The p% rule states that a cell, contained in a table of magnitudes, is sensitive if the value for any contributor can be calculated to within a given percentage. Cells that are identified as sensitive must be suppressed or otherwise avoided (through table design).

The p% rule calculates the distance (as a percentage) between the estimated value and the true value for the largest contributor in a table, as follows:

$$p = \frac{\hat{X} - X}{X} \times 100$$

X is the value for the largest contributor, $\hat{X} = Total - Y$ is the estimate of X, and Y is the value for the second largest contributor. If the value of p is less than the p% threshold then the cell is deemed sensitive. If a cell is sensitive, you must apply aggregation or suppression (with secondary suppression) to the table. If a contributor's value is negative, take the absolute value before calculating Total, X, and Y.

The value of the threshold is confidential and must not be made public, but it is built into the macro and visible to researchers.

Justification:

The justification for the p% rule follows directly from its description. If a cell is deemed not sensitive, then a business can estimate the value contributed by a competitor, but only with a sufficient level of uncertainty. The p value estimates this level of uncertainty.

Business	BP	Z	Caltex	Mobil	Total
Income ($millions)	50	100	150	200	500

Macro:

A macro that calculates the p value for each cell is available from Stats NZ.

Example:

A cell in a table contains five businesses with the following income values:

Assume that Caltex knows its own income and the total income for all businesses. If Caltex would like to estimate the income for Mobil, then the value of p is calculated as follows:

$$p = \left(\frac{(500 - 150) - 200}{200}\right) \times 100 = 75\%$$

Therefore, Caltex will overestimate the income for Mobil by 75 percent. If 75 percent was less than the p% threshold, then the cell total would be deemed sensitive and must be suppressed.

Age-standardisation of data

Special note for standardised rates which are particularly relevant when working with health-related data

Crude rates: When checking the crude rates for output using health-related data, apply the following steps:

- Crude rates should be calculated in the following way: Crude rate = rounded samples/rounded population*100000 (for rates per 100,000 population, could be per 1,000 or %)
- Check if the following information has been provided: population counts and sample counts
- Suppression under 6 has been applied to the raw counts. Apply RR3 to both samples and population counts
- If the sample count has been suppressed, then crude rate also needs to be suppressed

Standardised rates: Standardised rates (for example, age-standardised rates) are calculated from individual data with age group rates weighted using a reference population (often WHO population) and then summed. The final standardised rate is equivalent to a statistical analysis result so it should not be suppressed or rounded.

Maximum and minimum values

Maximum and minimum values are normally suppressed. Where a maximum or minimum value is not identifying, it may be considered for release.

Justification:

Maximum and minimum values are respondent values and may be outliers that pose a high risk of disclosure.

Examples:

Income, expenditure, and revenue are examples of sensitive variables. Maximum and minimum values for these variables are normally suppressed.

A researcher produces output about adults from a Stats NZ dataset, including maximum and minimum ages. The minimum age may be released as this is likely to be defined by Stats NZ. The maximum age will be suppressed as this may be an outlier.

A researcher creates derived scores from the Integrated Data Infrastructure (IDI) and produces maximum and minimum values for these scores. These maximum and minimum values may be considered for release on a case-by-case basis. The researcher will need to provide justification that the values are safe and are not identifying.

Underlying entities

Examples of entities include (and are not limited to) businesses, schools, and providers of mental health and children's action plan services.

Output based on a small number of underlying entities may breach the confidentiality requirements specified in the Statistics Act 1975 (regardless of whether the entities are named).

If your output contains any of the information listed in section 4.13 you must provide underlying entity counts for each cell. Use variables such as 'provider ID' and 'organisation ID' to count the number of distinct entities.

Rule 4.13.3 must be applied for outputs (related to Mental health) produced from the PRIMHD, Pharmaceutical, National Non-Admitted Patient Collect, and Hospital discharges datasets. Further information about entity variables for all datasets can be found in the applicable data dictionaries.

In general, Section 4.13 does not apply for outputs produced from datasets where underlying entities cannot be counted. If you are unsure whether you need to provide underlying entity counts in your output please email DatalabChecking-SharedMailbox@stats.govt.nz

Supporting information to the Microdata output guide (fifth edition)

Frequently asked questions – FAQs

What confidentiality rules should I apply when I am using multiple datasets?

When using more than one dataset with different confidentiality requirements, apply the most conservative confidentiality rules. For example, when using census data combined with Ministry of Education data, follow the rules in the IDI section of the Microdata Output Guide, in addition to the education and census specific rules (suppress counts below 6 and aggregate data to a minimum of two education entities (rules 4.11.4 and 4.13.2).

Note that unless otherwise specified, of the confidentiality rules listed in the microdata output guide apply to specific types of output, regardless of which dataset you are using. For example, if you calculate a percentage using any kind of IDI data, you must always follow rule 4.5.4 (derive the percentage from randomly rounded counts)

What supporting information should I provide with my output request?

Please provide any background and supporting information that may be useful for the output checkers. Examples include raw data, descriptions of variables, which confidentiality rules you have applied and details of previous discussions with Stats NZ.

What is raw data?

Raw data is data that has not been confidentialised and is not intended for release. It is used by the checkers to verify that the appropriate confidentiality has been applied. Examples include unweighted counts before random rounding was applied and underlying entity counts.

What are entities and how do should I count them?

Entities are organisations that are present in some datasets. Example of entities include (but are not limited to) businesses, schools, universities and providers of mental health services.

For example, to count the entities underlying the group of students below, add up the total number of distinct entities associated with the (fictional) SNZUID values.

SNZUID	Provider ID
123456	63744
234567	37926
345678	63744
456789	43657

Note that you will not be able to release each individual snzuid value, but you may release the unweighted count of students (4) after applying random rounding. In this case there are 3 distinct entities so rule 4.13.2 has been met.

Where should I store the files listed in my output request?

More information is provided about this by the Access2Microdata team once you are an approved researcher.

Place your files within the checking folder under your name. For example:

> MAA2019-99 Investigating the IDI Confidentiality Rules > Checking > JohnDoe

Once your files have been released they will be saved in your checked folder organised by date:

> MAA2019-99 Investigating the IDI Confidentiality Rules > Checked > JohnDoe > 31-12-19

Further rule breakdown

Further rule breakdown and examples will be coming for the next update, please contact datalabchecking-sharedmailbox@stats.govt.nz for assistance if you have any questions or recommendations.