Australian Bureau of Statistics

Statistical Spatial Framework

# SSF Guidance Material – Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing

When individuals or organisations provide information that is private[1], this information must be managed as sensitive information and confidentialisation methods must be used to protect that private information. The Statistical Spatial Framework (SSF)[2] identifies the importance of protecting private information that is geospatially enabled. These protections must be applied to private information stored in unit record datasets (where data is in its raw form) *and* to statistical information that is released from these datasets. Maintaining privacy is of critical importance whenever *any* data is released from a socio-economic dataset; however, when the data being released includes point-based locations (i.e. coordinates) or regional[3] breakdowns there are some specific issues that must be considered.

## Purpose

This paper describes where some common privacy risks occur when releasing statistical information from geospatially enabled socio-economic datasets. It will also provide some information on the options and techniques for confidentialising the statistical information released. The content of this paper is particularly relevant to organisations seeking to implement the Statistical Spatial Framework (SSF) for Australia, and is generally relevant to those organisations releasing information from geospatially enabled socio-economic datasets.

## Releasing information - privacy requirements

Releasing information from socio-economic datasets helps to realise the value of that dataset by:

- providing information for business processes and management,
- informing and engaging the public,
- promoting innovation and economic growth, and
- realising a commercial return.

[1] The ‘What is covered by privacy’ page on Office of the Australian Information Commissioner website provides more information about privacy.

[2] More information on the Statistical Spatial Framework (SSF) is available on National Statistical Service (NSS) website.

[3] For this paper, the term ‘region’ refers to a boundary set of any type (e.g. Suburbs, Local Government Areas, ASGS SA1s to SA4s, ASGS Remoteness Areas, etc.).

Release of information from a dataset must also comply with the provisions of the *Privacy Act 1988*, as well as any other Commonwealth or state and territory legislation relevant to the operation of your organisation or agency, or the information you are using. The Office of the Australian Information Commissioner provides a range of guidance material for government agencies and businesses on matters of privacy. A good starting point is the About privacy page on the Office of the Australian Information Commissioner website.

Release of information may also need to take into account ethical considerations, both professional and general. For some organisations and government agencies, release of information may also need to comply with organisational polices and/or government policies.

## De-identifying and confidentialising data

To protect the privacy of an individual's or organisation's information in socio-economic datasets, the data being released may need to be de-identified or confidentialised. This will prevent an individual or organisation from being identified within the data and the private information they have provided from being disclosed.

*De-identifying data* removes or modifies some or all of the identifying features, such as name, address, date of birth, and gender.

*Confidentialising data* seeks to minimise the risks of identification of an individual or organisation through the presence of very rare characteristics or the combination of unique or remarkable characteristics by applying some or all of the following methods:

- removal of directly identifying features or characteristics,
- careful design of the format of the released data, and/or
- modification to the content of the data released.

When data is released with a location or region component privacy risks can be heightened. The number of characteristics that are needed to uniquely identify an individual or organisation generally decreases as the size of the region diminishes. When data is released for point-based locations (i.e. with coordinates) identification is a relatively simple task.

The National Statistical Service (NSS)[4] has a Confidentiality Information Series that provides information on the obligations and identification risks relating to data releases, as well as methods that can be used to confidentialise data. The Confidentiality Information Series can be found on the NSS website. A more detailed paper on the ABS website – Research Paper: A Review of Confidentiality Protections for Statistical Tables, Jun 2005 (ABS Cat. No. 1352.0.55.072) – examines some of the confidentialisation techniques that can be used to protect privacy. The remainder of this paper focuses on the specific privacy risks for data with a location or region component.

[4] More information about the National Statistical Service is available on the NSS website.

## Point-based coordinate data and privacy

The Statistical Spatial Framework (SSF) recommends that unit records in socio-economic datasets be geocoded with location coordinates (i.e. latitude and longitude) and an Australian Statistical Geography Standard (ASGS) Mesh Block[5] code. This geocode information is usually obtained through geocoding the location address information for each statistical unit in the dataset. While de-identification of a dataset will generally remove the address information, the coordinate information and possibly the Mesh Block code could be combined with other information in the dataset to identify an individual or organisation. Due to the precision of coordinate information, it is generally recommended that it not be released in combination with other characteristic information contained in socio-economic datasets.

In specific instances, it may be necessary or desirable to release data at the unit record level. In these instances the data should be treated as 'microdata' and the specific guidance around the release of this type of data should be carefully considered. For further guidance see the Confidentiality Information Series Information Sheet 5: 'Managing the risk of disclosure in the release of microdata'.

The National Address Management Framework (NAMF)[6] includes provision for interchange of address and coordinate information between organisations where other information is not included. This is documented in the NAMF 'Address data interchange standard', which can be found on the NAMF webpage.

## Geographic region data and privacy

The Statistical Spatial Framework (SSF) recommends that regional data released or made available from any socio-economic dataset should include Australian Statistical Geography Standard (ASGS)[7] regions. ASGS regions are the common geography in the SSF. The SSF acknowledges that releases of data from these datasets may also include additional regional breakdowns that are not part of the ASGS (e.g. school catchments or Medicare local regions).

It is recommended that a careful assessment should be made before releasing information for ASGS Mesh Blocks and any other very small area or population regions. While the Mesh Blocks may be available through the address coding process, there are usually only a limited number of persons or organisations in each Mesh Block. These low numbers mean there is a high likelihood that individuals or organisations would be able to be identified through the use of only a limited number of characteristics, creating substantial risk of a breach of privacy through disclosure of private information. Similar concerns apply for other very small area or population regions. If data is released for Mesh Blocks, the information provided in the remainder of this paper is particularly relevant and should be carefully considered.

[5] More information on Mesh Blocks is available on the Australian Statistical Geography Standard (ASGS) webpage.

[6] More information is available on the National Address Management Framework (NAMF) webpage.

[7] More information is available on the Australian Statistical Geography Standard (ASGS) webpage.

It is generally recommended that data only be released where it is aggregated (i.e. summed or grouped together) for medium to large geographic regions. By aggregating data into regions the private information for an individual or organisation is combined with other private information; as more data is combined together it becomes increasingly difficult to identify the separate pieces of private information. The process of aggregating data by regions or by other classification groups (such as industry classification) is an important and fundamental confidentialisation tool that is used to protect privacy.

The region type selected for releasing data should have a large enough number of statistical units (e.g. persons or organisations) in each region to maintain confidentiality for the majority of the variables for which data is planned to be released. The size of the region must also be balanced against the regional analysis needs of the users of the data. Data confidentialisation techniques can then be used to manage the remaining instances where privacy risks exist.

Selecting a region type usually requires testing possible options to determine if any particular region type will contain large amounts of statistical information that requires further confidentialisation. The information provided below explains how to identify and manage these privacy risks.

Releasing data using geographic regions helps to manage privacy risks but does not completely resolve these issues. Geographic regions are only one set of variables in a dataset and the other characteristic variables can be used in combination with each other, or in combination with region, to identify an individual or an organisation. Therefore, each possible cell of data needs to be tested to assess the privacy risks.

## Aggregate data - identifying and managing privacy risks

A range of methods are available to identify and manage privacy risks when preparing data for release. These methods can be applied equally to data that is broken down by region or by other characteristic variables, or a combination of these.

The Confidentiality Information Series Information Sheet 4 - How to confidentialise data: the basic principles provides information on these methods. When applying these confidentialisation methods there are a few specific issues that need to be considered when the data to be released is broken down by region. These are discussed in the next section.

Other confidentialisation methods for managing privacy risks when releasing data in a geospatial format have been used by a range of organisations. These methods include:

- modelled characteristic information, which removes the risk of an individuals or organisations actual information being released, and
- data generalised to a grid system, including smoothing of data across a grid to remove extreme or unique values.

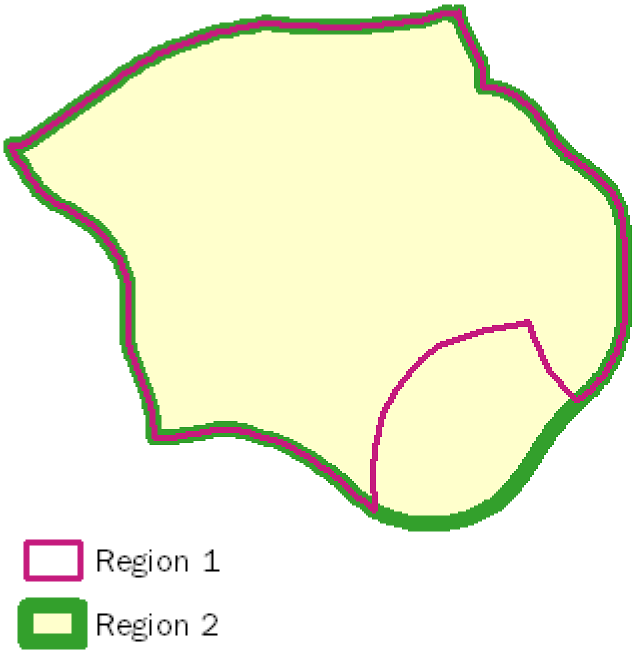
## Privacy risks for regionalised data

This section describes privacy risks that are specific to regionalised data released from socio-economic datasets.  It presumes readers are somewhat familiar with the privacy risks and confidentialisation methods outlined in the Confidentiality Information Series Information Sheet 4 - How to confidentialise data: the basic principles.

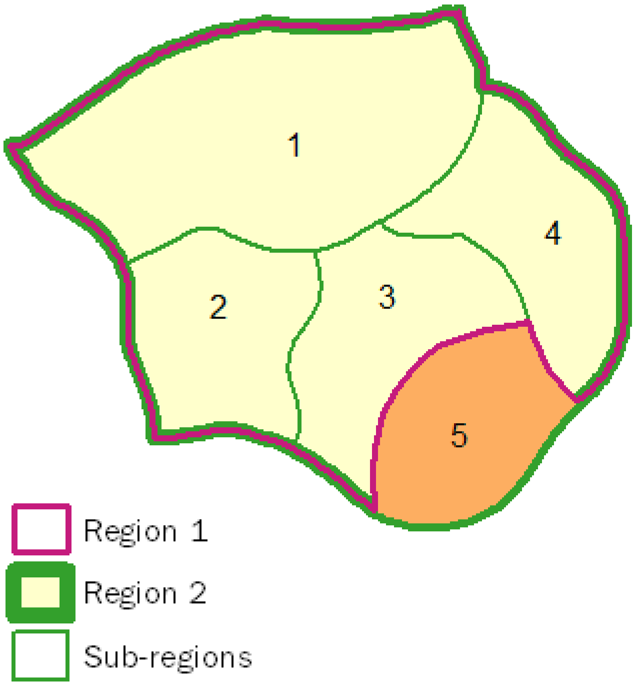
### Simple Geographic Differencing

Geographic differencing is the process where the same statistical data is obtained for two similarly shaped regions and the data from one region is subtracted from the other larger region – see Diagram A.  By using this method, it is possible to obtain data for the area that is not common to both regions.  The data obtained for this smaller area might result in a privacy breach through inadvertent disclosure of private information.

### Diagram A – Simple geographic differencing across two region types



### Diagram B – Geographic differencing across regions with sub-regions



A similar but more complicated example of geographic differencing is where data for a region is formed by adding together data from smaller sub-regions – see Diagram B. In this example, data for the region 1 in the previous example can be obtained by adding together the data for the 4 sub-regions. The data for the small area in sub-region 5 can then be differenced by subtracting the data for Region 1 from Region 2. The same privacy risks potentially arise from this example.

It can be seen from these examples that the release of data for different regions from a dataset may be used to obtain private information through geographic differencing. To ensure the confidentiality of data released from a dataset, each release of data on a region must be compared with other releases of data by region to identify each area of overlap where geographic differencing could potentially occur. Each of these areas of overlap must be assessed for privacy risks.

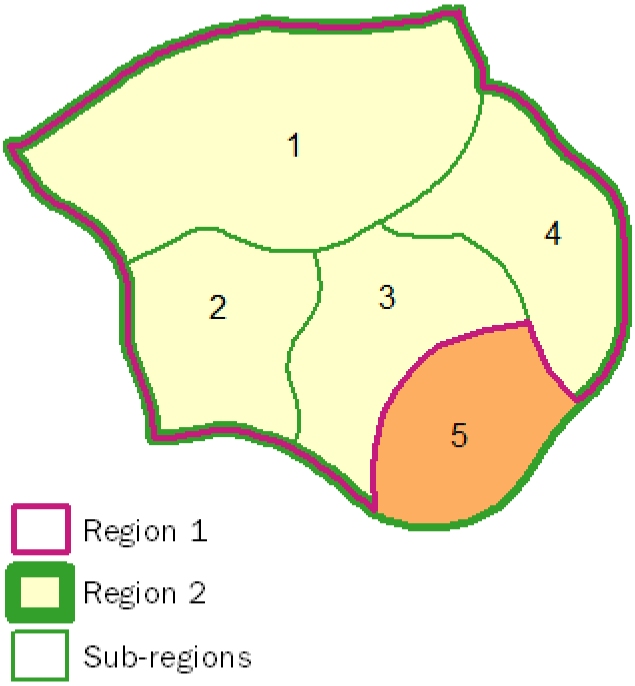
Methods that can be used to ensure the confidentiality of the data where overlap areas create privacy risks include:

- modifying the regions to be released to eliminate the areas of overlap,
- combining regions to eliminate the areas of overlap, and
- using techniques, such as category collapsing or suppression, which are described in the Confidentiality Information Series Information Sheet 4 How to confidentialise data: the basic principles (noting the issues associated with suppression outlined below).

## Suppression and Geographic Differencing

In an extension of the example in Diagram B, data is released for the sub-regions shown in Diagram C; however, the data for sub-region 5 is deliberately suppressed (not released) as it is assessed to contain data that is a privacy risk. In addition, the data for Region 2 is released; either as a higher level region, possibly in an earlier release, or as a total item with the sub-region data.

## Diagram C – Geographic differencing and suppression



The geographic differencing method from our previous example, demonstrates that the suppressed information for sub-region 5 can be easily obtained. To avoid this problem, two sub-regions must always be suppressed within any larger region that is released. By suppressing two sub-regions, any data obtained from geographic differencing could be from either of the suppressed areas and so is less likely to represent a privacy risk. These privacy risks can be largely eliminated by examining the characteristics information in the other sub-regions that could be supressed and carefully choosing sub-regions that are the best combination to ensure privacy can be protected.

The need to have two sub-regions supressed within any larger region applies to any future release of data for regions built from these sub-regions. This may mean restricting the subsequent release of data for larger regions that contain only one of the suppressed sub-regions.

An alternative approach to this problem is to combine or merge the sub-region that has the privacy risk with one of the other sub-regions (selected using a similar process to that described above). This method allows all of the data available to be released, while also obscuring the detail for the sub-region that contains the privacy risk.

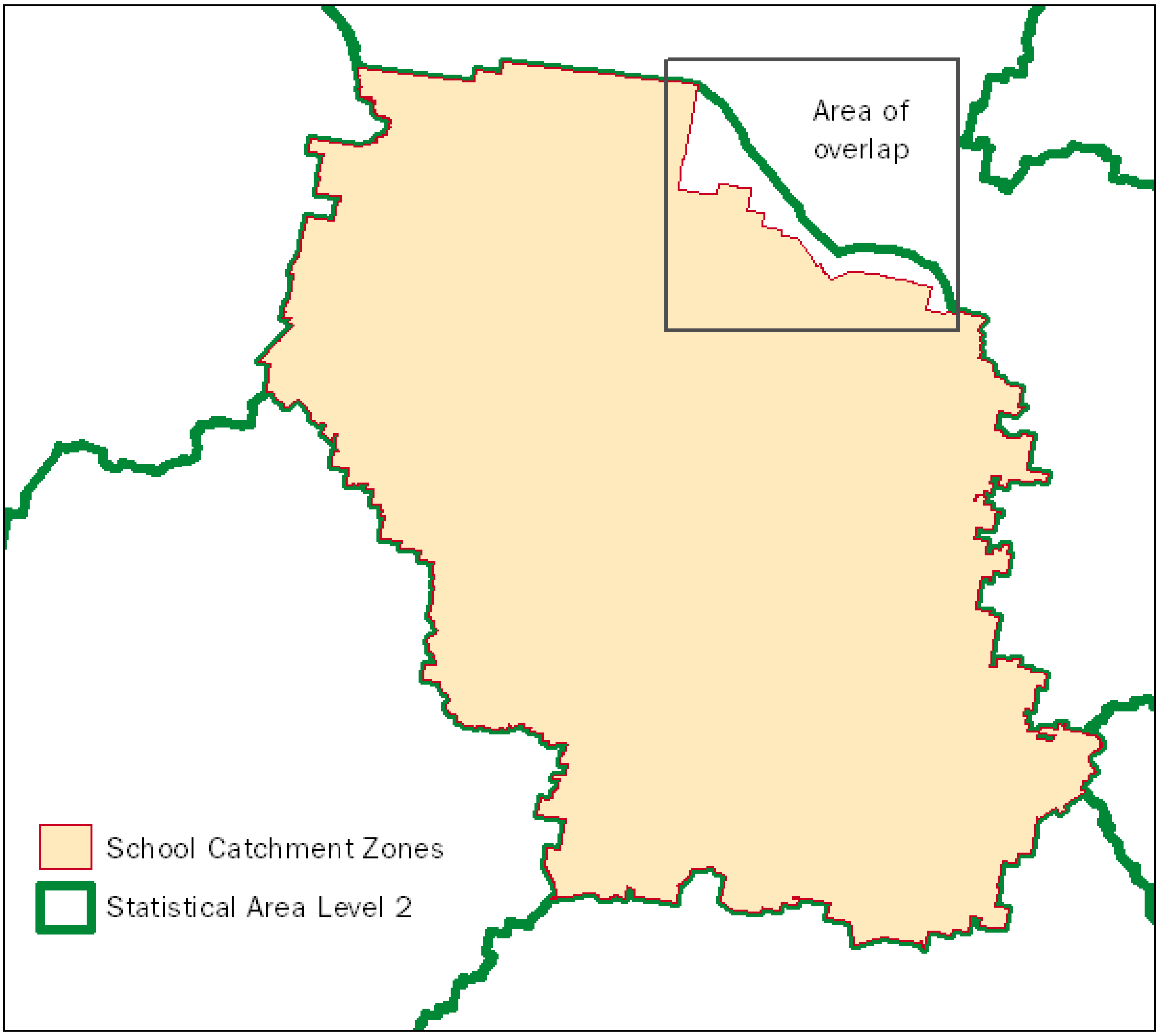
The complexity of these geographic differencing issues highlights the value in using nested, hierarchical regions, such as those included in the ASGS. Regional hierarchies of this type help to limit and manage this complexity. This is one of the reasons why the ASGS has been selected as the common geography in the Statistical Spatial Framework (SSF).

## Complex Geographic Differencing – overlapping regions

Using multiple region types (that are not nested, hierarchical regions) when releasing data can create a very large number of privacy risks to assess and manage. This is particularly relevant when location coordinate information is used to allocate unit records to the regions being released. Using location coordinates creates a high degree of certainty of identification if only a few statistical units (e.g. household or businesses) are geographically differenced within any of the overlapping areas.

To illustrate, consider a hypothetical example of releasing data on two region types – School Catchment Zones and ASGS Statistical Areas Level 2. As each of these boundaries is designed for a different purpose they will cover different areas, which will result in overlaps between regions. Any of these overlaps may pose a privacy risk due to the potential for geographic differencing. In some instances, these overlaps may be relatively small; in the hypothetical example below in diagram D, the small difference may result from both regions following the same suburb or local government boundaries for much of their extent. Where the area covered by these overlaps is small, the overlap area or areas may only include a very limited number of individuals who may be easily identifiable from the resulting geographically differenced data (other similar examples apply to data for organisations).

### Diagram D – Complex geographic differencing across 2 different region types



As with the simpler examples discussed above, the region being released must be compared with the regions included in previous releases to identify each instance of potential geographic differencing created by overlapping regions. This assessment may include a large number of overlap areas and may include complex cases where the boundaries of a number of individual regions, across several region types, create many sets of overlapping areas.

Identifying where regions overlap with each other is a relatively simple task using mapping or geographic information system (GIS) software. Identifying the level of risk that each one of these overlap areas represents is more challenging, the differenced data for each overlap area must be calculated and assessed for privacy risks.

Methods that can be used to ensure the confidentiality of the data where overlap areas create privacy risks include:

- modifying the boundary of regions to be released to eliminate the areas of overlap,
- combining regions to eliminate the areas of overlap, and
- using the techniques, such as category collapsing or suppression, which are described in the Confidentiality Information Series Information Sheet 4 How to confidentialise data: the basic principles (noting the issues associated with suppression outlined above).

## Glossary

A glossary of geographic and related terms used in this paper is available on the ABS website.

## Where can I get further information?

More information on the Statistical Spatial Framework can be found on the SSF webpage on the NSS website – www.nss.gov.au

Information about confidentiality is contained in the Confidentiality Information Series on the NSS website – www.nss.gov.au

More information about geographic boundaries and classifications can be found by visiting the ABS website – www.abs.gov.au/Geography

Information about using geographic regions is contained in the SSF Guidance Material “Using Geographic Boundaries and Classifications with Statistics” on the NSS website.

Any questions or comments on this paper or other statistical geography topics can be emailed to geography@abs.gov.au