DEPARTMENT OF
PRIMARY INDUSTRIES

The thinking behind
our everyday essentials

Analysis of attribute errors in the Victorian soil salinity GIS database in 2007

Victoria
The Place To Be

Published by: Department of Primary Industries, 2008
Primary Industries Research Victoria
Bendigo

May 2008

Also published on http://www.dpi.vic.gov.au

Authorised by the Victorian Government, Midland Highway, Epsom, Victoria

The National Library of Australia Cataloguing-in-Publication entry:

Author: Clark, R. M., 1955-

Title: Analysis of attribute errors in the Victorian soil salinity GIS database in 2007 / Rob Clark, Wayne Harvey, Steve Williams.

Publisher: Bendigo, Vic.: Department of Primary Industries, 2008.

ISBN: 9781741998153 (pbk.)

Notes: Bibliography.

Subjects: Soils, Salts in--Victoria--Databases.
Salinity--Victoria--Databases.
Geographic information systems.

Other Authors/Contributors:
Harvey, Wayne (Wayne David)
Williams, Steve (Steve Ross)
Victoria. Dept. of Primary Industries.

Dewey Number: 631.4160285

Summary

The Victorian soil salinity monitoring project, funded by the Department of Sustainability and Environment provides support for standardised collection of data showing the extent and severity of soil salinity. This project also manages the storage of the data in the Corporate Data Stores and distributes information products back to the community. The current soil salinity layer collates all available recorded soil salinity mapping in Victoria.

This report summarises a review and analysis of the data asset conducted to identify and understand errors within the data so that its management, condition and use can be improved.

Errors were assessed in the ArcGIS environment using a combination of statistical analysis and error trapping techniques. Intelligence associated with the original surveys was used to guide these approaches. Analysis of spatial error of mapped polygons was excluded from the review because of the inability to validate historical data with contemporary surveys due to the potential for the site extents to change over time (a result of changes in land management and climate). Consequently error analysis focused on validating site attribution.

The data is a collation of all mapping up to the present time and is not a snap shot at any one time. The dryland salinity discharge data asset is comprised of a series of mapped areas (polygons) with associated assessment data (attributes or fields). The attribute fields have changed over time as the ability to collect and process data improved and requirements of the data collectors and users evolved. There are three distinct epochs of data collection with different attributes, namely: pre 1995, 1995 to 2004, post 2004.

Analysis of the data shows that there are 9492 polygons mapped as saline across Victoria with a total of 66 444 attribute values. Of these 23 009 attribute values are missing or apparently incorrect. This represents almost 35% of all attribute values. Most errors occur in the Glenelg Hopkins CMA (9638 or 41% of the total errors) followed closely by the Mallee CMA (8120 or 36% of the total errors). The other CMAs have significantly less errors ranging from 1323 errors (6% of the total number of errors) in the North Central CMA region to only five errors in the North East CMA region.

The results show that more recent surveys tend to hold higher quality data (fewer errors) — an indication that the statewide coordination activity is proving effective. Exceptions to this rule occur where surveys have been undertaken without full adherence to the mapping methodology (such as in the Mallee). More recent surveys in the Wimmera also show high error rates but unlike the Mallee survey these are almost exclusively caused by a lack of assignment of a unique identifier for the sites (*dssite*). This omission has negligible impact and can easily be resolved.

This error analysis has highlighted the benefit of documenting all field surveys in the form of a brief report. Documentation of the survey method and results allow errors to be retrospectively addressed and would also provide useful information when assessment of change over time is the objective. It is essential that all such reports be published to ensure that they can be accessed in the future.

Use of existing reports to support rectification of errors/omissions in the two CMAs where most errors occur will reduce the total number of errors for the soil salinity database by around 11 500 (a decrease in the order of 50% of the errors/omissions).

Contents

Figures

Tables

Acknowledgement

The authors would like to acknowledge Leisa Macartney for her assistance in editing this report.

Analysis of attribute errors in the Victorian soil salinity GIS database in 2007

Rob Clark, Wayne Harvey, Steve Williams

1 Introduction

The Victorian soil salinity monitoring project is funded by the Department of Sustainability and Environment. It provides support for standardised collection of data showing the extent and severity of soil salinity, manages the storage of the data in the Corporate Spatial Data Library (CSDL) and distributes information products back to the community. The soil salinity layer collates all available recorded soil salinity mapping in Victoria.

A coordinated, comprehensive mapping of soil salinity has never been funded and so the data is not able to provide a comprehensive picture of salinity across Victoria at one time. Mapping exercises have generally been driven by local needs and funded from various state and federal initiatives. This makes it difficult to report on change in Victoria over the timeframes specified for the VCMC and SOE reporting. While acknowledging these limitations, the soil salinity layer is able to provide the most current record of soil salinity in Victoria and show how salinity has changed over varying intervals where an area has been remapped. The current data asset has been collected from projects driven by a number of organisations occurring from the 1970s to current day. Over that period the mapping methodology evolved until Matters (1987) formulated a standard. The standard developed by Matters has provided the base standard for field identification and mapping of soil salinity since 1987. Allan (1996) documented and published the standard method and Clark and Fawcett (in prep) updated the method to take into account changes in the recorded attributes, improved field methods for distinguishing primary from secondary sites and the advent of GPS technology in improving site delineation.

The above factors (in particular the long collection period and consequent data currency issues) all contribute to potential variability in the dataset. The current exercise reviewed the data asset to identify and understand errors within the data so that its management and use can be improved. The results of the review are documented in this report along with issues that could not be addressed within the scope of this exercise.

2 Dryland soil salinity mapping in Victoria

The data is a collation of all mapping to the present time and is not a snap shot at any one time. Where sites have been remapped we have included the latest version of the data and estimates of extent are based on the most recent data. Table 1 shows the current best estimate of affected soil salinity in Victoria.

Figure 1 shows the distribution of mapped dryland soil salinity in Victoria and it is evident that significant amounts of soil salinity have been mapped in all Catchment Management Authority (CMA) regions except East Gippsland. Little dryland soil salinity has been reported in the East Gippsland CMA region probably due to a combination of climate, soils, land use and land management practices and it is not expected that many new areas will be identified by future mapping programs.

Table 1 Summary of mapped dryland soil salinity

Catchment Management Authority	Salinity discharge [ha]
Wimmera	21 789
Glenelg Hopkins	27 435
North Central	27 114
Goulburn Broken	4 778
North East	1 311
Corangamite	25 162
Port Phillip	2 890
West Gippsland	24 160
Mallee	127 756
East Gippsland	273
Total	262 668

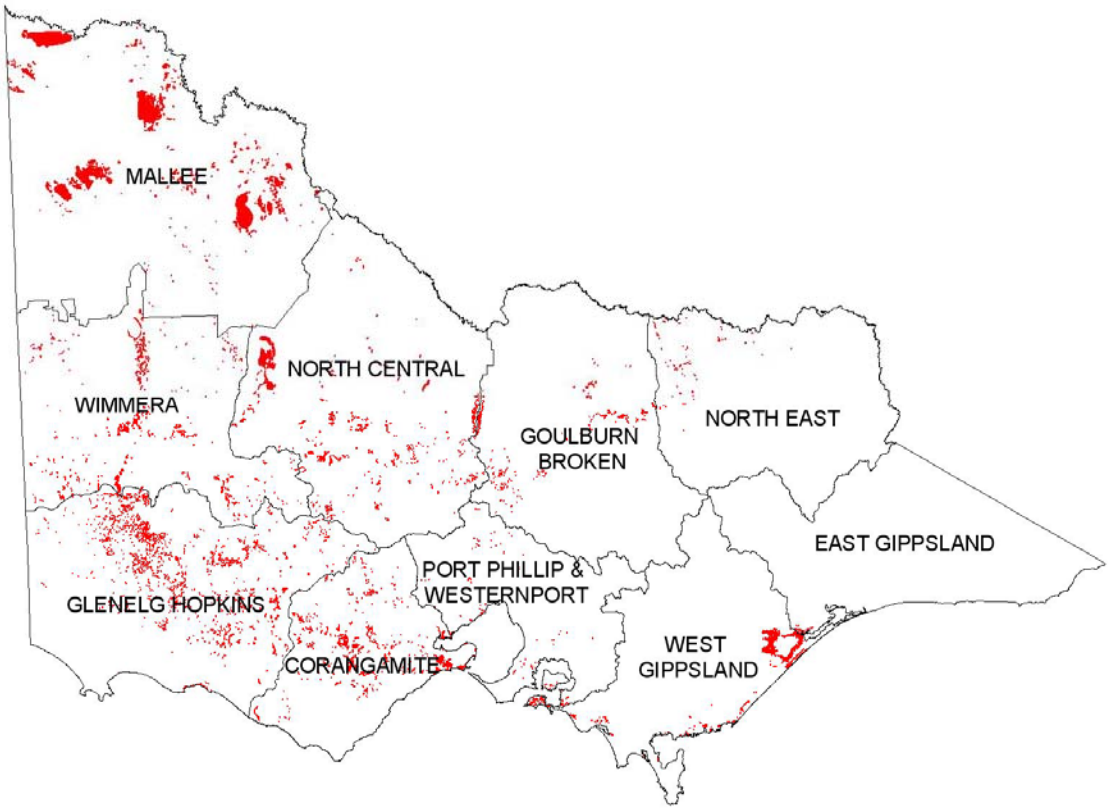


Figure 1 Mapped dryland salinity in Victoria

3 Description of dryland discharge data structure

To properly understand an analysis of the errors and omissions of attributes of the dryland salinity database, it is necessary to have some appreciation of the attributes attached to each polygon.

The dryland salinity discharge data asset is comprised of a series of mapped areas (polygons) with associated assessment data (attributes or fields). The attribute fields have changed over time as the ability to collect and process data improved and requirements of the data collectors and users evolved. There are three distinct sets of attributes: pre 1995, 1995 to 2004, post 2004.

3.1 Attributes recorded for each polygon prior to 1995

Valid field values for all attributes described below are listed in Appendix 1.

dssite Unique identification number given to each site

assessdate Date of field assessment (indicator of data currency)

dstype Contains code indicating primary or secondary salinity

severity Index of relative site salinisation calculated from field assessment of severity

The *severity* attribute for this and subsequent time periods is a single digit index indicative of soil salinity levels for a whole polygon. The class that occupies the largest spatial portion of a polygon is designated as the severity value, but if two or three classes occupy equal areas, the highest class is assigned. A severity value of four indicates an area that was considered to be saline but its severity was unknown or not recorded. Table 2 shows the EC_{se} range for each class used up until mid 2004 in dSm^{-1}.

Table 2 The four class salinity severity scheme used until mid 2004

Salinity level	Salinity class number	EC_{se} range [dSm^{-1}]
Non saline	0	
Slightly saline	1	2–4
Moderately saline	2	4–8
Highly saline	3	>8

3.2 Attributes recorded for each polygon in the period 1995 to mid 2004

Valid field values for all attributes described below are listed in Appendix 1.

dssite Unique identification number given to each site

assessdate Date of field assessment (indicator of data currency)

sev1 Percentage of a polygon rated as low level salinity (based on a 4 class system)

sev2 Percentage of a polygon rated as moderate level salinity (based on a 4 class system)

sev3 Percentage of a polygon rated as high level salinity (based on a 4 class system)

sev4 Percentage of a polygon identified as saline but the severity level not recorded

sev9 A non-saline polygon lying within a larger saline polygon (indicated by a value of '9')

severity Index of relative site salinisation calculated from field assessment of severity

dstype Indicates primary or secondary salinity

dsprocess	Indicates whether the process driving salinity at the site is discharge or waterlogging
datacapt	Lists the method used to delineate the spatial extent of the saline area and gives an indication of the likely spatial accuracy of the polygonal data.

After 1995, severity was little used but continued to be calculated to allow comparison with previously mapped sites if required. Instead, from 1995 it was decided to begin recording the field assessment of the percentage of each salinity class attributed to a polygon (*sev 1*, *sev2*, *sev3*) as a number up to three digits between 1 and 100. At this time *sev 4* still indicated salinity severity unknown and for the period 1995 to mid 2004, the rule *sev1+sev2+sev3+sev4* = 100 must be true at all sites else this constitutes an error.

The attribute *dsprocess* was added in 1995 to accommodate projects operating in areas where waterlogging was considered to the dominant process driving surface expressions of soil salinity rather than groundwater discharge. It is difficult to assess this attribute in the field as a sound understanding of the depth to ground water and the dominant groundwater processes in an area are required. Generally the decision is made after consultation with a hydrogeologist familiar with the survey area.

The attribute *datacapt* was added in 1995 to enable storage of an indicator of the likely spatial accuracy of the polygons.

3.3 Attributes recorded for each polygon after mid 2004

Valid field values for all attributes described below are listed in Appendix 1.

dssite	Unique identification number given to each site
assessdate	Date of field assessment (indicator of data currency)
sev1	Low level salinity (based on a 5 class system)
sev2	Moderate level salinity (based on a 5 class system)
sev3	High level salinity (based on a 5 class system)
sev4	Extreme level salinity (based on a 5 class system)
sev8	An area identified as saline but the severity level not recorded
sev9	A non saline polygon lying within a larger saline polygon
severity	Index of relative site salinisation calculated from field assessment of severity
dstype	Indicates primary or secondary salinity
dsprocess	Indicates whether the process driving salinity at the site is discharge or waterlogging
datacapt	Lists the method used to delineate the spatial extent of the saline area and gives an indication of the likely spatial accuracy of the polygonal data.

In mid 2004 after developing a regional salinity field guide for north western Victoria it was decided to add a fifth class to the salinity classification scheme by subdividing the existing *sev 3* (high level salinity) into high (*sev 3*) and extreme (*sev 4*) salinity to create a fifth class to accommodate the higher soil salinity levels common in some parts of the state. Table 3 shows the EC_{se} range for each class used since mid 2004 in dSm^{-1}. This change in classification reverts to the class structure described by Matters (1987) and the United States Department of Agriculture (United States Salinity Laboratory Staff 1954).

The severity index was originally applied to a four class system used up until mid 2004 (non, slightly, moderately and highly saline). Where it is calculated for sites mapped after mid 2004

using five severity classes (non, slightly, moderately, highly and extremely saline), amalgamation of the high and extreme categories will allow comparison with sites mapped prior to mid 2004.

Table 3 The five class salinity severity scheme used for field assessment since mid 2004

Salinity level	Salinity class number	EC_{se} range [dSm^{-1}]
Non saline	0	<2
Slightly saline	1	2–4
Moderately saline	2	4–8
Highly saline	3	8–16
Extremely saline	4	>16

The implications of the changes in approach and database schema are clear. For areas classified as non saline, slightly or moderately saline, there are no changes. While any areas now classified as either highly (class 3) or extremely (class 4) saline may be equated to areas classified as highly (class 3) saline in surveys before mid 2004.

However it is not possible to retrospectively deconstruct the highly saline class areas mapped prior to 2004 to identify extremely saline areas within that class. This does not reduce the ability to identify changes over time, but should improve the ability to identify changes at the upper end of the scale in the future.

The field naming convention for the five class scheme still follows the same format as the old four class scheme i.e. the *sev1* attribute records the assessment of class 1 salinity expressed as a percentage of the area of the respective polygon. In this time period at all sites the rule *sev1+sev2+sev3+sev4+sev8* = 100 must be true else this constitutes an error.

The other major change to the attributes for post 2004 surveys involves areas identified as saline but where no severity classification is recorded. Prior to mid 2004 such areas were classified as *sev4*. For surveys since mid 2004, they have been classified as *sev8*. To eliminate confusion, all *sev4* polygons for surveys prior to mid 2004 will be reclassified as *sev8* polygons to align with post 2004 surveys.

4 Overview of statewide salinity data errors

Errors were assessed in the ArcGIS environment using a combination of statistical analysis and error trapping techniques. Intelligence associated with the original surveys was used to guide these approaches. Analysis of spatial error of mapped polygons was excluded from the review because of the inability to validate historical data with contemporary surveys due to the potential for the site extents to change over time (a result of changes in land management and climate). Consequently error analysis focused on validating site attribution.

Errors or omissions for each attribute in the statewide soil salinity layer polygon file are summarised in Table 4 with the exception of the s*ev1*, *sev2*, s*ev3* and *sev4* attributes. Omissions in these fields may be valid as a site may be classified as 100% *Sev1* for instance so no values in the other *severity* classes would be expected. To determine if there are errors or omissions in the 's*ev* fields' the sum of these fields for each site is evaluated to see if it equals 100. If the sum does not equal 100 then by implication the estimate in the field was in error or a field estimate has been entered incorrectly.

As a first step, the analysis identified errors/omissions in the *assessdate* attribute for the whole dataset. During migration of the data into new and different data base platforms, format changes have occurred to the *assessdate* (a result of how different databases treat some date formats) which made groups of records inaccurate. These have been sorted into a group called 'omitted/wrong'. The remaining sites (where the *assessdate* appeared to be valid) were split into three other groups according to their assessment date, leaving four groups (Group 1 = omitted/ wrong *assessdate*, Group 2 = 1975–1994, Group 3 = 1995–mid 2004, Group 4 = post mid 2004). Stratification of the data by date is required as some attributes (*sev1*, *sev2*, *sev3*, *sev4*, *sev8*, *sev9*, *dsprocess*, *datacapt*) were not in operation during some of the time periods (as described previously) and this influences the error analysis. Percentage errors for all other attributes are calculated after sites were stratified according to their *assessdate* value.

Error estimates for *assessdate*, *dssite*, *severity* and *dstype* are applied to the entire dataset as these attributes should have been recorded for all sites. Further analysis of additional fields is confounded where the *assessdate* field was omitted or in error as there is no certainty regarding which attribute set is appropriate for these sites (other than by inferring assessment dates by comparison with sites in close proximity if it is known that multiple surveys had not been conducted in the area). The data in Table 4 is an aggregate of statistics calculated from up to four *assessdate* groups.

Table 4 Summary of errors in dryland salinity mapping information attributes

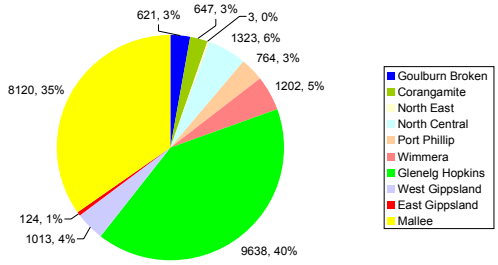
Attribute	Attribute description	Number of attribute records in error or omitted (%)
dssite	A unique site ID	2366 (25%)
assessdate	Date of field assessment	3561 (38%)
sev total	The sum of *sev1*, *sev2*, *sev3* and *sev4* (and/or *sev8*)	3382 (36%)
severity	Indicator of relative site salinisation	1163 (12%)
dstype	Type of salinity	5639 (59%)
dsprocess	The groundwater process driving the soil salinity	3391 (36%)
datacapt	The method used to delineate the extent of soil salinity	3430 (36%)

5 Analysis of errors for each CMA region

To assist the analysis of attribute errors/omissions, each CMA region was examined separately. Figure 2A shows the total number of errors/omissions for all attributes within each CMA region and also as a percentage of the total number of errors for all attributes across Victoria. Most errors occur in the Glenelg Hopkins CMA (9638 or 41%) followed closely by the Mallee CMA (8120 or 36%). The other CMAs have significantly fewer errors ranging from 1323 errors (6% of the total number of errors) in the North Central CMA catchment to only five errors in the North East CMA catchment.

A

Percentage of errors by CMA region



B

Total number of sites in each CMA region

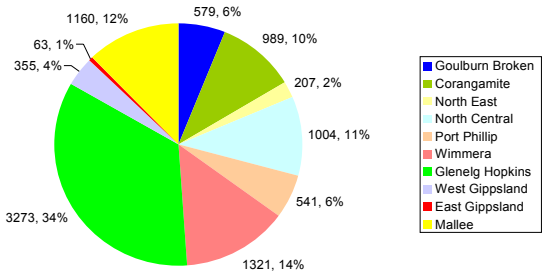


Figure 2 The percentage of errors and number of sites by CMA

Figure 2B displays the size of the data asset per CMA and represents the number of saline polygon sites in each CMA and as a percentage of the total number of sites across Victoria. If the numbers in Figure 2A are divided by the numbers in Figure 2B we derive the average error rate per record (or site) for each CMA (see Table 5 below).

Table 5 Average error rate per discharge site by CMA

CMA	Total number of errors	Total number of sites	Average error rate per site
Goulburn Broken	621	579+	1.07
Corangamite	647	989	0.65
North East	3	207	0.01
North Central	1323	1004	1.32
Port Phillip	764	541	1.41
Wimmera	1202	1321	0.91
Glenelg Hopkins	9638	3273	2.94
West Gippsland	1013	355	2.85
East Gippsland	124	63	1.97
Mallee	8120	1160	7.00

Figure 2A shows that 75% of all attribute errors lie within the Glenelg Hopkins and Mallee CMA regions. It would seem logical to target these two CMAs for the initial work into correcting errors/omissions in the soil salinity database. There is a high likelihood of reward for effort as in both these catchments the field surveys are relatively recent and well documented. For each CMA detailed descriptions of the errors and omissions, and the steps that can be taken to rectify them, are given in the following sections.

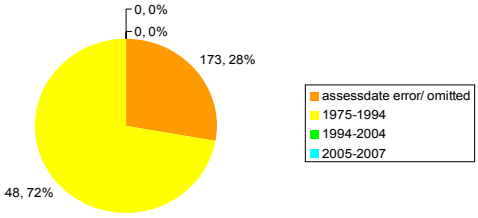
In the following sections, for each CMA region, this report has produced:

- A pie chart showing the number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole CMA region.
- A bar graph showing the number of sites where all attributes are correct/present (blue) versus the number of sites where any attributes are errors/omissions (red) for each *assessdate* group.
- A map showing the distribution of mapped soil salinity within a CMA catchment. Each site is colour coded and red represents sites with any errors/omissions and blue represents sites where all attributes are correct/present.
- A pie chart colour coded similarly to the accompanying map so that red represents the percentage (and number) of sites with any attribute errors or omissions and blue shows the percentage (and number) of sites where all attributes are accurate compared to the total number of sites within that CMA region.
- A break down of the number of each type of error/omission within each *assessdate* group.
- Recommendations for rectifying the errors/omissions within each group.
- An assessment of the ability to rectify the errors/omissions.

5.1 Goulburn Broken CMA region

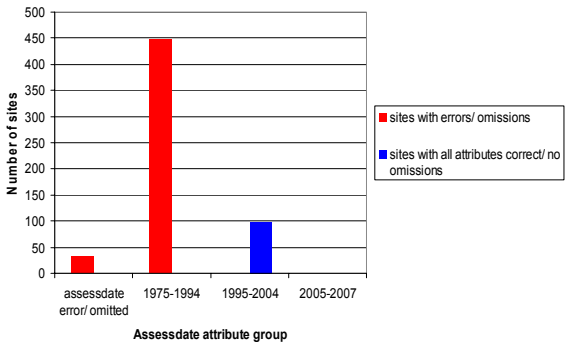
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole Goulburn Broken CMA region (621 attribute errors and 579 sites in total).



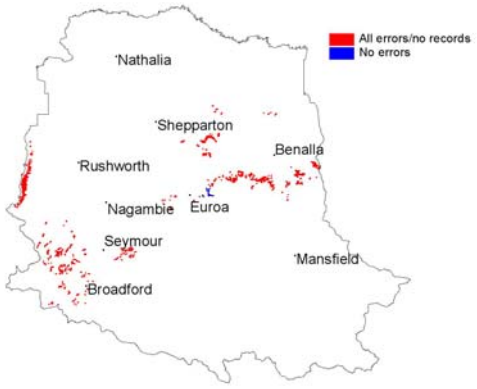
B

Number of sites with attribute errors/omissions and correct/no omissions for each *assessdate* group in the Goulburn Broken CMA region.



C

The distribution of mapped soil salinity in the Goulburn Broken CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions (including *assessdate* errors/omissions) or all correct/present records in the Goulburn Broken CMA region.

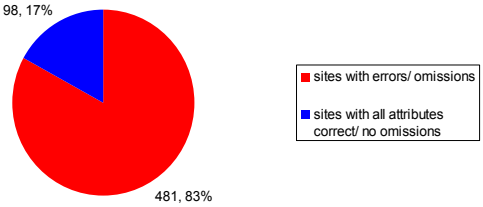


Figure 3 Graphical representation of the error analysis for the Goulburn Broken CMA region

The analysis in Figure 3A is based on total errors whilst the Figures 3B, 3C and 3D relate to analysis of errors viewed from a site basis (number of sites with one or more errors, number of sites without any errors). This reporting pattern is repeated in the sections for the subsequent CMAs.

The key observations regarding soil salinity data in the Goulburn Broken CMA region are as follows:

1. For the omitted/wrong date group (33 sites in total): 33 sites also had errors in *dstype, dsprocess, datacapt, sev_totals;* eight sites also had errors in *dssite,* giving a total of 173 errors/omissions.
2. For salinity mapped in the period 1975–94 (448 sites in total): 448 sites did not have a *dstype* attribute.
3. For salinity mapped in the period 1995–2004 (98 sites in total): all attributes were present and correct.

Recommendations for rectifying errors/omissions in Goulburn Broken CMA based on the analysis are:

1. For the omitted/wrong date class:

- Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion.
- Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.

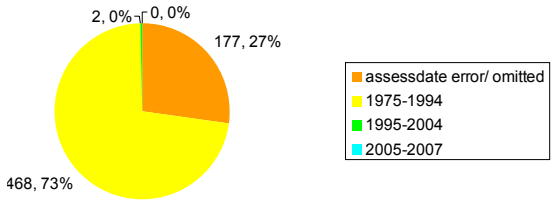
2. For salinity mapped in the period 1975–94:
    - Attempt to source the original data and associated reports to identify if the *dstype* attribute was recorded at the time of the original survey.

It may be possible to resolve the *assessdate* errors at 33 sites which may in turn reduce the number of attribute errors. However it is less likely that the 448 *dstype* attribute values for the 1975–1994 mapping will be found as there is little documentation available for these surveys.

5.2 Corangamite CMA region

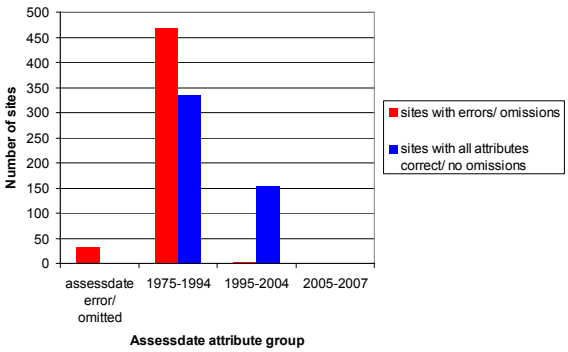
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole Corangamite CMA region (647 attribute errors and 989 sites in total).



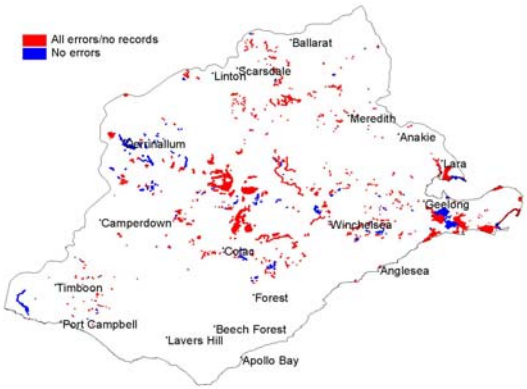
B

Number of sites with error/omissions and correct/no omissions for each *assessdate* group in the Corangamite CMA region.



C

The distribution of mapped soil salinity in the Corangamite CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/ present records in the Corangamite CMA region (989 sites in total).

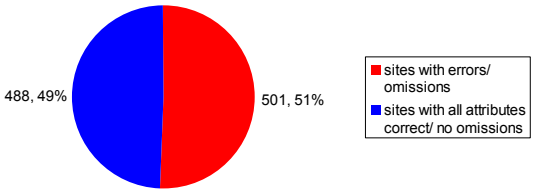


Figure 4 Graphical representation of the error analysis for the Corangamite CMA region

The key observations regarding soil salinity data in the Corangamite CMA region are as follows:

1. For the omitted/wrong date group (32 sites in total): 31 sites had errors in *dstype, dsprocess, datacapt* and *sev_totals*; also 18 sites had errors in *dssite* with *severity* having three errors (a total of 177 errors/omissions).
2. For soil salinity mapped in the period 1975–94 (803 sites in total): the *dstype* attribute had 468 errors.
3. For soil salinity mapped in the period 1995–2004 (154 sites in total): *total sev%* and *dsprocess* had one error each.

Recommendations for rectifying errors/omissions in Corangamite CMA region based on the analysis are:

1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion.
   - Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.

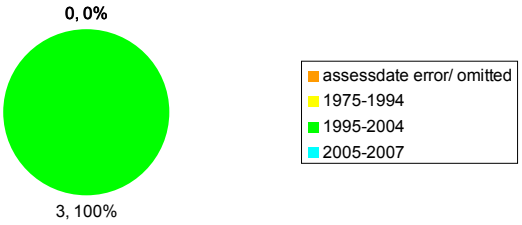
2. For salinity mapped in the period 1975–94:
    - Attempt to source the original data and associated reports to identify missing *dstype* records.

It may be possible to resolve the *assessdate* errors at 32 sites which may in turn reduce the number of attribute errors. However it is less likely that the 468 *dstype* attribute values for the 1975–1994 mapping will be found as there is little documentation available for these surveys.

5.3 North East CMA region

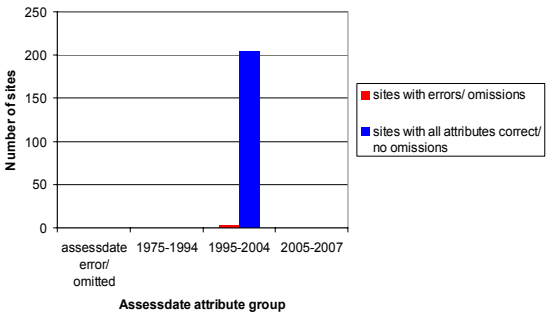
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole North East CMA region (three attribute errors and 207 sites in total).



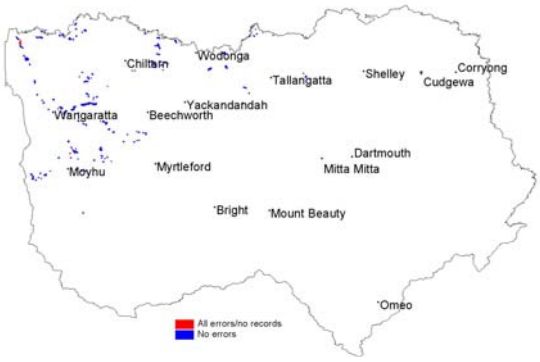
B

Number of sites with errors/omissions and correct/no omissions for each *assessdate* group in the Corangamite CMA region.



C

The distribution of mapped soil salinity in the North East CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/present records in the North East CMA region (207 sites in total).

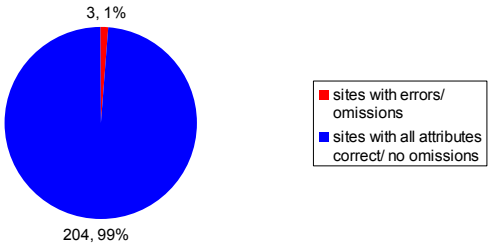


Figure 5 Graphical representation of the error analysis for the North East CMA region

The key observations regarding soil salinity data in the North East CMA region are as follows:

1. For salinity mapped in the period 1995-2004 (207 sites in total): *total sev%*, *dsprocess* and *datacapt* had one error each.

Recommendations for rectifying errors/omissions in North East CMA region based on the analysis are:

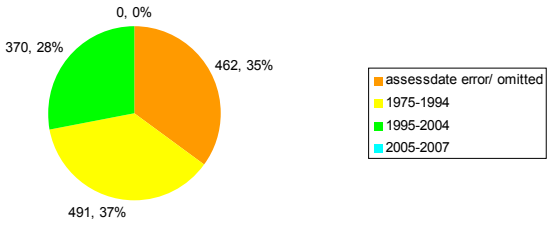
1. Check associated report to identify missing record.

There is a high probability that these errors can be rectified as the surveys are recent, staff responsible for completed the surveys are still employed in the region and the surveys are well documented.

5.4 North Central CMA region

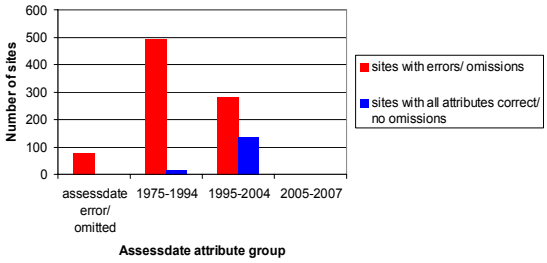
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole North Central CMA region (1323 attribute errors and 1004 sites in total).



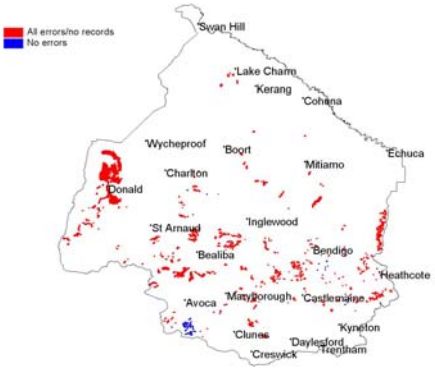
B

Number of sites with error/omissions and correct/no omissions for each a*ssessdate* group in the North Central CMA region.



C

The distribution of mapped soil salinity in the North Central CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/ present records in the North Central CMA region (1004 sites in total).

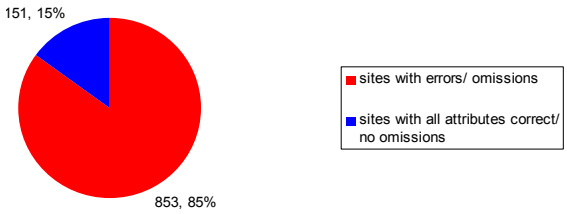


Figure 6 Graphical representation of the error analysis for the North Central CMA region

The key observations regarding soil salinity data in the North Central CMA region are as follows:

1. For the omitted/wrong date class (78 sites in total): 78 sites had errors in *sev_totals, dsprocess* and *datacapt*; also 77 sites had errors in *dssite.*
2. For salinity mapped in the period 1975–94 (506 sites in total): 491 sites did not have a *dstype* attribute.
3. For salinity mapped in the period 1995–2004 (420 sites in total): *dssite* attribute had 284 errors; *total sev%* had 66 errors; with *dsprocess* and *datacapt* each having 10 errors.

Recommendations for rectifying errors/omissions in North Central CMA region based on the analysis are:

1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion. Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.
2. For salinity mapped in the period 1975–94:

- Attempt to source the original data and associated reports to identify the missing *dstype* values.

3. For salinity mapped in the period 1995–2004:

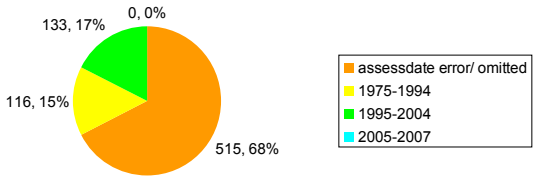
    - Attempt to source the original data and associated reports to identify the missing *dssite* numbers and the values for the *total sev%, dsprocess* and *datacapt* attributes.

Examination of previous versions of the database may make it possible to rectify the *assessdate* errors which may in turn reduce the errors for some of the other attributes. It is unlikely that it will be possible to find the missing *dstype* values for polygons mapped in the period 1975–94 as the surveys are not well documented. It may be possible to find the missing attribute values for the surveys from 1995–2004 as the surveys are relatively recent and there is an increased chance of locating the original survey records.

5.5 Port Phillip CMA region

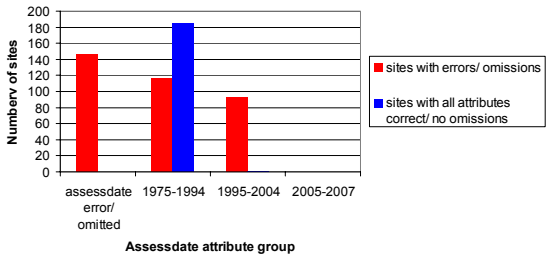
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole Port Phillip CMA region (764 attribute errors and 541 sites in total).



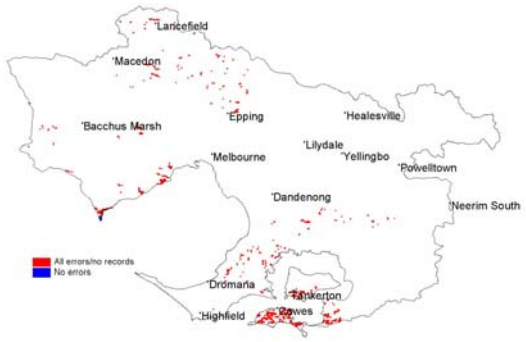
B

Number of sites with error/omissions and correct/no omissions for each *assessdate* group in the Port Phillip CMA region.



C

The distribution of mapped soil salinity in the Port Phillip CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/present records in the Port Phillip CMA region (541 sites in total).

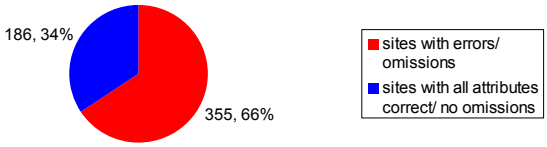


Figure 7 Graphical representation of the error analysis for the Port Phillip CMA region

The key observations regarding soil salinity data in the Port Phillip CMA are as follows:

1. For the omitted/wrong date class (146 sites in total): 131 sites had errors in *dssite;* 139 sites also had errors in *sev_total;* all 146 sites had errors/omissions in *assessdate*; and the *dstype, dsprocess* and *datacapt* attributes each had 33 errors.
2. For salinity mapped in the period 1975–94 (301 sites in total): 116 sites did not have a *dstype* attribute.
3. For salinity mapped in the period 1995–2004 (94 sites in total): *dssite* attribute had 93 errors and *total sev%* had 40 errors.

Recommendations for rectifying errors/omissions in Port Phillip CMA based on the analysis are:

1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion.
   - Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.
2. For salinity mapped in the period 1975–94:

- Attempt to source the original data and associated reports to identify the missing *dstype* numbers.

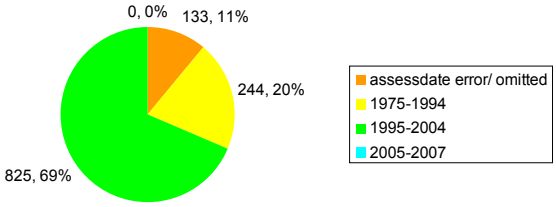
3. For salinity mapped in the period 1995–2004:
    - Attempt to source the original data and associated reports to identify the missing *dssite* numbers and the *total sev%* values.

Examination of previous versions of the database may make it possible to rectify the *assessdate* errors which may in turn reduce the errors for some of the other attributes. It is unlikely that the missing *dstype* values will be found for the 1975–94 survey data as the surveys were not well documented. It may be possible to locate the missing *dssite* numbers and the *total sev%* values for the 1995–2004 survey data as the surveys are relatively recent and there is an increased chance of locating the survey records.

5.6 Wimmera CMA region

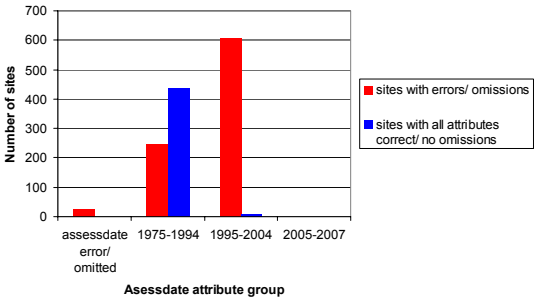
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole Wimmera CMA region (1202 attribute errors and 1321 sites in total).



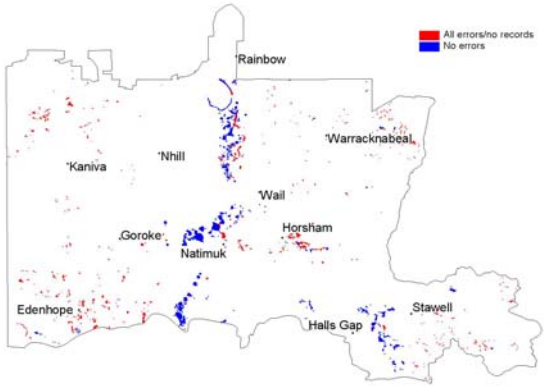
B

Number of sites with error/omissions and correct/no omissions for each *assessdate* group in the Wimmera CMA region.



C

The distribution of mapped soil salinity in the Wimmera CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/present records in the Wimmera CMA region (1321 sites in total).

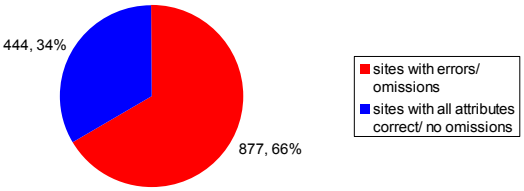


Figure 8 Graphical representation of the error analysis for the Wimmera CMA region

The key observations regarding soil salinity data in the Wimmera CMA region are as follows:

1. For the omitted/wrong date class (24 sites in total): 24 sites had errors in *assessdate, sev_totals, dstype, dsprocess* and *datacapt;* 13 sites also had errors in *dssite*, giving a total of 133 errors/omissions.
2. For salinity mapped in the period 1975–94 (680 sites in total): 244 sites did not have a *dstype* attribute.
3. For salinity mapped in the period 1995–2004 (617 sites in total): *dssite* attribute had 609 errors; *dstype* had nine errors; *dsprocess* had 85 and *datacapt* had 122 errors, giving a total of 825 errors/omissions.

Recommendations for rectifying errors/omissions in Wimmera CMA region based on the analysis are:

1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion. Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.

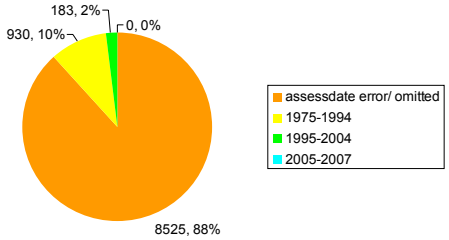
2. For salinity mapped in the period 1975–94:
    - Attempt to source the original data and associated reports to identify the rest of the *dstype* attribute numbers.
3. For salinity mapped in the period 1995–2004:
    - Attempt to source the original data and associated reports to identify the rest of the *dssite, dstype, dsprocess, datacapt*, and attribute values.

Examination of previous versions of the database may make it possible to rectify the *assessdate* errors which may in turn reduce the errors for some of the other attributes in that group. Wimmera salinity surveys have been reasonably well documented and there is some chance that the other missing attribute values may be located.

5.7 Glenelg Hopkins CMA region

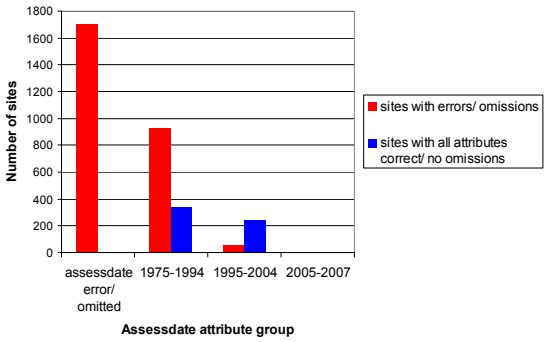
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole Glenelg Hopkins CMA region (9638 attribute errors and 3273 sites in total).



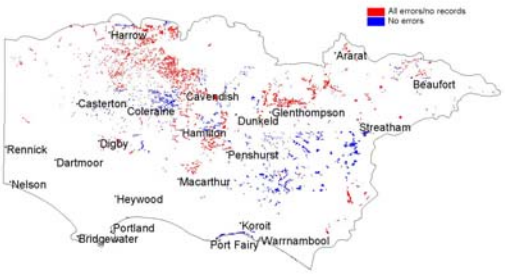
B

Number of sites with error/omissions and correct/no omissions for each *assessdate* group in the Glenelg Hopkins CMA region.



C

The distribution of mapped soil salinity in the Glenelg Hopkins CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/present records in the Glenelg Hopkins CMA region (1321 sites in total)

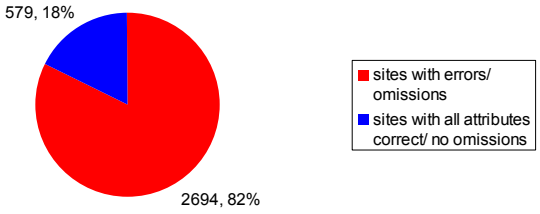


Figure 9 Graphical representation of the error analysis for the Glenelg Hopkins CMA region

The key observations regarding soil salinity data in the Glenelg Hopkins CMA region are as follows:

1. For the omitted/wrong date class (1708 sites in total): 66 sites had errors in the *dssite* attribute; 1708 sites also had errors in *assessdate, dstype; dsprocess* and *datacapt*; 1627 errors are present in *total sev%.*
2. For salinity mapped in the period 1975–94 (1269 sites in total), 930 sites did not have a *dstype* attribute.
3. For salinity mapped in the period 1995–2004 (296 sites in total): *dssite* attribute had seven errors; *dstype; dsprocess* and *datacapt* had 56 errors; and *total sev%* had eight errors.

Recommendations for rectifying errors/omissions in Glenelg Hopkins CMA region based on the analysis are:

1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion. Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.
2. For salinity mapped in the period 1975–94:

- Attempt to source the original data and associated reports to identify the *dstype* attribute values.

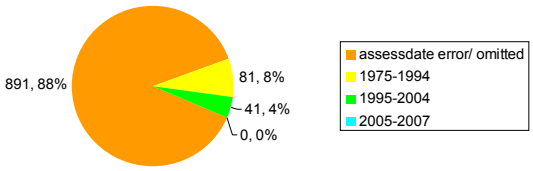
3. For salinity mapped in the period 1995–2004:
    - Attempt to source the original data and associated reports to identify the *dstype, dsprocess, datacapt* and *total sev%* attribute values.

All salinity surveys in the Glenelg Hopkins CMA region are well documented. It is likely that around 70% of the attribute errors will be able to be rectified. However initial inspection of the reports suggests that *dstype* values were not recorded at the time of survey.

5.8 West Gippsland CMA region

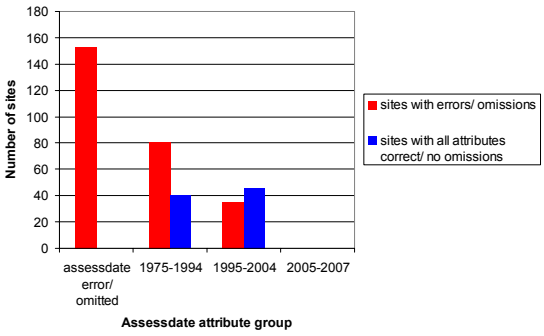
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole West Gippsland CMA region (1013 attribute errors and 355 sites in total).



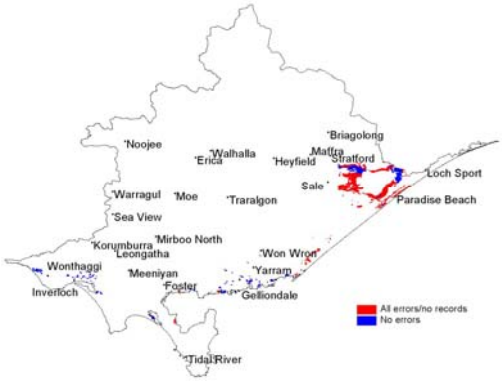
B

Number of sites with errors/omissions and correct/no omissions for each *assessdate* group in the West Gippsland CMA region.



C

The distribution of mapped soil salinity in the West Gippsland CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/ present records in the West Gippsland CMA region (355 sites in total)

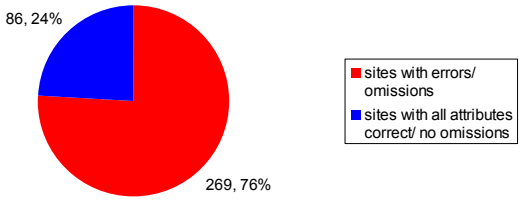


Figure 10 Graphical representation of the error analysis for the West Gippsland CMA region

The key observations regarding soil salinity data in the West Gippsland CMA region are as follows:

1. For the omitted/wrong date class (153 sites in total): 134 sites had errors in *dssite*; 153 had errors/omissions in *assessdate*; 152 sites in *datacapt* and *total sev%* had errors, while the *dstype* and *dsprocess* attributes both had 150 errors.
2. For salinity mapped in the period 1975–94 (121 sites in total): 81 sites did not have a *dstype* attribute.
3. For salinity mapped in the period 1995–2004 (81 sites in total): the *dssite* attribute had 35 errors; *dstype* and *dsprocess* had one error; *datacapt* and *total sev%* each had two errors/omissions.

Recommendations for rectifying errors/omissions in West Gippsland CMA region based on the analysis are:

1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion. Attempt to source the original data

and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.

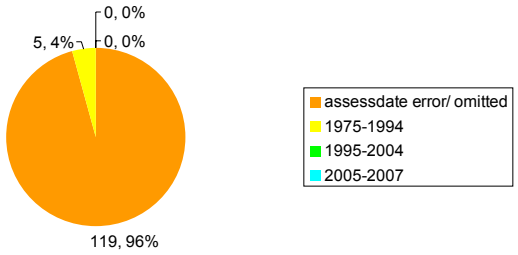
2. For salinity mapped in the period 1975–94:
    - Attempt to source the original data and associated reports to identify the *dstype* attribute records.
3. For salinity mapped in the period 1995–2004:
    - Attempt to source the original data and associated reports to identify the missing *dstype, dsprocess, datacapt and total sev%* attribute records.

Examination of previous versions of the database may make it possible to rectify the *assessdate* errors which may in turn reduce the errors for some of the other attributes in that group. West Gippsland salinity surveys are generally not well documented and it may be difficult to locate missing attribute values.

5.9 East Gippsland CMA region

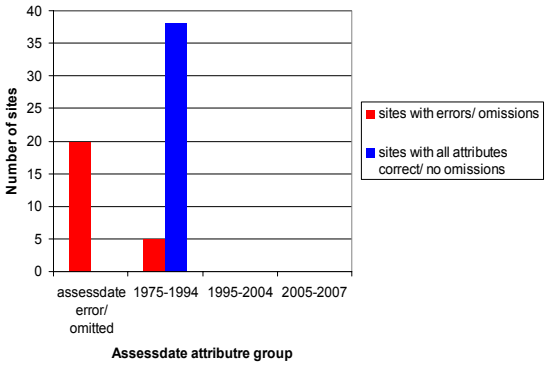
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole East Gippsland CMA region (124 attribute errors and 63 sites in total).



B

Number of sites with error/omissions and correct/no omissions for each *assessdate* group in the East Gippsland CMA region.



C

The distribution of mapped soil salinity in the West Gippsland CMA region.

D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/ present records in the West Gippsland CMA region (63 sites in total)

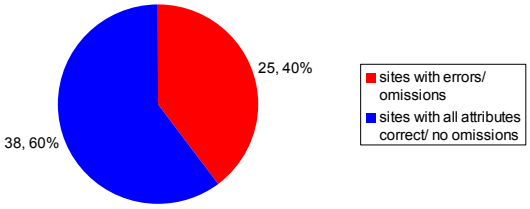


Figure 11 Graphical representation of the error analysis for the East Gippsland CMA region

The key observations regarding soil salinity data in the East Gippsland CMA region are as follows:

1. For the omitted/wrong date class (20 sites in total): 19 sites had errors in *dssite; assessdate, datacapt, total sev%, dstype* and *dsprocess* each had 20 errors, giving a total of 119 error/omissions.
2. For salinity mapped in the period 1975–94 (43 sites in total): five sites did not have a *dstype* attribute.

Recommendations for rectifying errors/omissions in East Gippsland CMA region based on the analysis are:

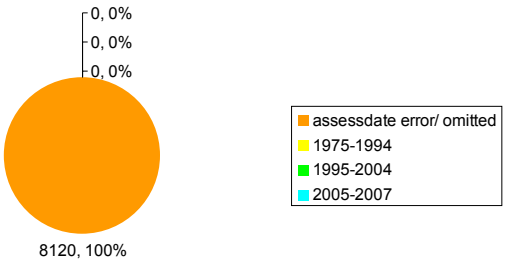
1. For the omitted/wrong date class:
   - Check previous versions of the GIS library to determine if the *assessdate* errors/omissions result from format conversion. Attempt to source the original data and associated reports to identify if the attributes *dstype, dsprocess, datacapt, sev_totals* and *dssite* were recorded at the time of the original survey.
2. For salinity mapped in the period 1975–94:
   - Attempt to source the original data and associated reports to identify the missing *dstype* values.

Examination of previous versions of the database may make it possible to rectify the *assessdate* errors which may in turn reduce the errors for some of the other attributes in that group. East Gippsland salinity surveys are generally not well documented and it may be difficult to locate missing attribute values.

5.10 Mallee CMA region

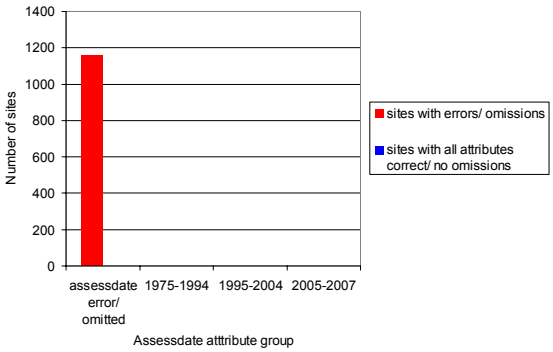
A

The number of attribute errors/omissions for each *assessdate* group and the percentage that each group forms of the total errors/omissions for the whole Mallee CMA region (124 attribute errors and 8120 sites in total).



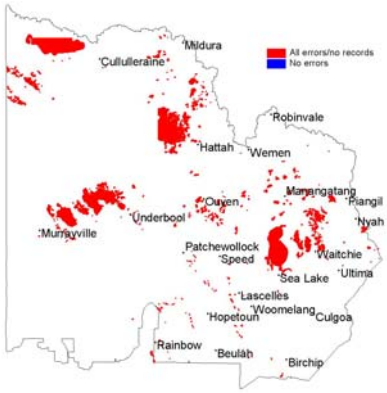
B

Number of sites with error/omissions and correct/ no omissions for each *assessdate* group in the Mallee CMA region.



C

The distribution of mapped soil salinity in the Mallee CMA region.



D

The percentage (and total number) of saline sites with any attribute errors/omissions or all correct/ present records in the Mallee CMA region (63 sites in total)

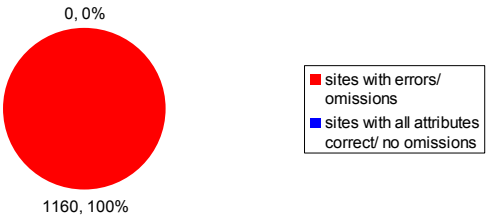


Figure 12 Graphical representation of the error analysis for the Mallee CMA region

The key observations regarding soil salinity data in the Mallee CMA region are as follows:

1. No attribute data came with the layer

Recommendations for rectifying errors/omissions in Mallee CMA region based on the analysis are:

1. *dssite, assessdate* and *datacapt* attribute can be populated by us from the 'Mapping the Mallee's saline land using available special data' report dated May 2007. The *dsprocess* field may be populated based on consultation with hydrogeologists with local knowledge.

It will not be possible to populate all attribute fields as the data was not collected at the time of survey.

6 Conclusion

Analysis of the data shows that there are 9492 polygons mapped as saline across Victoria with a total of 66 444 attribute values. Of these 23 009 attribute values are missing or apparently incorrect. This represents almost 35% of all attribute values. Table 6 ranks the CMAs in order of the number and percentage of errors. Most errors occur in the Glenelg Hopkins CMA (9638 or 41% of the total errors) followed closely by the Mallee CMA (8120 or 36% of the total errors). The other CMAs have significantly fewer errors ranging from 1323 errors (6% of the total number of errors) in the North Central CMA region to only five errors in the North East CMA region.

Table 6 Ranking the CMA regions in order of percentage attribute error/omissions

CMA	Total number of sites (%)	Total number of errors (%)
Glenelg Hopkins	3273 (34.5%)	9638 (41.1%)
Mallee	1160 (12.2%)	8120 (34.6%)
North Central	1004 (10.6%)	1323 (5.6%)
Wimmera	1321 (13.9%)	1202 (5.1%)
West Gippsland	355 (3.7%)	1013 (4.3%)
Port Phillip	541 (5.7%)	764 (3.3%)
Corangamite	989 (10.4%)	647 (2.8%)
Goulburn Broken	579 (6.1%)	621 (2.6%)
East Gippsland	63 (0.7%)	124 (0.5%)
North East	207 (2.2%)	3 (0%)
Total	9492	23455

The results show that more recent surveys tend to hold higher quality data (fewer errors) — an indication that the statewide coordination activity is proving effective. Exceptions to this rule occur where surveys have been undertaken without full adherence to the mapping methodology (such as in the Mallee). More recent surveys in the Wimmera also show high error rates but unlike the Mallee survey these are almost exclusively caused by a lack of assignment of a unique identifier for the sites (*dssite*). This condition has negligible impact and can easily be resolved.

The Glenelg Hopkins CMA contains 34% of the saline polygons, however 41% of the total attribute errors/omissions occur in the Glenelg Hopkins CMA region. This suggests that the Glenelg Hopkins data has a slightly higher rate of errors than might be expected. Similarly, the Mallee contains 12% of all mapped saline polygons, but 36% of all attribute errors/omissions which is a much higher rate of attribute error/omission than might be expected.

Table 6 shows that just over 77% of all attribute errors lay within the Glenelg Hopkins and Mallee CMA regions. It would seem logical to target these two CMAs for the initial work correcting errors/omissions in the soil salinity database. There is a high likelihood of reward for effort as, in both of these catchments, the field surveys are relatively recent and well documented. The reports accompanying the Glenelg Hopkins field surveys show that the *total sev%*, *dstype*, *dsprocess* or *datacapt* attributes were not recorded at the time of survey. It will not be possible to retrospectively allocate a value for the *dstype* attribute, but the data recorded in the reports should make it possible to allocate values for the other missing attributes. This should allow almost 7000 (over 70%) of attribute errors/omissions in the Glenelg Hopkins CMA region to be rectified. In the recent Mallee survey, which was largely a modelling exercise based on the available spatial information supplemented by a small amount of field work, no attributes were recorded for any of the polygons. Based on the companion report (Grinter and Mock 2007) it will be possible to

retrospectively allocate attribute values for the *dssite*, *assessdate*, *dsprocess* and *datacapt* attributes. However, the *severity*, *total sev%* and *dstype* attributes cannot be retrospectively allocated. This should allow over 4500 (over 55%) of attribute errors/omissions in the Mallee to be rectified. Rectifying errors in these two CMA regions will reduce the total number of errors for the soil salinity database by around 11 500 (a decrease in the order of 50% of the errors/omissions).

Other regions will be examined in order of the number of errors. The ability to rectify attribute errors/omissions will primarily depend on the availability of reliable source data, be it earlier versions of the GIS database to allow correction of *assessdate* errors caused by migration between GIS environments or hard copy records of individual sites for *dstype* and *total sev%* attributes. It may be possible to rectify some *assessdate* errors by comparison with nearby sites where the *assessdate* attribute is known. However, this assumes a good knowledge of the time of all surveys in the region and will only be used as a last resort.

Omissions of the *dssite* attribute constituted 11.3% (2650 sites) of the total number of attribute errors/omissions. These missing attribute values have no impact on the ability to analyse the data and in effect may be left blank as comparison over time may be addressed by spatial overlay of polygons.

This error analysis has highlighted the benefit of documenting all field surveys in the form of a brief report. Documentation of the survey method and results allows errors to be retrospectively addressed and would also provide useful information when assessment of change over time is the objective. It is essential that all such reports be published to ensure that they can be accessed in the future. Any corrections that are made to the soil salinity database in the course of this work will be documented in a published report for future reference. The aims of this error analysis and the following correction of errors are to improve the data quality and the confidence of users of the soil salinity database.

References

Allan MJ (1994) An assessment of dryland salinity in Victoria. Department of Conservation and Natural Resources, Victoria

Allan MJ (1996) Method for assessing dryland salinity in Victoria. Department of Natural resources and Environment, Victoria

Clark R, Fawcett J (in prep) Method for field assessment of dryland salinity in Victoria using GPS technology. Department of Primary Industries, Victoria

Gardiner B (2001) Salinity discharge mapping in the Corangamite region. Department of Primary Industries, Victoria

Grinter V, Mock I (2007) Mapping the Mallee's saline land using available spatial data. Agriculture Victoria, Attwood, Victoria.

Matters JM (1987) Dryland salinity component: criteria and methodology of assessment. Land Protection Division, Victoria.

United States Salinity Laboratory Staff (1954) Diagnosis and improvement of saline and alkali soils. Agricultural Handbook No.60. US Department of Agriculture, Washington

Appendix 1 Attribute descriptions and valid field values

dssite

A unique name for each site comprised of a string of text or integers.

Each site should be assigned a unique name. This will allow other attributes to be attached to the correct polygon.

assessdate

Record the date of the field assessment using six integers in the date format mm/dd/yy.

sev1

Percentage of the whole site considered to have low salinity levels

(Class 1 from Table 1Table 1). No more than 3 integer characters, with a value between 0 and 100.

Sev1 + Sev2 + Sev3+ Sev4+ Sev9 must equal 100.

sev2

Percentage of the whole site considered to be moderately saline (Class 2 from Table 1Table 1). No more than 3 integer characters, with a value between 0 and 100.

Sev1 + Sev2 + Sev3+ Sev4+ Sev9 must equal 100.

sev3

Percentage of the whole site considered to be highly saline (Class 3 from Table 1Table 1). No more than 3 integer characters, with a value between 0 and 100.

Sev1 + Sev2 + Sev3+ Sev4+ Sev9 must equal 100.

sev4

Percentage of the whole site considered to be extremely saline (Class 4 from Table 1Table 1). No more than 3 integer characters, with a value between 0 and 100.

Sev1 + Sev2 + Sev3+ Sev4+ Sev9 must equal 100.

sev9

Percentage of the whole site considered to be saline, but where the severity level cannot be classified. No more than 3 integer characters, with a value between 0 and 100.

Sev1 + Sev2 + Sev3+ Sev4+ Sev9 must equal 100.

This is also used for existing sites where the salinity class was not recorded at the time of survey. Allan (Allan 1994) said that salinity class had been recorded at only 57% of sites when the database was compiled in the early 1990s.

severity

A single digit index indicative of soil salinity levels for a whole polygon calculated from the Sev1, Sev2, Sev3 and Sev4 field estimates. This attribute is automatically generated in the office when field data is loaded in the GIS.

dstype

Classification of the site as either primary or secondary salinity using one or two text characters (N, I, NI). Where:

N = natural or primary salinity, I = induced or secondary salinity and NI indicates sites where both primary and secondary salinity exists together.

dsprocess

Characterisation of the groundwater process at the site as either discharge or waterlogging using one or two text characters (D, W, DW). Where D = a discharge related groundwater process, W = a waterlogging related groundwater process and DW indicates sites where both discharge and waterlogging processes occur together.

It may not be possible for field staff to record this attribute at the time of survey. It requires a good knowledge of the depth to groundwater and the dominant groundwater processes in the survey area. Consultation with a hydrogeologist familiar with the study area may be required.

datacapt

This attribute describes the method used to map the soil salinity using a one-integer code. The codes and their definitions are listed in Table 7. It gives an indication of the accuracy of the mapping for a site.

Table 7 *datacapt* attribute codes and their definition

Code	Method description
1	Global Positioning System (GPS) – differentially corrected
2	GPS – uncorrected
3	Aerial Photo Interpretation allied with field work
4	API – no field work
5	Base map used in conjunction with field work
6	Base map with no field work
7	Remote sensing in conjunction with field work
8	Remote sensing – no field work
9	Derived digital data – spatial modelling using GIS datasets with some field work
10	Derived digital data – spatial modelling using GIS datasets with no field work