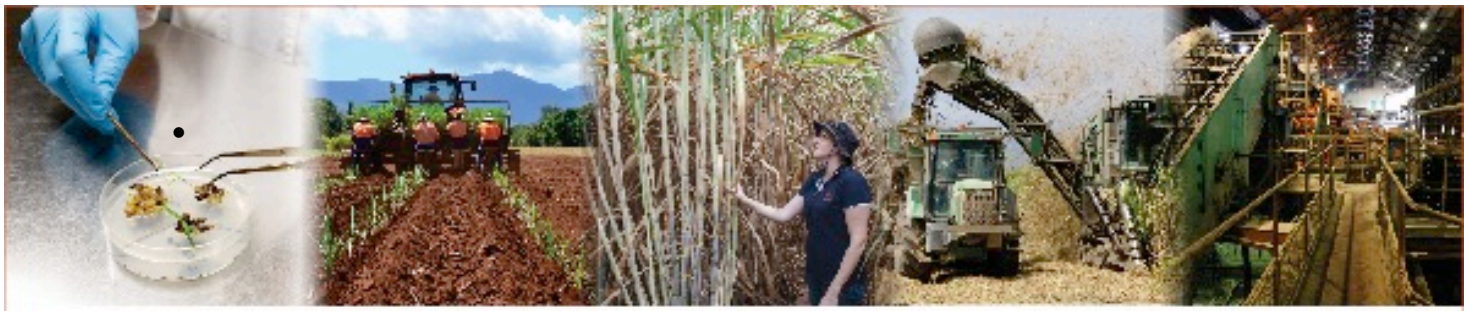


# FINAL REPORT 2015/045

## Sugar Industry Productivity and Data Recording Spatial Data Hub For Research and Extension

| | |
|---|---|
| **Final report prepared by:** | Robert Crossley |
| **Chief Investigator(s):** | Robert Crossley |
| **Research organisation(s):** | Agtrix |
| **Co-funder(s):** | Sugar Research Australia |
| **Date:** | 21 May 2018 |
| **Key Focus Area (KFA):** | Knowledge and technology transfer and adoption |

Sugar Research Australia

AGTRIX Harvest Management Systems

Please cite as: Crossley, R.(2018). Sugar industry productivity and data recording spatial data hub for research and extension. Final Report Project 2015/045. Sugar Research Australia Limited, Brisbane.

# ABSTRACT

The Australian sugar industry was an early adopter of Geographic Information Systems (GIS), and has considerable spatial data of where the crop has been grown. In some cases, data extends back for more than 20 years. When combined with the productivity data kept by the milling organisations, the data represents a considerable resource that could be used for research projects such as historical productivity analysis and bio-security response.

This data was difficult for research providers to access and use, however, as it was fragmented amongst multiple databases and archived files, and was stored in different formats using different codes to indicate varieties and classes.

The Sugar Data Hub project collated available data together into a single common spatial database, and enhanced the data by relating its productivity data for previous years regardless of previous block names, and to other data sources such as soils and weather.

Although the original concept of the project was to store this data centrally and provide access to that data by agreement from the data owners, privacy concerns precluded this model of distribution. Instead, this data was provided back to the owners for distribution to the research community.

Data quality varied across the industry, depending largely on the effort that the mill and farmers/ harvesters took in accurately consigning the bins to the mill.

Considerable benefits can be derived from the data collated in this project to the industry, including (a) consistency of data between regions, (b) the ability to access historical production data for a location, and (c) the ability to relate the production data to other spatial information such as soils, agro-climatic regions and GPS data from harvesters.

# EXECUTIVE SUMMARY

The Australian sugar industry was an early adopter of Geographic Information Systems (GIS), and have considerable spatial data of where the crop has been grown. In some cases, the data extends back for more than 20 years. When combined with the productivity data kept by the milling organisations, the data represents a considerable resource that could be used for research projects such as historical productivity analysis and bio-security response.

This data had been collated by different milling organisations over time and was fragmented amongst multiple databases and archived files, and was stored in different formats using different codes to indicate varieties and classes. This made it difficult for researchers to access and use.

The Sugar Data Hub project aimed at collating the industry data into a single common spatial database, and make it more accessible and easier to use for research..

**Introduction and Objectives**

The initial intention of this project aimed to build a secure 'Data Hub' to store all Australian sugarcane industry production data, and streamline the data owner's ability to provide this data to researchers securely. The data would include a spatial representation of the historical data of paddocks where sugar was grown, as well as production from those paddocks. This data was then to be provided to approved stakeholders such as researchers directly once permission was gained from the data owners.

The concept was changed in response to feedback about data privacy concerns during a series of regional workshops during December 2015. As a result, data was collated from the milling organisations, was still to be processed into a single common spatial database for the whole industry, as per the original concept, but the data access component was changed so that each mill data entity's data was not provided to anyone except to the data owner for them to distribute directly.

While under the new project concept, the researchers would not be able to retrieve data directly from a central hub, when this data is supplied from the different data owners it would have a common format and include additional attributes such as soils that were added to the data as part of the data collation process.

**Data processing**

For the purposes of this project, 15 mill data entities were defined to source data from, which were a combination of both organisations and regions. Of these, 13 mill data entities provided data for the project.

Data was sourced from the mill data entities and collated into a common format across all mills, with codes for variety and class converted to a common set using a standard translation table. The spatial data was often retrieved from copies made of historical spatial data made by cane officers at the end of the season, and included data in many formats in some cases.

These various sources were collated into a purpose-built SQL Server spatial database designed to include data across all regions and seasons.

The data that was collated was:

- ☐ The spatial extents of the paddocks for each season, in a format suitable for a GIS
- ☐ The paddock identifier (typically farm/ block/ sub-block), matching the paddock and rake data)
- ☐ Attributes, variety, class, age (for multiyear crops)

- ☐ Area of paddock that was fallow, cut, stood-over, ploughed out or taken for plants
- ☐ Date Time Harvested or Date Time Crushed
- ☐ Tonnes sugar cane
- ☐ Tonnes sugar

Once the data was collated, additional analyses were applied including:

- ☐ Soil type (from QDPI soils mapping)
- ☐ Closest weather station (from BOM active weather stations)
- ☐ Proportion overlap to previous year's data (even different organisations)
- ☐ Estimated previous harvest date

**Benefits for research**

The project has collated a substantial repository of historical data that is now available in a common format to much of the industry. This has been provided back to the participating milling organisations who provided the data, and can be requested directly from those organisations. While tabular productivity data is typically available for most regions, the value of linking that data to the spatial data and the standardisation of the data collation will enable researchers to:

1. Access consistent data between milling organisations and regions, enhancing the ability to do analyses across regions or in regions that have multiple milling organisations operating,
2. Relate the data to other spatial layers such as soils, weather data, agro-climatic regions, and harvester tracking, enabling refinement of recommendations such as varietal performance on different soils types in agro-climatic regions,
3. Access historical production for a particular location, regardless of what the location was called, or which organisation processed the cane in the past,
4. Understand productivity of areas that have been lost to the industry over time and target incentives that might return some areas to cane,
5. Analyse areas which may benefit from infrastructure to enhance the overall efficiency of the supply chain such as new loading zones/ sidings.

Historically, there are examples of the use of this type of data in these types of studies, but the effort of data collation has typically been repeated for each study. Further, the understanding of the limitations of the data, for example, due to poor consignment accuracy may have led to misinterpretation of the data.

**Conclusions and Recommendations**

The data repository may be maintained in the future to include data from each year as it becomes available, and the process of updating this data to include current year will be comparatively straight forward. This may be done by each organisation independently using the translation tables collated as part of this project, or be done centrally with data provided by the processors as was done in the project to date.

# TABLE OF CONTENTS

# 1. BACKGROUND

## 1.1. Introduction

The Australian sugar industry has been using Geographic Information Systems (GIS) to collate sugarcane crop information spatially for more than 20 years in some cases. It is the most comprehensive spatial dataset of any agricultural industry in Australia. This task originally related to a legislative requirement to map assigned land, but continued after this legislative requirement was discontinued, as the data had become a critical part of the sugar industry's business for planning and logistic optimisation.

The work has resulted in considerable data resources that could be used for research, but it was difficult for researchers to access and use that data quickly. The data for the industry is currently managed by eight milling organisations, but historical management dictated that the data was managed differently in different regions, in some cases even by the same organisation. The spatial data was fragmented amongst databases and archived files, and was stored in different formats using different codes to indicate varieties and classes.

In the past, each researcher was expected to collate this data, resulting in increased load on each research project using this data, and potentially the data being used inappropriately due to a lack of understanding about its limitations (e.g. poor consignment accuracy).

## 1.2. Benefits to Industry

While tabular productivity data is typically available for most regions, the value of linking that data to the spatial data and the standardisation of the data collation will provide researchers with a capacity to:

1. Access consistent data between milling organisations and regions, enhancing the ability to do analyses across regions or in regions that have multiple milling organisations operating,
2. Relate the data to other spatial layers such as soils, weather data, agro-climatic regions, and harvester tracking, enabling refinement of recommendations such as varietal performance on different soils types in agro-climatic regions,
3. Access historical production for a particular location, regardless of what the location was called, or which organisation processed the cane in the past,
4. Understand productivity of areas that have been lost to the industry over time and target incentives that might return some areas to cane,
5. Analyse areas which may benefit from infrastructure to enhance the overall efficiency of the supply chain such as new loading zones/ sidings.

## 1.3. Previous Research

The work builds on the previous association of Agtrix with the industry. Agtrix has supplied customised GIS systems to up to 85% of the industry milling sector over the last 24 years. Recent corporate changes, staff retirements and changes to Windows software may make the data less available in the future.

# 2. PROJECT OBJECTIVES

The project objectives that were originally proposed at the project inception were:

1. Engage the main data providers and stakeholders of the industry to establish the Privacy and Intellectual Property conditions required to enable participation for the major data

providers, and establish the acceptable protocols and mechanisms needed to allow data to be provided when requested.

2. Communicate with the industry what data may be stored in this data hub, and who may access that data, the types of data stored and the frequency and mechanism of updating the data repository.
3. Build on the existing infrastructure and systems to provide the data repository, data verification and quality assessment, and data migration pathways to be able to store the industry data, and provide access to the stakeholders that the industry agrees to.
4. Capture historical data of the spatial data representing paddocks where sugar was grown and production that is accessible to the industry for research purposes.
5. Implement data update processes that the industry can keep the data up to date into the future.
6. Implement data request protocols to facilitate the provision of data to the various data users, ensuring the rights of the original data owners are protected in the same manner that is done currently when data is provided electronically now.

Note: As an outcome of industry consultation at a series of workshops, the concept of the hub changed from being a physical repository that could be accessed by 3rd parties with permission of the data owners, to an improved spatially relevant data set that was provided back to the data suppliers for distribution at their discretion. This was approved as a modification to the original project objectives by the project steering committee.

# 3. OUTPUTS, OUTCOMES AND IMPLICATIONS

## 3.1. Outputs

As a result of this project and others, the sugar industry has begun discussions about data privacy and conditions required when providing data to 3rd parties. Over the course of this project, attitudes have changed to reflect the increased awareness of the responsibilities of data managers to protect privacy. While this caused delays in progressing the project, the discussions and resolutions have produced a better understanding of the issue and requirements when supplying data to 3rd parties, as well as the responsibility of those 3rd parties to protect the data.

The major output from the project is the spatial productivity data that has been collated for the majority of the sugar industry in a common format. Further analysis was done on the data to allocate the dominant soil type, agro-ecological zone and closest weather station to each paddock, as well as historical relationships to previous paddocks.

This data was then provided to the data owners in a format that suited their requirements (GIS files or SQL Server database).

The process of collating the industry data required the compilation of a translation table that allocated the codes from each data provider to a common set of codes for variety and cane class that were used in the database. This will be provided to Sugar Research Australia as per the project agreement.

## 3.2. Outcomes and Implications

The data collated in the project will provide a valuable asset over the tabular productivity data that is typically provided to researchers. The value of linking that data to the spatial data and the standardisation of the data collation will enable researchers to:

1. Access consistent data between milling organisations and regions, enhancing the ability to do analyses across regions or in regions that have multiple milling organisations operating.

2. Relate the data to other spatial layers such as soils, weather data, agro-climatic regions, and harvester tracking, enabling refinement of recommendations such as varietal performance on different soils types in agro-climatic regions.
3. Access historical production for a particular location, regardless of what the location was called, or which organisation processed the cane in the past. This will provide a valuable data set for historical analysis of report sensing imagery and its relationship to productivity.
4. Enable analysis of impacts of harvesting practices on subsequent yields where GPS tracking is available historically.
5. Understand productivity of areas that have been lost to the industry over time and target incentives that might return some areas to cane.
6. Analyse areas which may benefit from infrastructure to enhance the overall efficiency of the supply chain such as new loading zones/ sidings.

Historically, there are examples of the use of this type of data in these types of studies, but the effort of data collation has typically been repeated for each study. Further, the understanding of the limitations of the data, for example, due to poor consignment accuracy, may have led to misinterpretation of the data.

# 4. INDUSTRY COMMUNICATION AND ENGAGEMENT

## 4.1. Industry engagement during course of project

### 4.1.1. Steering Committee

A steering committee was engaged at the commencement of the project including representatives from the Australian Sugar Milling Council, CANEGROWERS, Sugar Research Australia and growers. Regular project meetings were held with this steering committee through the project, with the objective to:

1. Engage with industry to guide the project outcomes.
2. Approve changes to the project as a result of industry feedback, and to approve continuation of the project at a stop/ go milestone that depended on the level of industry involvement.
3. Help enlist support for the project, particularly from the milling sector that was to provide the data and CANEGROWERS who had to approve of the release of that data in some circumstances.

The support of this group was critical for the success of the project, as well as providing a profile of the outcomes of the project amongst key stakeholders.

### 4.1.2. Regional Project Workshops

A series of regional workshops were held at the beginning of the project to inform the industry of the objectives and to enlist participation from the data owners to provide data to the project. These workshops were held in Cairns, Ayr, Mackay, Bundaberg and Brisbane. In addition to these, meetings were held with most milling organisations directly and some grower organisations.

The objective of these workshops was to provide an opportunity to present cases where the concept of a data hub would provide researchers with a valuable resource, and to get feedback from the data owners (milling organisations and growers) on concerns about the use of that data.

The feedback from these workshops and follow up meetings guided the project to make changes to the proposed method of distribution, and provided awareness of the availability and use of this data once it was collected.

### 4.1.3. Industry Promotion

The project was also explained and promoted to industry at various opportunities including:

1. Agtrix's FarmMap conferences – attended by those responsible for spatial data in all milling sectors except Wilmar and Tully Sugar
2. Poster presented at the 2018 ASSCT conference (Crossley, Anderson and Andrews 2018)
3. Milestone reports to Sugar Research Australia
4. Direct contact with researchers that would use historical data in their work such as Andrew Robson (University of New England) and Joanne Stringer (SRA).

## 4.2. Industry communication messages

The Sugar Data Hub has collated available industry spatial and productivity data for the sugar industry, going back at least 10 years for most regions.

The spatial nature of the data allows analyses that were not possible previously, including:

1. analysing historical production for a paddock or location, regardless of what it may have been called in the past,
2. relating the data to soils or climate stations.

This data has been collated into a common format across all regions, and then passed back to the original data providers. This will allow researchers to use the data across different regions without having to translate data between data formats or codes used for varieties and class.

To access this data, researchers need to contact the data owners to ensure that the necessary privacy conditions are met with the supply of that data.

# 5. METHODOLOGY

The objective of the project was to collate industry productivity data from the sugar growing areas of Australia into a single database, including a spatial representation of the historical extents of paddocks where sugar was grown, as well as production from those paddocks. Once collated, the data would be more accessible and useable for researchers who needed to use that data.

The methodology used to accomplish this task is described in 3 phases:

1. Sourcing the data
2. Data Processing
3. Data Distribution

## 5.1. Sourcing the data.

The collation of the data for the project was reliant on provision of the data from the industry. Industry spatial and productivity data has been collated by the milling organisations for more than 20 years in some cases, but the structure and form of that data has changed due to mill ownership and computing system changes, and even varying data management between different regions.

For the purposes of this project, 15 mill data entities were identified to source data from, which were a combination of both organisations and regions (Table 1).

Regional workshops were then held in Gordonvale, Ayr, Mackay, Bundaberg and Brisbane in 2015 to explain the project objectives and to get approval of those data entities. This was followed up with further meetings with stakeholders where requested.

The intention of this project proposed in the initial meetings was to build a secure 'Data Hub' to store all Australian sugarcane industry production data, and streamline the data owner's ability to provide this data to researchers securely. This data was to be provided to approved stakeholders such as researchers directly once permission was gained from the data owners. Strong feedback from the industry expressing concerns about data privacy and new legislation was received during those workshops and attendees proposed an alternative strategy.

As a consequence, the concept of the Sugar Data Hub was changed. The modified version presented later to industry still included collating data from the milling organisations, and processing that data into a single common spatial database for the whole industry. However, the data access component was changed so that each mill data entity's data was not provided to anyone except for the data owner, and they would have the responsibility of distributing the data directly (Figure 1).

While under the new project concept, the researchers would not be able to retrieve data directly from a central hub, it would have a common format and include additional attributes such as soils that were added to the data as part of the data collation process, even where this data is supplied from the different data owners.

Approval to proceed on this basis was given by the Project Steering Committee, and support was sought from the 15 Mill Data Entities, of which 13 agreed to provide data (Table 1). There was a stop/ go Milestone that required a majority of the industry to be willing to participate in the project before the project would proceed, and this Milestone was met.

**Table 1. Mill Data Entities identified as data providers, the current owners and whether they agreed to participate in the project.**

| Mill Data Entity | Description | Current Data Owner | Participate in Project |
|---|---|---|---|
| Mossman | Farms serviced by Mossman Central Mill | Mackay Sugar | Yes |
| Tablelands | Farms serviced by Tablelands Mill farms | MSF Sugar | Yes |
| Mulgrave | Farms serviced by Mulgrave Mill farms | MSF Sugar | Yes |
| South Johnstone | South Johnstone Mill farms, including those historically serviced by Babinda and Mourilyan Mills | MSF Sugar | Yes |
| Tully | Farms serviced by Tully Mill farms | Tully Sugar | No |
| Herbert | Farms serviced by Victoria, Macknade Mills in Herbert Region | Wilmar | Yes |
| Burdekin | Farms serviced by Pioneer, Invicta, Inkerman and Kalamia Mills in Burdekin Region | Wilmar | Yes |
| Proserpine | Farms serviced by Proserpine Mill | Wilmar | Yes |
| Mackay | Farms serviced by Farleigh, Racecourse, Marian and Pleystowe Mills in Mackay Region | Mackay Sugar | Yes |
| Plane Creek | Farms serviced by Plane Creek Mill | Wilmar | Yes |
| Isis | Farms serviced by Isis Central Mill | Isis Central | Yes |
| Bundaberg | Farms serviced by Fairymead, Farleigh, Bingera and Millaquin Mills | Bundaberg Sugar | No |
| Maryborough | Farms serviced by Maryborough Mill | MSF Sugar | Yes |
| Rocky Point | Farms serviced by Rocky Point Mill | Heck Family | Yes |
| NSW | Farms serviced by Condong, Broadwater and Harwood Mills | Sunshine Sugar | Yes |

## 5.2. Data processing

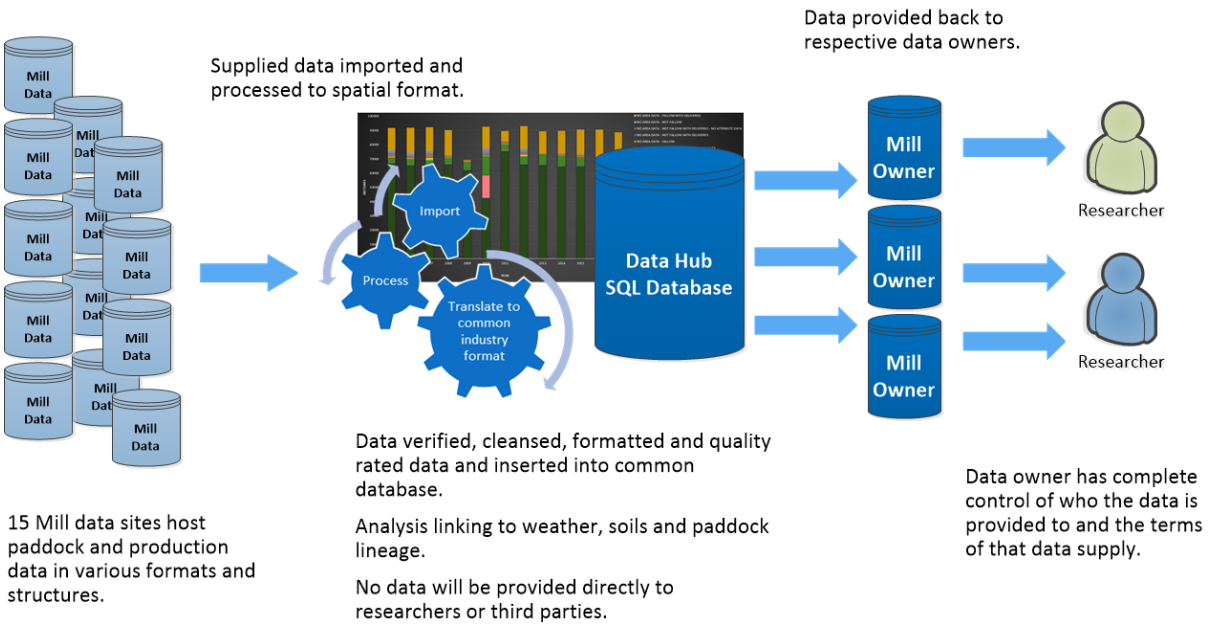



**Figure 1 Data processing workflow used for collating and distributing data in the project.**

## 5.3. Data Collation and Processing

There were three forms of data requested from each mill entity:

1. Spatial Data plus attributes. The spatial data is the mapping data for all the paddocks that sugar was grown for each season, stored in a format that includes the mapped extent of that paddock, and is accessible using Geographic Information Systems (GIS). Typically, this data also included the attributes of variety and class.
2. Mill Productivity Data, which comprised two components:
    a. Block data showing the end of season fate of each block in terms of the area that was harvested to supply cane to the mill from that paddock (excluding area used for plant, ploughed in or stood over), plus variety and class.
    b. Productivity Data relating to the delivered cane to the mill from each block, including date of delivery, weight and CCS.

Specifically, the data that was collated was:

1. The spatial extents of the paddocks for each season (location and area)
2. The paddock identifier (typically farm/ block/ sub-block), matching the paddock and rake data).
3. Attributes, variety, class, age (for multiyear crops)
4. Area of paddock that was fallow, cut, stood-over, ploughed out or taken for plants
5. Date Time Harvested or Date Time Crushed
6. Tonnes sugar cane
7. Tonnes sugar

This data was typically sourced by Agtrix visiting the milling organisation site to work with the person that knew where historical spatial data was held, usually the Chief Cane Inspector, and copied any historical data directly or arranged a transfer at a later date.

Data where this data was sourced included historical Farm Map directories, archives of season data kept in personal directories for each season, current FarmMap databases that include the more recent historical data, and data that had already been collated by the organisation.

In many cases, some of the older databases were stored in a format that is no longer accessible on Windows 10. This required a computer to be setup with Windows 7 to open those old data files to then save into a workable file format.

The sourced data was then scripted to create temporary import tables of a common format across all milling organisation using their own codes for variety and class. This data was then loaded into a Microsoft SQL Server database designed to include data across all regions and seasons, and perform much of the data manipulation and spatial analysis required for the project.

Further data manipulation was done to insert that data into the generic data hub formats, including:

1. Using a translation table to provide a generic code for variety and class. This translation table has been provided to SRA as an Excel spreadsheet as part of the project outputs.
2. Adding dominant soil type, closest weather station, agro-ecological region from public sources of that data.
3. Spatially analysing previous paddocks that were historically overlapping that paddock using spatial queries on MS SQL Server.

These are described in more detail in Table 2.

## 5.4. Data Distribution

No direct access has been provided to 3rd parties of the Sugar Data Hub Database during the project, nor will there be in the future.

The processed data for each data provider was provided back to each data owner at the completion of the project, and they will have discretion as to who and how they provide the data to any 3rd party directly.

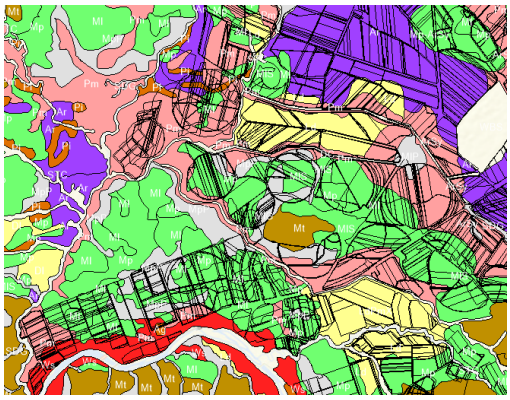
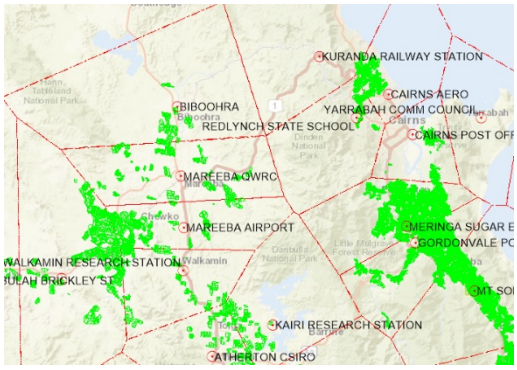
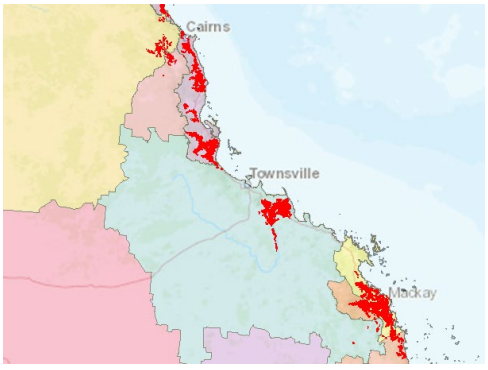
The data provided to each data entity included:

1. Table of Paddock data (See Appendix 6.2), including:
    a. area, variety and class of cane,
    b. tonnes cane harvested
    c. total sugar produced
    d. date of harvest
    e. dominant soil type
    f. closest weather station
    g. agro-ecological zone
2. Paddock relationships to previous paddocks
3. Monthly weather data for relevant weather stations
4. Collated soils mapping for Queensland

This data was provided as either a Microsoft SQL Server database, ESRI shape files or MapInfo tab tables.

At the completion of the project and after the data has been distributed, this data will be backed up onto encrypted back up media, and removed from the Agtrix networks. The backups will only be restored if further processing is requested in the future, and then only for the period required.

**Table 2 Explanation of additional data added to the spatial data through spatial analysis in the Sugar Data Hub Project.**

| | |
|---|---|
|  | **Proportion overlap to previous years data (even different organisations).** Each year's paddocks were intersected with paddocks from previous years to produce a table defining the relationship between paddocks and the paddocks that occupied that space previously.<br><br>This relationship table may be used to determine previous production on paddocks, regardless of what the paddocks were named in the past. |
|  | **Soil Mapping Unit (from QDPI soils mapping and Sunshine Sugar mapping).** Soils data from mapping done from Queensland Department of Primary Industries was downloaded from the Queensland Open Data Portal and compiled into one dataset. Where data from adjadcent studies overlapped, the largest scale data available was used. Soils mapping supplied by Sunshine Sugar through collaborative work with their growers was added to cover the sugar areas in NSW. The dominant soil mapping code was written to each paddock. |
|  | **Closest weather station (from BOM active weather stations).** Active weather stations were identified from Bureau of Meteorology (BOM) data, and the location of those stations were written to a GIS table. Veronoi polygons were created from these points (areas corresponding to where each weather station would be closest), and the closest weather station was written to each paddock record by intersection. Monthly average weather data from those stations were included in the data that was provided from the project. |
|  | **Agro-Ecological Zones (from Williams, J., Hook, R. & Hamblin, A. (2002)).** Zones of similar growing conditions were intersected with the paddocks to write the agro-ecological zone for each paddock. These zones may be used to compare data across similar climatic conditions, but which are in different organisations' data. |

# 6. RESULTS AND DISCUSSION

## 6.1. Industry Engagement and Participation

Industry data was sought from the fifteen mill data entities, which were a combination of the nine milling organisations and different regional data management sets within each organisation.

A series of regional workshops held throughout the sugar growing areas in Australia to explain the project objectives and to seek approval from those milling entities to participate. This was followed up with further meetings with stakeholders where requested.

The initial intention of this project proposed was to build a secure 'Data Hub' to store all Australian sugarcane industry production data, and to provide researchers with direct access to that data securely with the data owner's permission. The concept of the Sugar Data Hub was changed in response to strong feedback from the industry expressing data privacy concerns received during the regional workshops. The modified concept still included collating data from the milling organisations and processing that data into a single common spatial database for the whole industry, but the data access component was changed so that each mill data entity's data was only provided to the data owner, and they would distribute data to researchers directly.

Requests for participation were sought from the nine milling organisations that now managed the 15 data entities identified. Of these, seven organisations representing 13 data entities agreed to provide data to the project. The organisations participating and the years that they were able to supply data is provided in Figure 2 and Table 3.

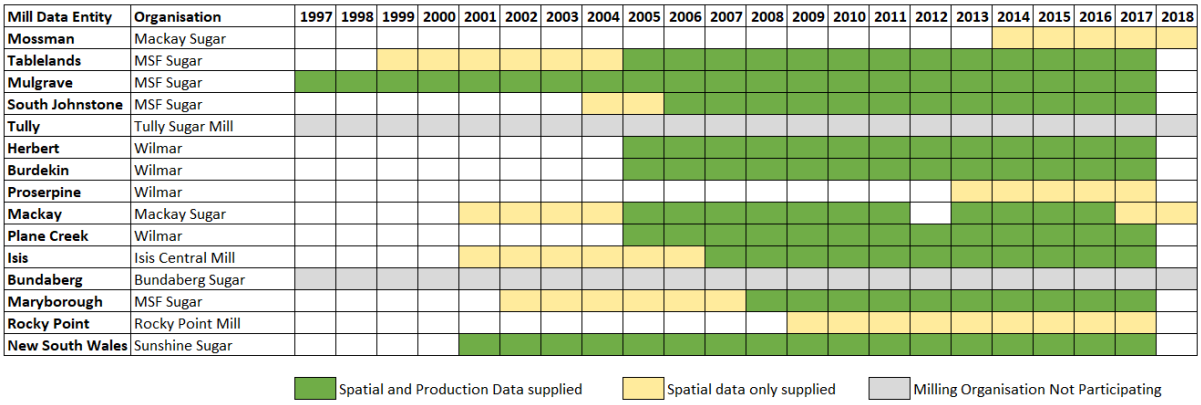



**Figure 2. Data provided by the milling organisations that own the data for the 15 data entities defined for the project.**

**Table 3. Mill Data Entities used for this project and current status of data collation.**

| Code | Mill Data Entity (Owner and Mills) Listed from north to south | Status |
|---|---|---|
| MOS | Mackay Sugar – Mossman | Spatial data collected for 2 years only, staff in process of collating it |
| TAB | MSF Sugar - Tablelands Mill | Spatial data from 1999, productivity data from 2006 |
| MUL | MSF Sugar - Mulgrave Mill | Spatial and productivity data from 1997 |
| SJM | MSF Sugar - South Johnstone Mill (includes Babinda and Mourilyan) | Spatial data from 2004, productivity data for South Johnstone Mill only from 2006, no productivity data from Babinda and Mourilyan Mills (prior to 2010) |
| TUL | Tully Central Mill | Not Participating |
| HBT | Wilmar – Herbert (Victoria, Macknade) | Spatial and productivity data from 2005 to 2017 |
| BKN | Wilmar – Burdekin (Kalamia, Inkerman, Pioneer, Invicta) | Spatial and productivity data from 2005 to 2017 |
| PPN | Wilmar – Proserpine | Spatial data from 2013 to 2017, awaiting supply of further spatial data, productivity data supplied 2005 to 2017. |
| MKY | Mackay Sugar (Racecourse, Farleigh, Marian, Pleystowe) | Spatial data from 2001 to 2018, productivity data from 2005 to 2016. |
| PCK | Wilmar – Plane Creek | Spatial and productivity data from 2005 to 2017 |
| ICM | Isis Central Mill | Spatial and productivity data from 2001 to 2017 |
| BDY | Bundaberg Sugar – (Millaquin, Fairymead, Bingera) | Not Participating |
| MSF | MSF Sugar (Maryborough) | Spatial and productivity data from 2002 to 2017, productivity data from 2008 to 2017. |
| RKY | Rocky Point | Spatial Data since 2009, awaiting supply of productivity data |
| NSW | Sunshine Sugar (Condong, Broadwater, Harwood) | Spatial and productivity data from 2001 to 2017 |

## 6.2. Data Collated

The data received from the data entities provided a substantial data set of productivity data for the Australian sugar industry. A total of 1,877,083 data records were collated from the 13 data entities, with each record representing a paddock for a particular year. This represented 6,163,269 hectares of data from the industry (Figure 3).

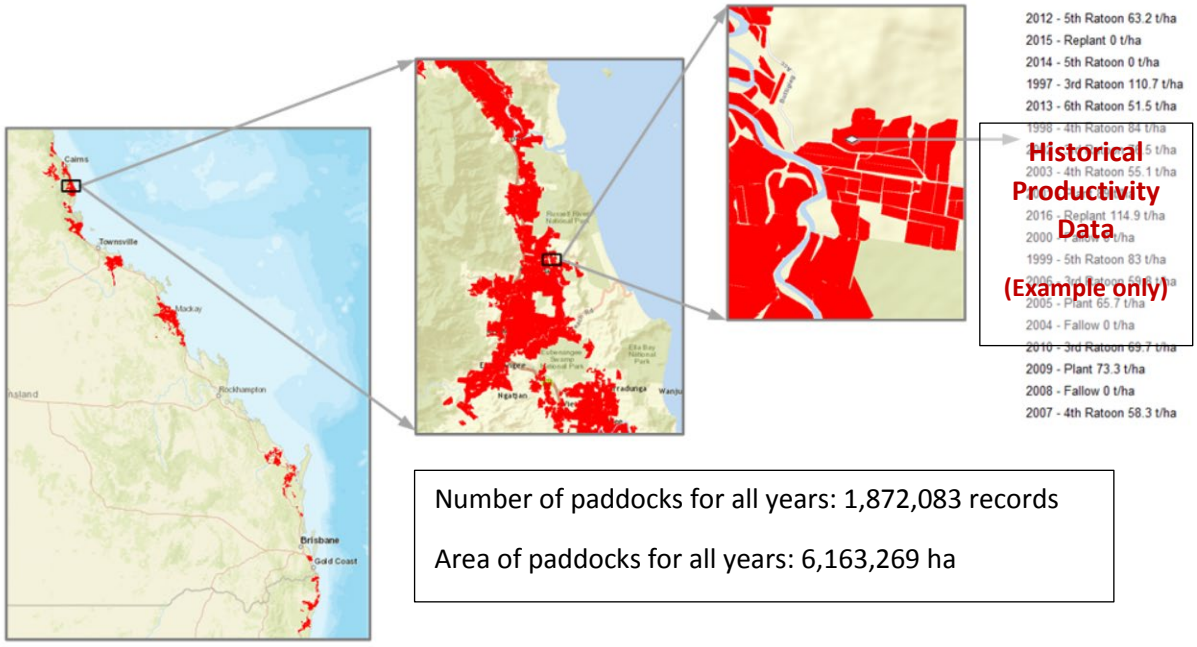



**Figure 3. Graphic showing the extent of the total coverage of data collated in the project, showing more detail of paddocks with higher zoom levels, and the multiple years of data accessible by location (Note: data displayed is an example only, and is not sourced from the location shown).**

The number and area of paddocks collated each season is presented in Figure 4. Most data entities were able to supply spatial data for the last 10 years, with some others providing data from previous seasons as far back as 1997. There was an anomaly in the data for 2012 due to data not being provided for Mackay Sugar, which may be remedied with Mackay Sugar after the project completion. Some productivity data was not available due to mill closures such as Mourilyan and Babinda where, although the spatial data had been archived by a staff member, the productivity data was not available for those mills centrally.

The data came in many formats, including MapInfo Tab (Native and Access formats), Microsoft SQL Server Spatial database, and ESRI shape files for the spatial data, and typically Excel spreadsheets for the non-spatial data such as the block or production tables. The codes used for variety and class, as well as the structure of how paddocks were identified, changed over time for some mill entities. This caused issues when matching production data to the spatial data in some cases, but these issues were worked around.

The collation of this data was timely, as much of this data was sourced from data that had been collated by individuals over the years, and may have been difficult to source if those individuals left the industry. Further, data was stored in formats that required using an older version of Windows operating system, as the Office 365 suite did not support the drivers required to open the Access based spatial data that was provided.

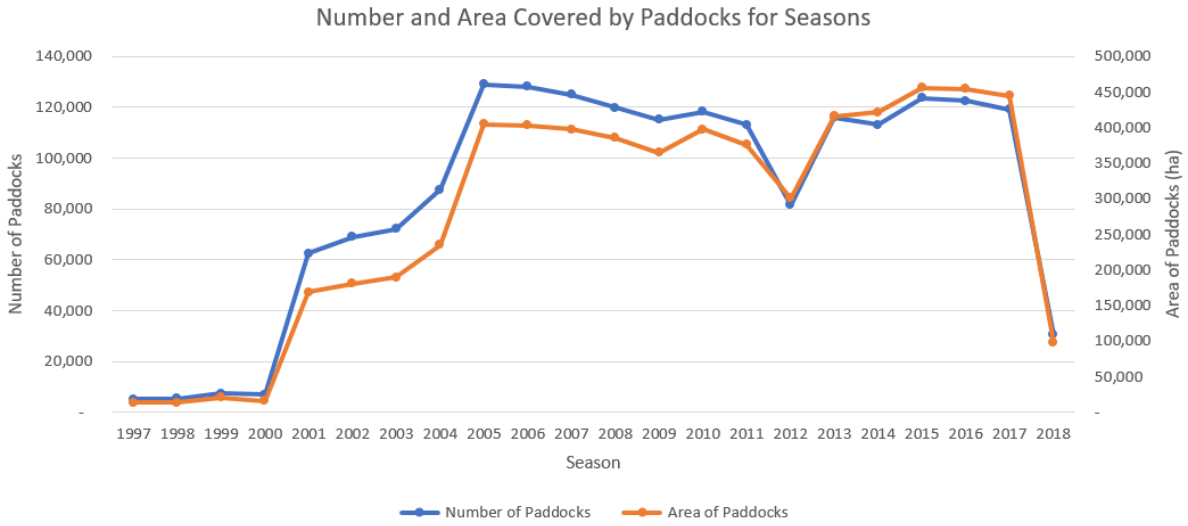



**Figure 4. Number and area of paddocks collated for seasons in the Sugar Data Hub Database.**

## 6.3. Data Quality

The data from each organisation and season was analysed for its internal accuracy. The main criteria for the analysis was the consistency of the data between the spatial data and the mill production data. Commonly, milling organisations record production data in a separate system to the spatial data, and at the start of the season after the seasons mapping is completed, the spatial data is loaded into the milling organisations receivals and production recording system. Any changes to the farm/ paddock data after the start of the season are often not reflected back into the spatial data, and some discrepancies between the spatial data and the production data will exist as a result.

This analysis highlighted the seasons for each data entity that data was missing or mismatched, and was used during the project as a quality assurance check that the data had been processed correctly. The categories used in this analysis are provided in Appendix 2, Table A2.1. The breakdown of the areas in the database that fall into each of these categories is provided in Figure 5.

The data presented in Figure 5 is for the whole dataset across all milling entities. The data quality for each milling entity is discussed in a separate sub-report provided to SRA as Appendix 3, which has an equivalent graph for each milling entity. This sub-report also outlines some parameters that influence the confidence of the accuracy of the production data, such as:

1. Degree to which harvester tracking was used to identify daily harvested areas
2. Rigour in validating consignment accuracy by cane supply officers
3. General indication of the accuracy of consignment from those who collect it.

This review is deemed confidential and will not be made available publicly.

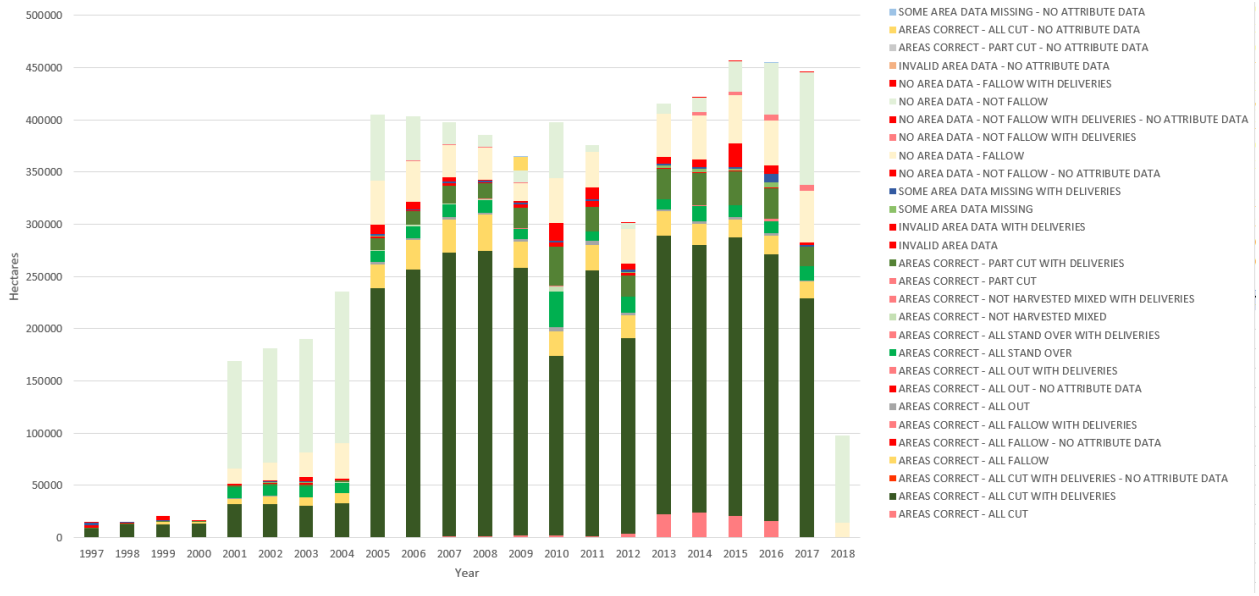



**Figure 5. Graph showing data quality for each year from all data entities. A breakdown of this data for each data entity is provided to SRA, but is considered confidential, and not part of the public report.**

## 6.4. Analyses

This project was not intended to perform analyses of the data, but examples of how this data could be used include:

1. Analysis of production across agro-climatic zones, which may include multiple milling areas or partial mill areas (Figure 6).
2. Retrieve historical production regardless of what paddocks were called in the past (example shown in Figure 3).
3. Analyse impact of time and weather conditions between seasons from historical data and spatial relationship of paddocks.

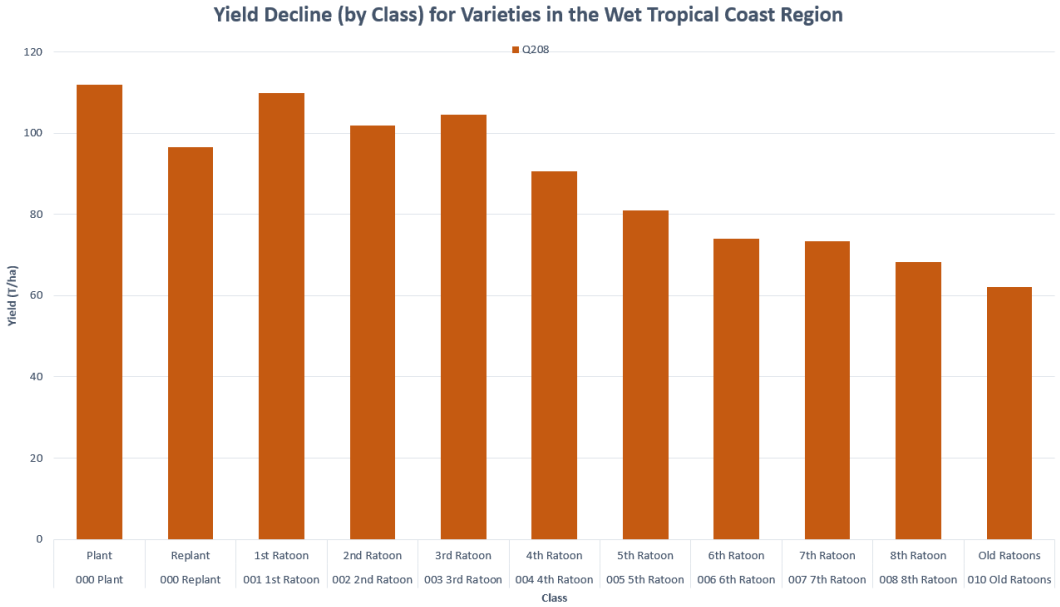



**Figure 6. Yield associated with each cane class for Q208Φ variety for the 2015 to 2017 seasons, for paddocks that were located in the Wet Tropical Coast agro-climatic zone.**

## 6.5. Further Work

The Sugar Data Hub project has created a substantial database of historical spatial production data up until the 2017 season.

There are some data missing in some cases as mill staff could not organise to supply the data in the time frame of the project. Future work may include filling in missing or improved data on a commercial basis or as part of the on-going relationship with those data owners.

Alternatively SRA may continue to facilitate this data collation effort at regular intervals, either speculatively or as part of defined projects that use this data for research.

# 7. RECOMMENDATIONS FOR FURTHER R,D&A

Engaging with industry to collate data as was done in this project can be challenging. Factors that influence the ease of this process include:

1. Timing – Sugar mills have an annual cycle of tasks that cause dramatic changes on the workload of different staff. Any data requests immediately prior to the start of the season or during the season may be met with resistance due to the high workload managing the harvest and cane supply. Further, immediately after the harvest season many staff will take time off for an extended period until mid-January.
2. Responsibility – Typically the staff who are responsible for providing the data are not directly involved in using this data for productivity analysis. This may create difficulties in gaining priorities for the extraction of this data depending on the structure of the organisation.
3. Permission for 3rd party access to data – The ability to supply data to a 3rd party can be influenced by the cane supply agreements that are in place with the growers, and in some cases, supply of data was held up due to the number of parties that need to provide approval.
4. Prior relationships – Typically, provision of data is simpler where the group requesting the data has a prior relationship with the data owner, particularly if that the request is for additional data for an existing project.

Industry organisations such as the Australian Sugar Milling Council, SRA and CANEGROWERS could facilitate this process by promoting discussion on protocols and practices that need to be addressed in the provision of data to researchers, and the responsibilities of those researchers in terms of protection of privacy.

While the Sugar Data Hub project has collated a lot of historical data for the industry, this data will become dated in time. Continued annual reviews to keep the data current would enable researchers to use this data into the future. This may be done on a commercial basis for each data owner directly, or coordinated for the industry by a central agency such as SRA. While the processes required to collate the data were scripted by Agtrix during the project to streamline the processing, other organisations could perform the same tasks using the translation tables and 3rd party datasets provided to SRA (soils and agro-ecological zones) and the data owners as part of this project.

# 8. PUBLICATIONS

Crossley R., A. Law and M. Anderson (2018) “Spatial Data Hub – A Valuable Resource for the Sugar Industry.” Proc. Australian Society of Sugar Cane Technologists, Mackay, 2018.

# 9. REFERENCES

Williams, J., Hook, R. & Hamblin, A. (2002) Agro-ecological regions of Australia. Methodology for their derivation and key issues in resource management (CSIRO Land and Water, Canberra). Verified on-line 24th May 2018. www.clw.csiro.au/publications/general2002/Agro-eco_report_nomap.pdf

# 10. APPENDICES

## 10.1. Appendix 1 Metadata Disclosure

**Table 1 Metadata disclosure 1**

| | |
|---|---|
| Data | Collated Industry Data – Data Hub |
| Stored Location | Agtrix – ROB-XPS, SQL Server Database |
| Access | Restricted Access, Microsoft Network Security, Password Protected |
| Contact | Robert Crossley |

**Table 2 Metadata disclosure 2**

| | |
|---|---|
| Data | Industry Supplied Data |
| Stored Location | Agtrix – Virtual Machine – Will be deleted after project completion |
| Access | Restricted, Microsoft Network Security, Password Protected |
| Contact | Robert Crossley |

**Table 3 Metadata disclosure 3**

| | |
|---|---|
| Data | Industry Translation Codes in Excel format |
| Stored Location | Spreadsheet file |
| Access | Restricted. |
| Contact | Peter Samson |

## 10.2. Appendix 2 Data Quality Description

**Table A2-1. Quality status categories evaluated for each record, and used to evaluate data quality and identify where attention is needed.**

| Data Quality Tag | Explanation | Area (ha) | % Total |
|---|---|---|---|
| AREAS CORRECT - ALL CUT | Mill area data matches the spatial area, the area identified as harvested was more than 98% of the paddock and there was production recorded against it. | 19,901 | 0.3 |
| AREAS CORRECT - ALL CUT WITH DELIVERIES | Mill area data matches the spatial area, the area identified as harvested was more than 98% of the paddock and there was production recorded against it. | 2,760,110 | 48.2 |
| AREAS CORRECT - ALL CUT WITH DELIVERIES - NO ATTRIBUTE DATA | Mill area data matches the spatial area, the area identified as harvested was more than 98% of the paddock and there was production recorded against it. Variety and class attributes were not available. | 11 | 0.0 |
| AREAS CORRECT - ALL FALLOW | Mill area data matches the spatial area, and all of the paddock was fallow (more than 98% of the paddock). | 292,071 | 5.1 |
| AREAS CORRECT - ALL FALLOW - NO ATTRIBUTE DATA | Mill area data matches the spatial area, and all of the paddock was fallow (more than 98% of the paddock). Variety and class attributes were not available. | 15 | 0.0 |
| AREAS CORRECT - ALL FALLOW WITH DELIVERIES | Mill area data matches the spatial area, and all of the paddock was fallow (more than 98% of the paddock). Deliveries were recorded against the paddock. | 40 | 0.0 |
| AREAS CORRECT - ALL OUT | Mill area data matches the spatial area, and all of the paddock was either taken for plant, ploughed out or for undifferentiated (more than 98% of the paddock). | 23,116 | 0.4 |
| AREAS CORRECT - ALL OUT - NO ATTRIBUTE DATA | Mill area data matches the spatial area, and all of the paddock was either taken for plant, ploughed out or for undifferentiated (more than 98% of the paddock). Variety and class attributes were not available. | 5 | 0.0 |
| AREAS CORRECT - ALL OUT WITH DELIVERIES | Mill area data matches the spatial area, and all of the paddock was either taken for plant, ploughed out or for undifferentiated (more than 98% of the paddock). Variety and class attributes were not available. Deliveries were recorded against the paddock. | 222 | 0.0 |
| AREAS CORRECT - ALL STAND OVER | Mill area data matches the spatial area, and all of the paddock was stood over (more than 98% of the paddock). | 215,238 | 3.8 |
| AREAS CORRECT - ALL STAND OVER WITH DELIVERIES | Mill area data matches the spatial area, and all of the paddock was stood over (more than 98% of the paddock). Deliveries were recorded against the paddock. | 61 | 0.0 |
| AREAS CORRECT - NOT HARVESTED MIXED | Mill area data matches the spatial area, and there is a mixture of outcomes for the cane in that paddock – ploughed in, fallow, plant and standover. | 10,457 | 0.2 |
| AREAS CORRECT - NOT HARVESTED MIXED WITH DELIVERIES | Mill area data matches the spatial area, and there is a mixture of outcomes for the cane in that paddock – ploughed in, fallow, plant and standover. Deliveries were recorded against the paddock. | 20 | 0.0 |
| AREAS CORRECT - PART CUT | Mill area data matches the spatial area, the area identified as harvested was more than 98% of the paddock and there was no production recorded against it. | 1,778 | 0.0 |
| AREAS CORRECT - PART CUT WITH DELIVERIES | Mill area data matches the spatial area, there was area identified as harvested in the paddock, but was less than 98% of the paddock and there was production recorded against it. | 451,254 | 7.9 |

| Data Quality Tag | Explanation | Area (ha) | % Total |
|---|---|---|---|
| AREAS CORRECT - PART CUT WITH DELIVERIES - NO ATTRIBUTE DATA | Mill area data matches the spatial area, there was area identified as harvested in the paddock, but was less than 98% of the paddock and there was production recorded against it. Variety and class attributes were not available. | 1 | 0.0 |
| INVALID AREA DATA | Mill area data on fate of cane is more than 102% of the spatial area – probably paddock changed during season and the area in the mill database was updated and not the mapped area. | 5,312 | 0.1 |
| INVALID AREA DATA WITH DELIVERIES | Mill area data on fate of cane is more than 102% of the spatial area – probably paddock changed during season and the area in the mill database was updated and not the mapped area. Deliveries were recorded against the paddock. | 869,688 | 15.2 |
| NO AREA DATA - FALLOW | No area data on fate of cane in paddock (cut, plant, standover, plough out) supplied from Mill | 447,077 | 7.8 |
| NO AREA DATA - FALLOW WITH DELIVERIES | No area data on fate of cane in paddock (cut, plant, standover, plough out) supplied from Mill | 432 | 0.0 |
| NO AREA DATA - NOT FALLOW | No area data on fate of cane in paddock (cut, plant, standover, plough out) supplied from Mill | 429,219 | 7.5 |
| NO AREA DATA - NOT FALLOW - NO ATTRIBUTE DATA | No area data on fate of cane in paddock (cut, plant, standover, plough out) supplied from Mill | 95,671 | 1.7 |
| NO AREA DATA - NOT FALLOW WITH DELIVERIES | No area data on fate of cane in paddock (cut, plant, standover, plough out) supplied from Mill | 20,244 | 0.4 |
| NO AREA DATA - NOT FALLOW WITH DELIVERIES - NO ATTRIBUTE DATA | No area data on fate of cane in paddock (cut, plant, standover, plough out) supplied from Mill | 48 | 0.0 |
| SOME AREA DATA MISSING | Mill area data on fate of cane is less than 98% of the spatial area – either area missing or paddock changed during season and the area in the mill database was less than the mapped area. | 5,849 | 0.1 |
| SOME AREA DATA MISSING WITH DELIVERIES | Mill area data on fate of cane is less than 98% of the spatial area – either area missing or paddock changed during season and the area in the mill database was less than the mapped area. | 75,204 | 1.3 |

**Table A2-2 Listing of the fields of the data provided to each organisation.**

| Field | Description |
|---|---|
| C_ORG_CODE | Organisation Code |
| C_FM_ORG_CODE | Organisation Code used in FarmMap |
| C_SEASON_ID | Harvest Season |
| C_LINKCODE | Structured Paddock Identifier |
| C_FARMCODE | Structured Farm Identifier |
| F_AREA_HA_FM | Area from Mapping Data |
| C_ORG_VARIETY_CODE | Preferred Variety Code (Used by Organisation) |
| C_ORG_VARIETY_DESC | Preferred Variety Description (Used by Organisation) |
| C_ORG_CLASS_CODE | Preferred Class Code (Used by Organisation) |
| C_ORG_CLASS_DESC | Preferred Class Description (Used by Organisation) |
| C_ORG_AGE_CODE | Preferred Age Code (Used by Organisation) |
| C_ORG_AGE_DESC | Preferred Age Description (Used by Organisation) |
| C_ORG_PLANT_CODE | Preferred Plant Code (Used by Organisation) |
| C_ORG_PLANT_DESC | Preferred Plant Description (Used by Organisation) |
| C_IND_VARIETY_CODE | Industry Generic Variety Code |
| C_IND_VARIETY_DESC | Industry Generic Variety Description |
| C_IND_CLASS_CODE | Industry Generic Class Code |
| C_IND_CLASS_DESC | Industry Generic Class Description |
| C_IND_AGE_CODE | Industry Generic Age Code |
| C_IND_AGE_DESC | Industry Generic Age Description |
| C_IND_PLANT_CODE | Industry Generic Plant Code |
| C_IND_PLANT_DESC | Industry Generic Plant Description |
| C_DOM_SOIL_MAP_CODE | Dominant mapping code for paddock |
| I_DOM_PCNT_COVER | Percentage of the paddock that is covered by the dominant mapping unit |
| SOIL_CONCEPT | Description of the dominant soil in the mapping code |
| SOIL_DESC | A more general description of the major soil in the mapping unit |
| AGRO_ECOL_TYPE | A general agro ecological type based on Williams et al. 2002 |
| AREA_HA | Area of Paddock |
| AREA_CUT | Area nominated as cut |
| AREA_FAL | Area nominated as fallow |
| AREA_PLANT | Area nominated as being cut for plant material |
| AREA_PLOUT | Area nominated as being ploughed out |

| Field | Description |
|---|---|
| AREA_SOVER | Area nominated as standover |
| AREA_OUT_UNDIFF | Area nominated as not being harvested for production, but no further information was available |
| FIRST_HARVEST_DATE | Date that the first harvest activity or delivery was recorded |
| LAST_HARVEST_DATE | Date that the last harvest activity or delivery was recorded |
| TOTAL_TONS | Total tonnes delivered |
| TOTAL_TONS_SUGAR | Total Tonnes of sugar produced as measured by CCS (Tonnes cane * CCS) |
| TOTAL_FIBRE | Total Tonnes of fibre produced as measured by CCS (Tonnes cane * fibre%) |
| C_YEAR_HARVEST_STATUS | Comment of validation status for paddock. |
| I_PCNT_AREA_DELIVERED | Percentage of area that was harvested for production |
| F_CALC_YIELD | Calculated yield (Total tonnes/ area cut) |
| I_DAYS_SINCE_LAST_HARVEST | Days since paddock was last harvested |
| MI_STYLE | Style to draw shape in GIS |
| MI_PRINX | Unique Integer ID |
| SP_GEOMETRY | Spatial object |

## 10.3. Appendix 3 Data Quality For Milling Entities - CONFIDENTIAL