**Geoscience Australia**

**Building Exposure Development – Enhancing address filtering for improved building exposure information**

**Intern Project – Winter 2015**

**Xavier Butcher**

Contents

[Keywords: 2](#_Toc423617137)

[Introduction 2](#_Toc423617138)

[Materials 3](#_Toc423617139)

[Method 3](#_Toc423617140)

[Results 4](#_Toc423617141)

[Discussion 5](#_Toc423617142)

[Further Recommendations 8](#_Toc423617143)

[Conclusion 10](#_Toc423617144)

[Acknowledgements 10](#_Toc423617145)

## Keywords:

Address, Building, Built Environment, Cadastre, Classification, Confidence, Data, Exposure, Flat Number, Geoscience, Guidelines, Points, Rules.

## Introduction

The aim for this project was to improve the accuracy and quality of Geoscience Australia’s (GA) modelling of the built environment by enhancing business rules and processes used to determine the number of dwellings and building location across the ACT, with a further check in Tasmania. This involves analysing the Geocoded National Address File (G-NAF) to identify the combinations of address attributes occurring within a single cadastre parcel, and determining business rules to apply to get the best representation of the built environment.

Exposure information is integral for understanding the buildings and people at risk during natural hazard events. This project involves analysing the G-NAF to enhance the validation and integration process, improving National Exposure Information System (NEXIS) and supporting GA’s role in ensuring Australia’s Community safety.

At the end of this project, pending approval, any modifications or additions to business rules will be developed and integrated into the NEXIS application.

## Materials

* G-NAF Geocoding Metadata Kit 2015
* ESRI ArcGIS
* Google Earth
* Google Maps

## Method

1. Create Geodatabase
2. Locate and import relevant data sets (including ACT\_Building, ACT\_GNAF, ACT\_Child, ACT\_Cadastre and NEXIS\_LAND\_USE\_ACT)
3. Understand all data fields and metadata
4. Select possible fields for further analysis (including Shape ID, Alias Principal, Confidence, Level Geocoded Code, Number First, Number First Suffix, Building Name, NEXIS CAD ID, Flat Number, G-NAF Street Confidence, Street Class Code, Street Name, Street Type and Reliability Code)
5. Run an Intersection Tool between ACT\_Cadastre and ACT\_GNAF. This creates a new layer that allows parcels of land with multiple address points to be seen more clearly
6. Using Google Earth/Maps, identify which of these parcels are incorrectly classified (i.e. have more or less buildings than what should be there).
7. Of those parcels, see if there is any correlation between certain fields that would affect the classification, and if so, create a set of rules that would allow for proper classification.
8. Create a frequency table to ensure all values covered, use selection queries to identify attributes
9. Import the Wind Regions Settlement data file, which adds a layer of symbology to distinguish between settlement types, including small towns, cities and rural areas. Use selection queries to sample areas from right around the state.
10. Import the NEXIS Land Use data file, which adds a layer of symbology to distinguish between land use types, such as commercial, industrial and residential.
11. See if those rules apply State (or country) wide
12. If so, assimilate into current rules, if not, discuss if it’s viable as a guideline

## Results

See attached spreadsheet for full results. An explanation of results will be shown here:

* **Sample size**:
  + First 72 CAD ID’s alphabetically,
  + Highest 16 frequency occurrences,
  + Further check in Tasmania by sampling via location (23 entries)
  + Total 111
* **Sample area**:
  + In depth ACT, check in Tasmania
  + 208,778 possible individual address points in ACT, 310,993 in TAS
* **Fields analysed**:
  + NEXIS CAD ID (Unique cadastral parcel identifier),
  + Number of address points (per parcel),
  + Number of buildings (per parcel),
  + Suggested fields (for further analysis),
  + Suggested rules (showing suggested rules to fix errors in that particular parcel)
* **Frequency table fields**:
  + Frequency (Number of occurrences),
  + Alias Principal (Value “A” or “P”, defines whether an address exists or not),
  + Confidence (How many data sets an address point appears in. *Key*: -1 =0 data sets, 0=1 data sets, 1=2 data sets, 2=3 (all) data sets),
  + Level Geocoded Code (Range from 0-7. 6=Geocode for locality and street, but not parcel. 7= Geocode for locality, street and parcel),
  + GNAF Street Confidence (Confidence of the existence of the street (same key as confidence above))
  + Reliability Code (Precision of geocode rated 1-6. Lower number is more accurate)
  + Description (Explanation of what the situation is regarding that particular parcel)
* **Other important field definitions**
  + Street Class Code (Binary existence of street. “C” = confirmed, “U” = unconfirmed),
  + Flat Number (Unit Number)
  + Number First (Street Number)
  + Building name (Usually found in apartment blocks; e.g. Endeavour Gardens)
  + Number First Suffix (Values found after street number; e.g. 36A)

## Discussion

I would firstly like to preface this discussion of results by noting that I have had only a limited time to not only develop my intermediate level ARCGIS skills, but also to understand the G-NAF metadata before I even started analysing the actual data. The 2 weeks I have been here have also been quite hectic with regards to workshops, seminars and meetings which I felt were in my best interest to attend in place of working on this project. Had I had more time, I would have developed a much more thorough sample size not only across the ACT, but nationwide as well. However I cannot determine if my rules are a fair representation on a broader nationwide scale. This is due to different states having their own variations with respect to the data collection and presentation. I chose to spend most of my time chasing down the rabbit hole of 1 rule, which unfortunately in the end is not feasible. I hope that whilst I can produce no major ground-breaking rule; that my efforts have helped in the sense that someone else won’t chase down the same hole I did.

The entire aim of this project is to define rules which aid in the correct conversion of address points to building points. The current rules define if a building is valid or not, so if a point breaks these rules, it is removed:

* Alias Principal starts with “A”
* Confidence is less than 0
* Has no NEXIS CAD PID
* Has no NEXIS Use II
* NEXIS Block-size is below defined threshold
* Local Footprint is below defined threshold
* Has a duplicate Locality ID

Most of the misclassification I encountered in my results occurred when multiple address points were assigned to a parcel where there are a fewer number of buildings. As such, most of the rules I created are a direct response to this. There is currently a rule in place however, that dictates the filtering of multiple address points based on the confidence. The current rule is best described in a table:

|  |  |  |  |
| --- | --- | --- | --- |
| **0** | **1** | **2** | **Action** |
| Yes | No | No | Keep all addresses |
| No | Yes | No | Keep all addresses |
| No | No | Yes | Keep all addresses |
| Yes | Yes | No | Keep confidence 1, remove confidence 0 |
| Yes | No | Yes | Keep confidence 2, remove confidence 0 |
| No | Yes | Yes | Keep confidence 2, remove confidence 1 |
| Yes | Yes | Yes | Keep confidence 2, remove confidence 1 and 0 |

I have decided that this rule is a fair way to classify data; hence I will be keeping this rule in my final recommendations. However the street class code rule I introduce should run prior to this one.

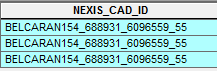
Of all the fields I analysed, the most important fields for general accuracy were Confidence and Alias Principal. Confidence values less than 0 removed approximately 1.3% of data. Alias values of “A” removed approximately 1% of data. As such, I have decided to keep both of the above rules, as they significantly cull the amount of address points that are just completely wrong.

The first rule I will add (for general accuracy purposes) is the Street Class Code, such that any value with an ‘Unconfirmed’ street is disregarded. Approximately 0.5% of all address points were unconfirmed. Based on my results, I posit that this rule will further help to reduce the number of address points that are completely wrong.

This next recommendation is not so much of a rule, but more of a guideline. Based on my results, a number of incorrectly classified points could be removed by using the GNAF Street Confidence <0 rule. However, it simply removed such a large portion (approximately 20%) of data. Upon further investigation I found that there were simply too many correctly classified addresses that were being removed. I modified the rule to be where: Confidence <1 and GNAF Street Confidence <0, and even then it was removing approximately 7% of the data, a healthy proportion of which was correctly classified data. Hence, I think would be unwise to incorporate this rule in my final product, however I do think it is an important field to take into consideration on a case by case basis. Again, had I had more time, I would have investigated this rule to see exactly how accurate it was.

By far the most common error I found in my results was when there were more address points than buildings on the land parcel. This requires the use of Google Earth/Maps to identify how many flats there are on the property. A significant portion of these occurred when there were 2 dwellings on the property, but there was 1 extra point. Often, the best way to define which 2 of the 3 points were correct was to examine the Flat Number or Number First Suffix. I found that the 2 correct points had values in either of these columns, and the surplus point had a null value in these fields. Hence, in my results it can be seen that I often suggested to ‘Remove Flat Null’ or ‘Remove Number First Suffix Null’. However, removing either of these values by itself doesn’t take into consideration properties where there is only 1 dwelling on it.

Where multiple address points have identical NEXIS CAD ID’s, and the number of buildings is less than the number of address points, and the address points have values in either the Flat Number or Number First Suffix, remove the address point with the ‘Null’ value in this field. This rule I have found works best in residential areas, so intersecting the NEXIS Land Use file will make for better results.



🡪



Unfortunately, this rule was the one that turned out to be unfeasible. It was quite a long process to automate my rule. It involved many intersects, joins, frequencies, field calculations and python scripts. My rule kept becoming more and more obscure, and in the end the thing that finally killed it was just how terrible the G-NAF data surrounding Flat Numbers and Number Suffixes was. Whilst my rule did get rid of a respectable portion of bad addresses, there were simply too many cases where the data was just too poorly formatted to be practical. So many subdivisions used a combination of both the Flat Numbers and Number Suffixes, which logically doesn’t make sense. For instance, one subdivision had 2 dwellings on it; one unit had 1 address point with a Flat Number 1 and a Number Suffix A. The other unit had 3 address points, none of which had a Flat Number, and each point had Suffix A, B or Null. This was just getting too complicated, so I decided to stop going ahead with this rule.

This last recommendation is again more of a guideline. In some circumstances where there are large building complexes (15+ dwellings), there were some rare cases where incorrectly classified data could be removed based on whether or not the address point had the correct building name. For instance BELCARAN1636\_688772\_6096765\_55 had 10 excess address points than dwellings; these were able to be removed by eliminating the address points with a null field in the Building Name. This guideline works in a similar way to the Flat Number Null rule, in that there has to be more address points than the number of dwellings. However, it is much less concrete, and should be used as a guideline only in a case by case basis.

To summarise, this is my final list of rules I will recommend to be implemented as to remove incorrect data:

* Alias Principal starts with “A”
* Confidence is less than 0
* Has no NEXIS CAD PID
* Has no NEXIS Use II
* NEXIS Block-size is below defined threshold
* Local Footprint is below defined threshold
* Has a duplicate Locality ID
* Street Class Code is “U”

With further guidelines (on a case by case basis) relating to:

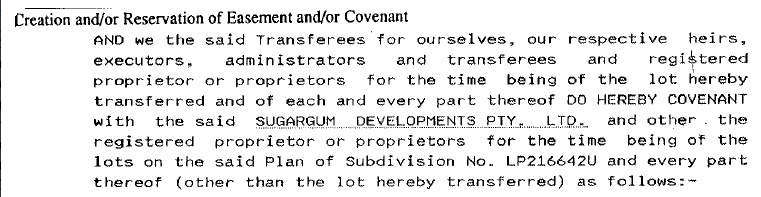
* In special cases outlined above, Flat Number or Number First Suffix is ‘Null’
* GNAF Street Confidence is less than 0
* Building Name is ‘Null’

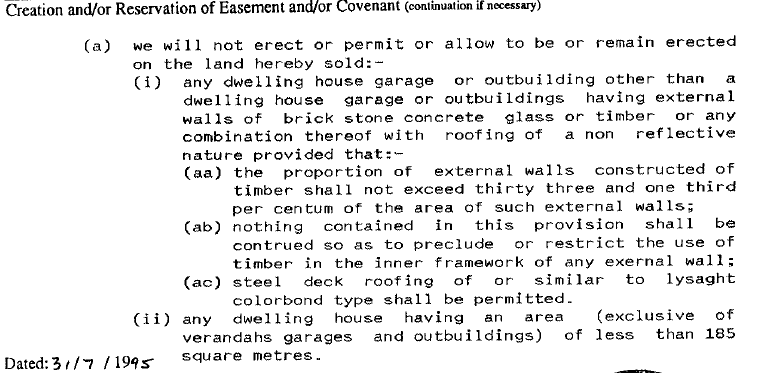
## Further Recommendations

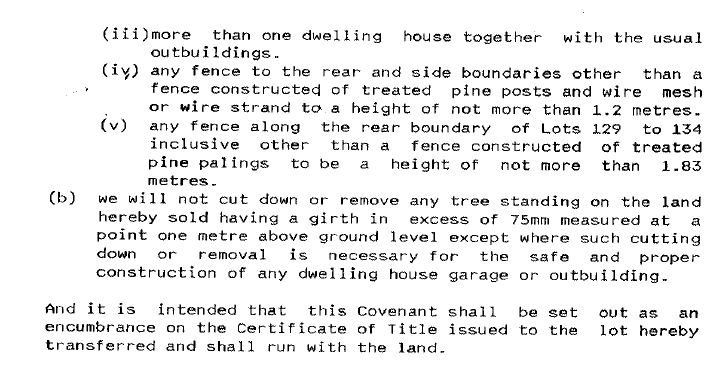
As I am studying a Surveying degree, my recommendations and rules demonstrate a cadastral bias, as that is what I am more knowledgeable about. Hence, it was my idea that another filter that could be added to enhance conversion from address points to building points might be restrictions placed on land title certificates. I cannot speak about ACT cadastral law, but in my home state of Victoria, each parcel of land has a certificate of title registered with Land Victoria. On this certificate, restrictions (known as Covenants) can be placed by a land owner relating to characteristics of the building on the title. Covenants are often placed when a parent parcel is subdivided into (usually 2) smaller child parcels. Covenants restrict such things as height or storey limits on a property, as well as the number of dwellings that can be built on the child parcel. In the image below (specifically point iii), a covenant is restricting the number of dwellings on the property. This is to restrict the building of an apartment block or secondary dwelling on the child parcel.

The application of this information is such that if a particular parcel of land had a covenant restricting the number of dwellings to be placed, and the GNAF had record of numerous address points on that block, then it should be discarded. It could be argued that the surveyor or builder wilfully (or otherwise) ignored this covenant and proceeded to build anyway so should be included in the GNAF. However, checking title restrictions is such a fundamental process in surveying and building that this should not ever be the case.

Unfortunately, no datasets currently exists for this information, and the amount of time and money required to produce this would far outweigh the benefits of having it. Perhaps this could be a potential solution in the future, but at this stage, more traditional datasets are far more effective.







## Conclusion

This project has been very helpful in developing my intermediate skills on ARCGIS and helping me understand what is it the Built Environment and Exposure Section do. I have seen that some of the skills learnt on ARCGIS are very relevant in the workplace. I hope that my work is useful to GA in some way, and that it either gets applied or at least reviewed. I am proud of the quality and quantity of work I have achieved this fortnight.

If I was to criticize this project in any way, it would be to note the routine nature of the methodology used to attain a respectable sample size. It is indeed challenging and time consuming.

## Acknowledgements

I would especially like to thank Lauren Power and Brad Cook for their continuous input and guidance on this project. Their help was much appreciated, and I’m sure that they are looking forward to not having to answer my rudimentary questions or verifying my attribute search queries. Their guidance also extended beyond the workplace, often giving me advice on things to see and do in Canberra. I would also like to thank Gerard Stewart for lending me his wealth of experience. Lastly of course, I need to thank Section Leaders, Branch Heads and Recruitment for allowing me to undertake this internship.