Imputation of Dwelling Occupancy for Census 2021

Australian Bureau of Statistics
Informing Australia's important decisions

Background – Census Research

- Census Futures is leading research to…
  - Improve Census data quality
  - Reduce the cost and burden of Census
  - Add new information to the Census

- Ably assisted by Methodology Futures
  - Future-focussed methods development
  - Cross cutting
  - Machine learning a focus

Background – Census 2016

- PES approx. 44% of non-responding dwellings incorrectly imputed as occupied (PES) – about 500,000 persons
- Hasn't affected population estimates because PES is used to adjust for over and undercount at broad levels
- But some small areas, particularly those with secure apartment buildings, have higher counts than they should.

Background - Census 2016

- Persons are imputed to non-responding (NR) *occupied* dwellings

- Occupancy determination depends on intelligence from Census field staff -> a tendency to determine as ‘occupied’ if little information to go on

- -> too many ‘occupied’ NR dwgs -> over-imputation of persons

- Solution for 2021: better determination for occupancy status of NR dwgs

➡ Machine Learning model plus admin data

Data Sources

- OUTCOME:
- Census 2016 – provides occupancy label

- INPUTS:
- Government data – Taxation, Social services, Medicare
- Other – electricity
- Previous Census data items at small area level
- Geography – remoteness

- ABS Data Linkage centre de-identifies the data as the first step
- Census Futures use de-identified data to create dwelling level integrated dataset for analytical purpose

Census Data

- Occupied on Census night
- Unoccupied on Census night
  - Vacant
  - Away
- Admin data is good for predicting *vacant* dwellings

- Responding ‘occupied’ dwellings
- Non-responding dwellings - labelled ‘occupied’ or ‘unoccupied’ based on *strong* field intelligence
- Non-responding dwellings – labelled ‘occupied’ or ‘unoccupied’ – based on *weak* field intelligence

Field Information

1. Strong Unoccupied Field Information

   eg: *No, advised by resident* or *No, empty dwelling, no furniture*

2. Strong Occupied Field Information

   eg: *Yes, advised by resident* or Yes*, advised by neighbour*

3. Weak Field Information

   eg: *No, looks unoccupied* or Yes*, looks occupied*

Machine Learning Model

- Build (train and test) a supervised ML model on 2016 linked Census-Admin data – predicting occupied/unoccupied

- Use non-responding dwellings with *strong* information only.

- Predict/impute occupancy using 2016 model for 2021 non-responding dwellings with *weak* information

- Need to generalise from (i) 2016 to 2021, (ii) strong NR to weak NR

Machine Learning Model - Details

- XGBoost (based on prediction performance)
  - Logistic regression, CART, random forests
  - Using R
- Help from QUT
- Separate state models
- Feature selection
  - throw them all in
  - seeing which datasets we can do without due to cost
- Upsampling due to class imbalance
- Threshold choice

Results

- Sensitivity, specificity, ROC, AUC
- Results vary by state

State	Sensitivity	Specificity	AUC
VIC (best)	0.783	0.710	0.849
NT (worst)	0.614	0.583	0.651

New Opportunities and Challenges

Australian Bureau of Statistics

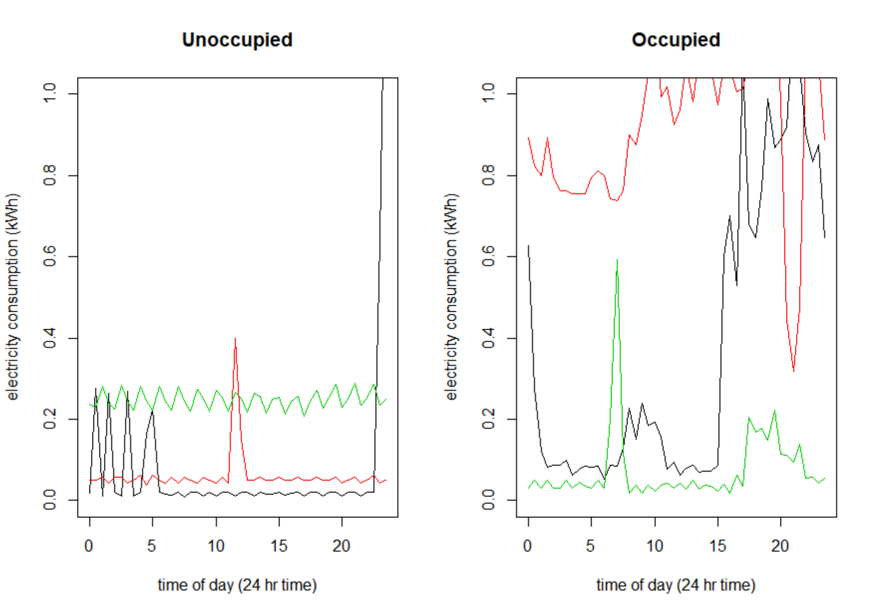
COULD HAVE CENSUS 2021 DATA TO TRAIN MODELS

- Have been training and validating on 2016 Census and admin data
- Assumes we can generalise to 2021
- Can’t use small area effects
- But might have 2021 Census data in time for imputation models
- Fewer generalisation assumptions needed
- *Could* use small area effects

New Opportunities and Challenges

Australian Bureau of Statistics

SMARTMETER ELECTRICITY DATA



Smartmeter Electricity Data

- Australian electricity supplied by a relatively small number of private providers
- Half hourly daily readings of dwelling electricity consumption in KW/h
- Coverage is most of Victoria currently – should increase rapidly over the next few years
- Powerful predictor of both ‘away’ and ‘vacant’ types of unoccupied
- Indicates some unreliability of Census occupancy label
- Potential to use by itself to determine occupancy status
- *Use of smartmeter data is research only at present and it’s not clear how/if we’ll be able to use it.*
- *Census Admin Data Privacy Impact Assessment is currently underway - both final PIA report and the ABS response will be published on the ABS website very soon. The outcomes from this PIA will influence the decision about how electricity data (in particular) will be used for Census 2021.*

Lessons learnt

- XGBoost requires careful parameter tuning
- Help from external experts is valuable
- Data access for external experts is a pain
- Data (reliable labels, good predictors) is more important than model/algorithm
- Use appropriate measures to compare models (e.g. AUC)

Conclusions and Further Work

- Fuller exploration of the value of smartmeter data
  - Help improve Census occupancy label
  - Occupancy prediction based on smartmeter data only

- Fit XGBoost models which include small area effects