Commute data methodology improvements (IDI_20201020 edition)

Richard Fabling

Independent Researcher

June 2021

Project funded by Waka Kotahi NZ Transport Agency. These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the IDI please visit https://www.stats.govt.nz/integrated-data

The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Outline

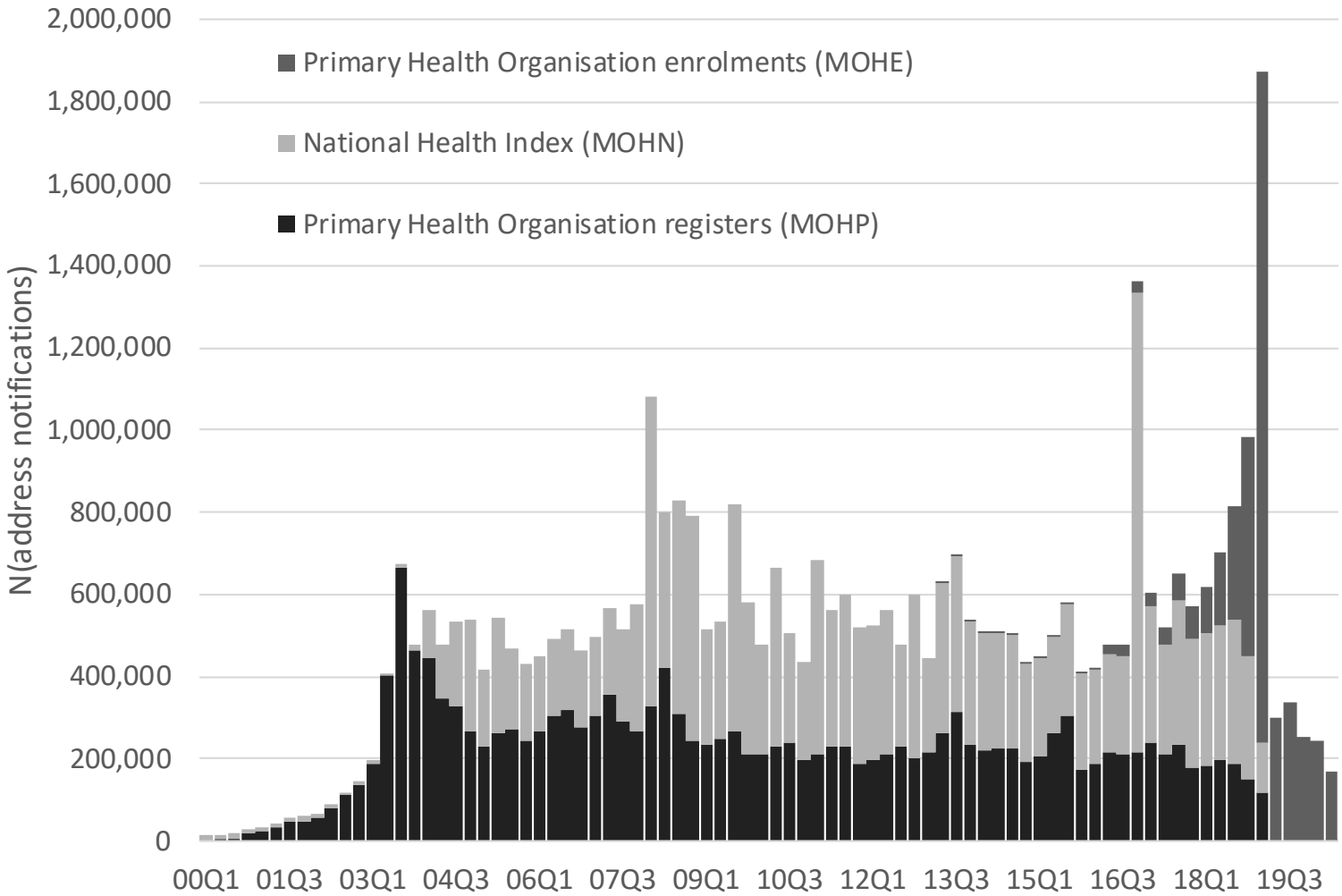
- Fabling & Maré (2020) develop a methodology for using the IDI to measure commute distance of NZ employees
- These slides explain changes made to that methodology for the 20201020 (and subsequent) IDI instance
  - Incorporating new residential address sources
  - Pooling NZTA addresses from across IDI instances
  - Adapting the methodology for a second Census (CEN'18)
  - Revising the weighting for missing commutes
  - Adjusting distance for within-meshblock (MB) commutes
  - Revising IDI Sandpit table contents to manage data size
- These notes assume familiarity with the Fabling & Maré (2020) methodology and are an addendum to that paper

Fabling & Maré (2020) "Measuring commute patterns over time: Using administrative data to identify where employees live and work," Motu Working Papers 20-05

New address sources

- Department of Internal Affairs: Births (DIAB); Civil Unions (DIAC); Deaths (DIAD); & Marriages (DIAM)
- Housing New Zealand Tenancies (HNZT)
- Ministry of Health PHO enrolments (MOHE)
- MOHE is the largest new address source, replacing the previously available health series (MOHN, MOHP)
- Census comparison implies that the DIA & HNZ address sources are all high quality (match rate $\geq$80%), and are assigned tier 1 status if coded to *x*-*y*
    - The MOHE series largely post-dates CEN'18, restricting our ability to formally test series quality. We assume that address quality is similar to the health series that MOHE replaces, and continue to pool health addresses into one series (MOH)

Ministry of Health residential address series



All residential addresses that are coded to *x*-*y*, restricted to individuals who are ever employed from 2005 onwards (ie, the population of interest).

Comparison of admin data to Census residential addresses

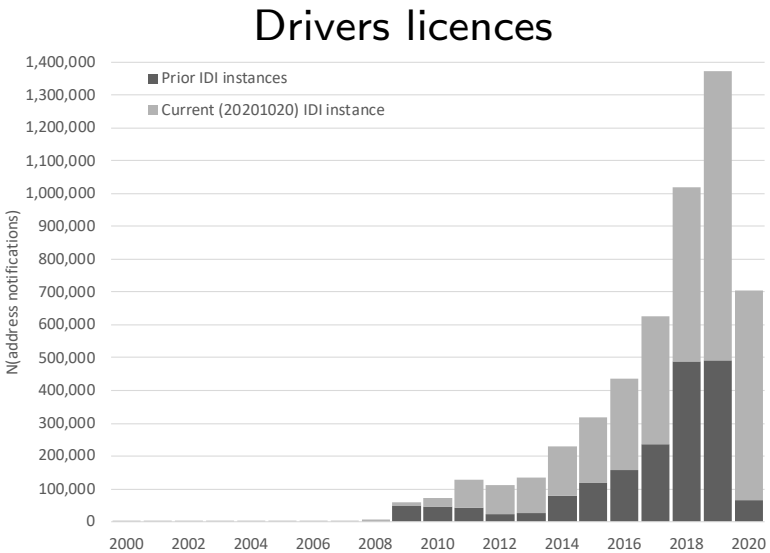
	N(addresses)			Census match rate		Change in
	Raw	Tier 1	p(tier 1)	2013	2018	match rate
ACCC	21,220,344	4,381,104	0.206	0.772	0.882	0.110
CEN	5,448,222	5,438,766	0.998			
DIAB	1,987,386	1,548,015	0.779	0.880	0.893	0.014
DIAC	5,148	5,073	0.985	0.824	0.800	-0.024
DIAD	100,908	98,232	0.973	0.963	0.864	-0.099
DIAM	702,465	688,566	0.980	0.821	0.838	0.017
HNZT	75,276	70,806	0.941	0.905	0.910	0.005
IRDA	42,066,324	16,436,304	0.391	0.786	0.818	0.031
MOES	375,333	0	0.000	0.679	0.712	0.032
MOH	40,713,291	15,240,345	0.374	0.867	0.882	0.016
MSDP	1,519,014	0	0.000	0.463	0.734	0.271
MSDR	11,330,643	9,865,416	0.871	0.811	0.826	0.015
NZTD	5,222,913	4,125,792	0.790	0.922	0.906	-0.016
NZTM	2,894,058	2,721,852	0.940	0.948	0.926	-0.021

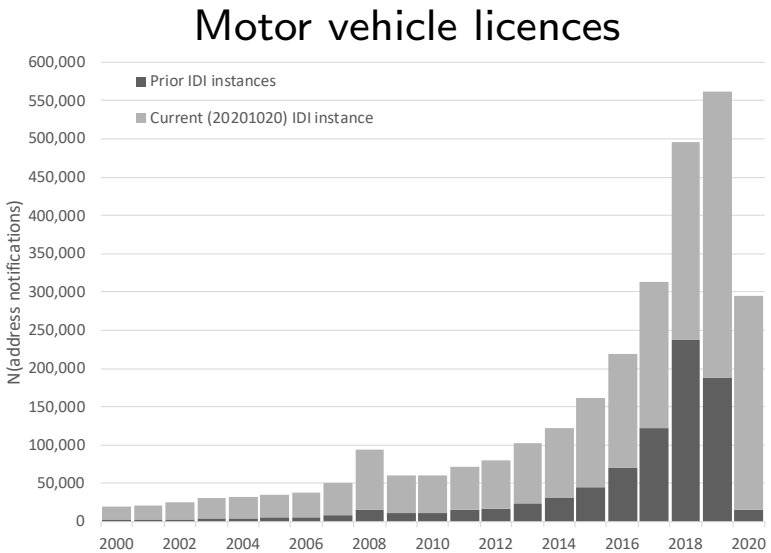
Matches include admin addresses in MBs adjacent to the Census MB. The match rate is conditional on having at least one admin address for the specific source in the year leading up to a Census (ie, prior year April to Census year March). The analysis is restricted to individuals employed on Census month (according to the Fabling-Maré labour tables), and to people who have an x-y coded (ie, tier 1) Census address.

Pooling NZTA addresses across IDI instances

- Residential address data supplied to Stats NZ by NZTA is a snapshot of current addresses
- Consequently, when an individual renews their drivers licence (say), their previous drivers licence address disappears from the address data
- However, we can make use of previous IDI archives to recover the lost address information, regaining a large number of high quality historical residential addresses
- Using previous IDI archives means that we rely on MB'18, avoiding the need to allocate addresses that have been split in more recent MB instances
  - Relying on MB'18 removes the need to recalculate MB adjacency and commute distances for new MB boundaries
  - MB'20 is almost identical to MB'18. Further, MB splits may not increase address precision because the IDI mapping to MB often assigns individuals to multiple adjacent MBs (which we group)

Additional addresses gained from pooling NZTA data





All residential NZTA addresses, restricted to individuals who are ever employed from 2005 onwards (ie, the population of interest). Address notifications that appear in both current and prior IDI instances are attributed to the current instance.

Stacking improves longitudinal dimension of NZTA data

	Proportion of individuals			
	Drivers license		Motor vehicle	
N(addresses)	Unstacked	Stacked	Unstacked	Stacked
1	0.984	0.648	0.963	0.743
2	0.016	0.266	0.032	0.210
3	0.000	0.070	0.004	0.039
4	0.000	0.014	0.001	0.006
5+	0.000	0.002	0.000	0.002
N(individuals)	4,455,483		2,240,589	

All individuals who are ever employed from 2005 onwards (ie, the population of interest), and who have at least one NZTA residential address in the relevant series in the current IDI instance (20201020).

- Consistent with the NZTA data being a snapshot of most recent notified addresses, only a small proportion of individuals have multiple addresses in the 20201020 IDI ("unstacked" columns)
- After stacking, at least a quarter of individuals have more than one address observations

Adapting method to include second Census

- Census 2018 requires two source-specific rules
  - The current IDI instance includes admin-imputed Census responses in the address notification table. These have been removed, since they do not reflect actual Census responses
  - The standard address methodology removes sequentially repeating addresses from within source, to avoid over-relying on address sources that may repeat unverified address notifications (dubbed "pinging")
  - The logic for removing repeated addresses doesn't apply to Census, so the repeated address rule is not applied in this case
    - Not exempting CEN'18 would have resulted in losing 1,178,826 Census addresses of employees who are at the same Census address in 2013 and 2018

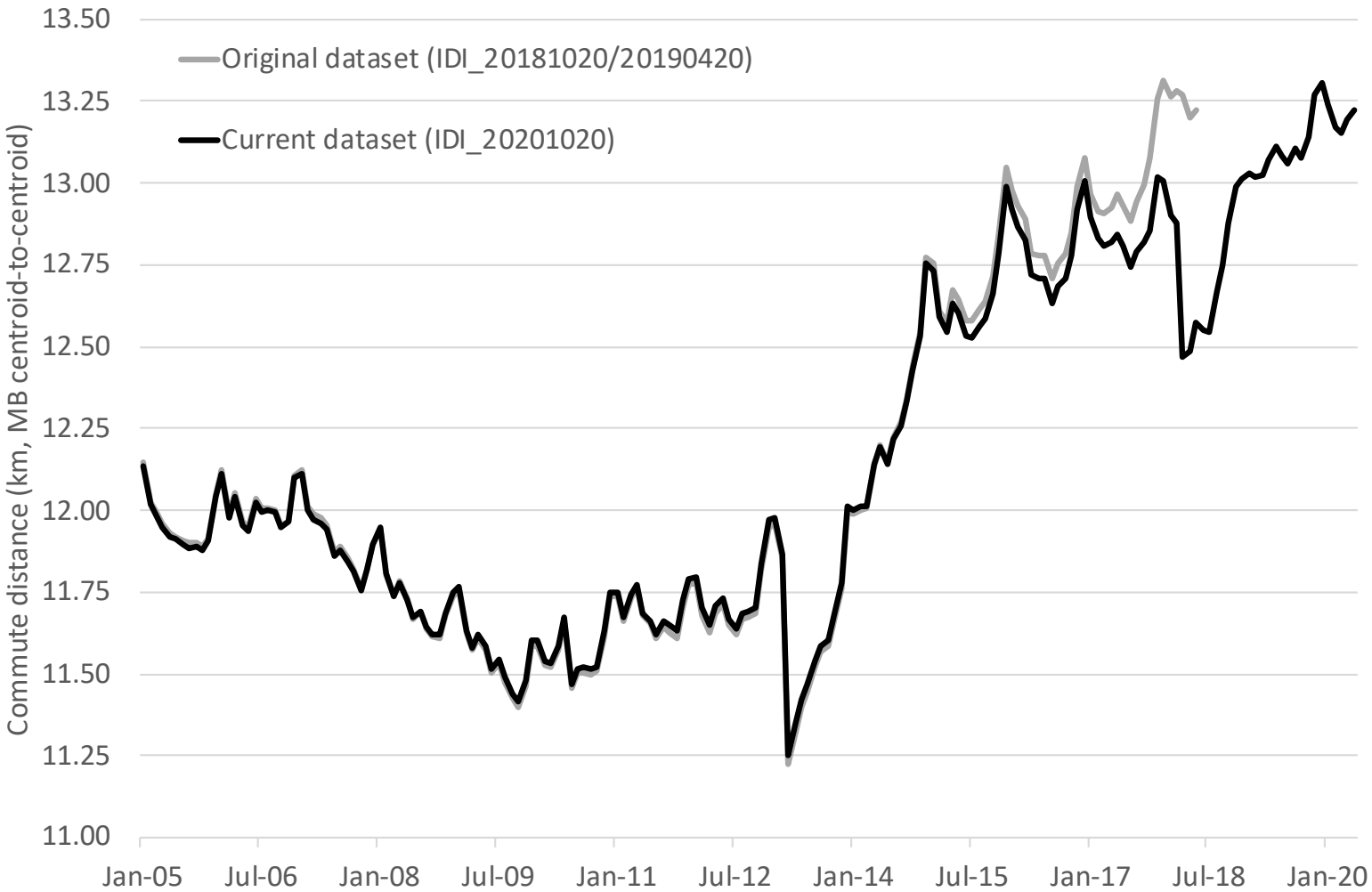
Revising weighting for missing commutes

- In the original methodology
  - Missing commutes account for 8% of total full-time equivalent (FTE) employment, consisting of teachers (4.6%) – who are not allocated to job locations – and workers whose known employer addresses aren't within 200km of their residential address (2.7%/0.7% for single-/multi-location firms)
  - Missing commutes were imputed as being the mean commute of other workers who live in the same MB
- However, Census analysis suggests that teachers tend to have shorter commutes than non-teachers
- Furthermore, mean commutes are substantial higher than median commutes, implying that mean imputation likely overestimates the true distance of missing commutes and, therefore, overestimates the median commute
- The new methodology
  - Applies a residential MB weight (total FTE/observed commute FTE) to all observed commutes from that MB, preserving the distribution of commute distance within the MB

Other changes to commute data

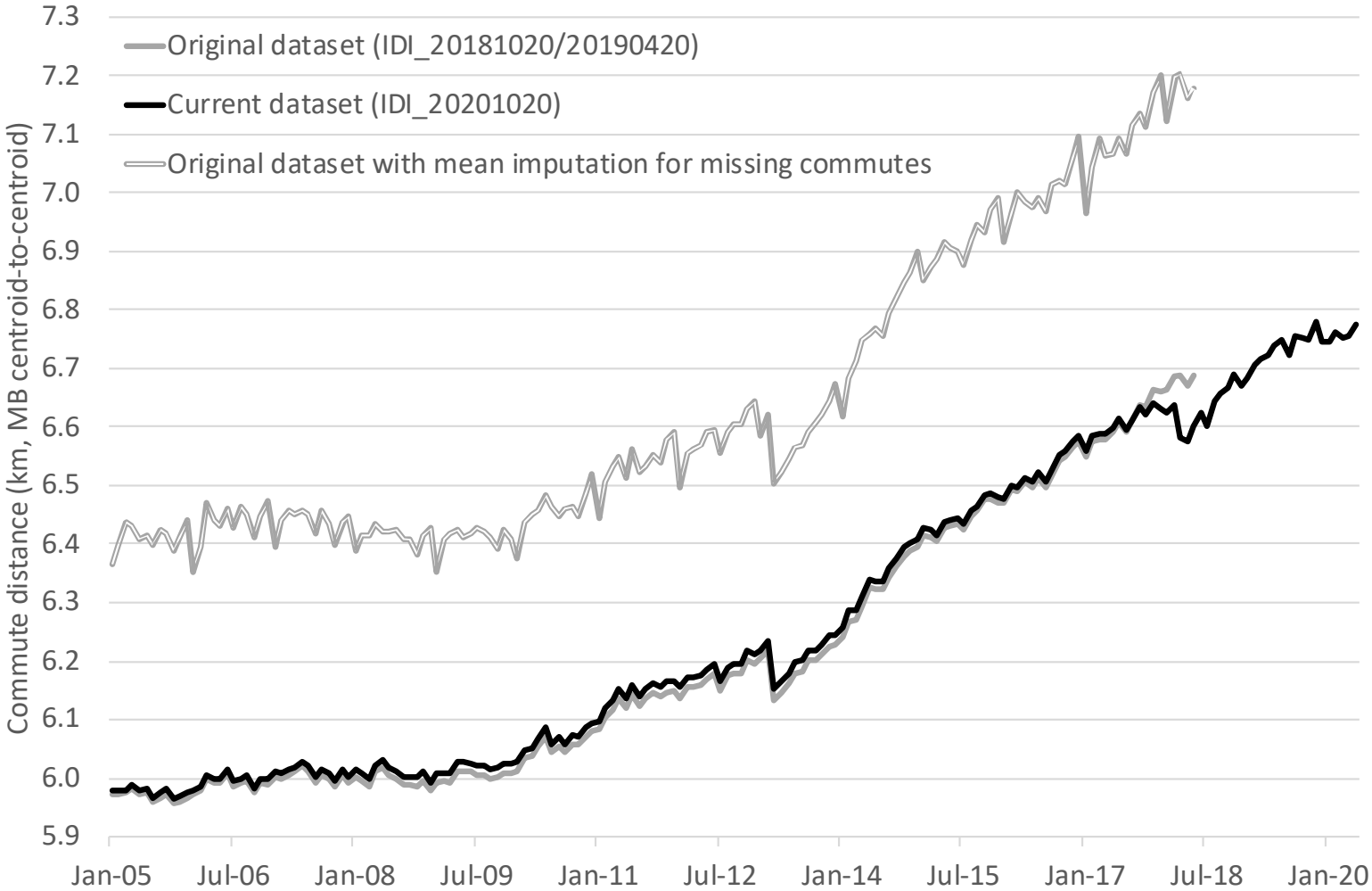
- Within-MB commute distance has been changed from zero to $\sqrt{\text{MB land area}/\pi}$, reflecting that commutes are likely longer within larger MBs
- To make clear that we no longer use the MB instance associated with the current IDI instance, all MB variables have been relabelled from "MB" to "MB18"
- To manage dataset size on the IDI Sandpit
  - Residential address tables are restricted to individuals who are ever employed from 2005 onwards (previously all employees & working proprietors)
  - Commute data are restricted to 2005 onwards (previously November 2003 onwards)
  - Tables feeding into the construction of the commute dataset are restricted to 2005 onwards (previously April 1999 onwards)
  - Variable types have been changed: MB18 (`char(7)`→`int`); distance/FTE/weights (`float`→`real`); flags (`tinyint`→`bit`)

Mean commute



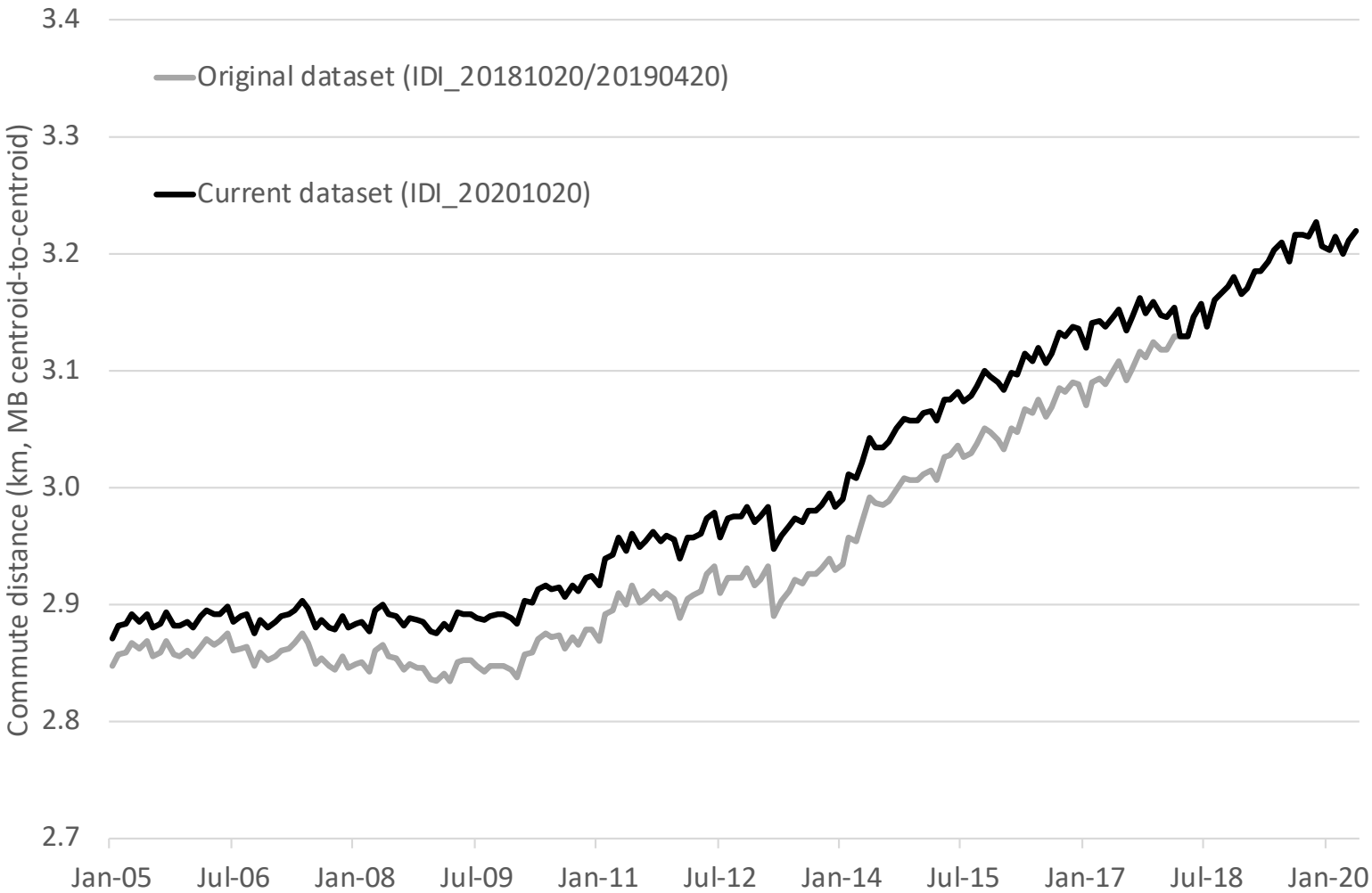
For comparability, both the original and current dataset series use the new residential MB weighting methodology for missing commutes.

Median commute



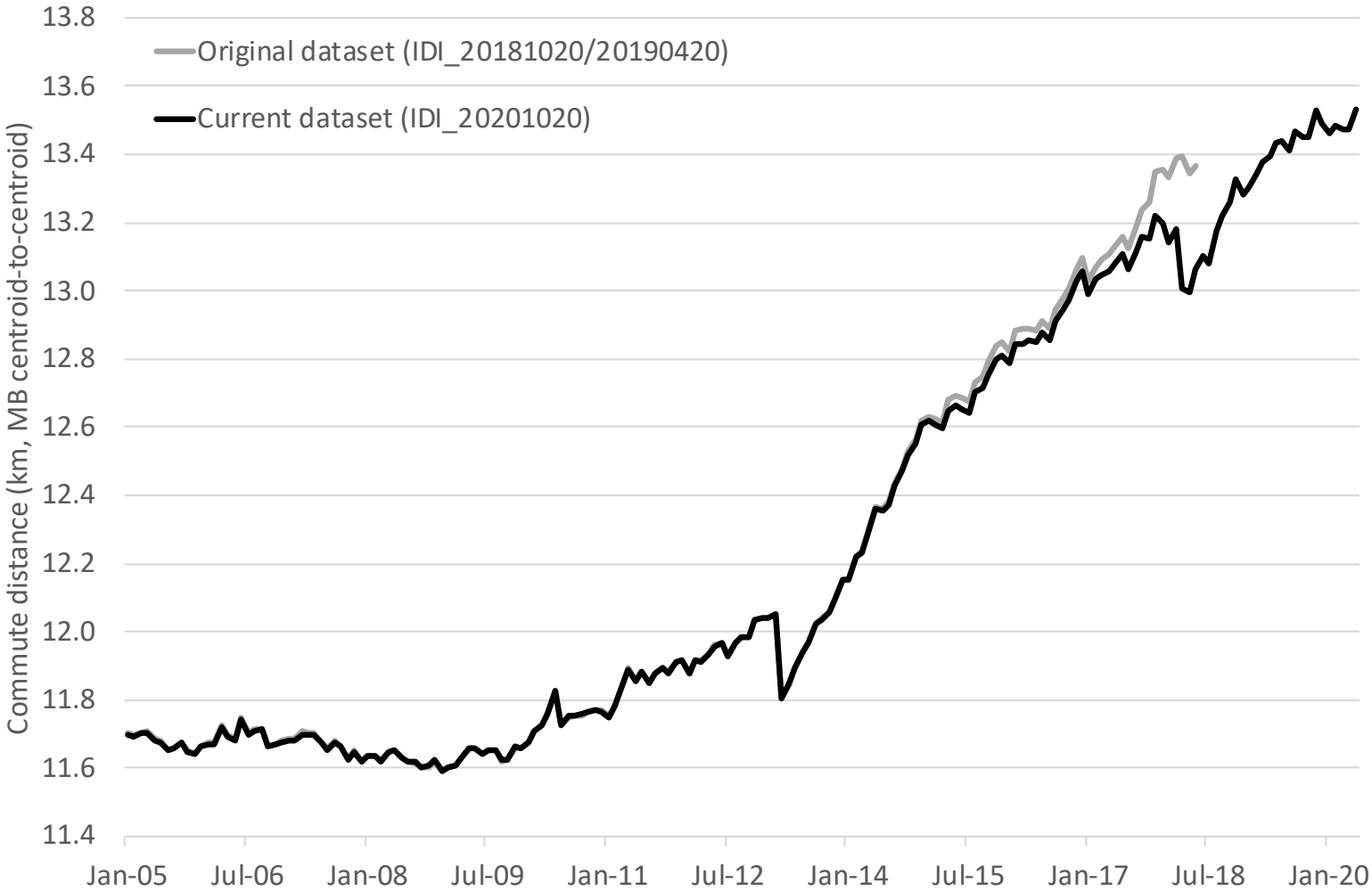
For comparability, both the original and current dataset series use the new residential MB weighting methodology for missing commutes (two solid lines). The hollow line uses the original dataset with mean imputation weighting for missing commutes (as reported in the original paper).

25th percentile commute



For comparability, both the original and current dataset series use the new residential MB weighting methodology for missing commutes.

75th percentile commute



For comparability, both the original and current dataset series use the new residential MB weighting methodology for missing commutes.

Reasons for changes in aggregate commute statistics

- Median plot shows the substantial impact mean imputation has on (incorrectly) raising the median commute
- 25th percentile increase in commute distance is caused by change in measurement of within-MB commute distance
- Other series are generally consistent up to the end of 2016
- From 2017, the new and old series start diverging
  - From March 2018, this effect is directly due to Census and reflects the same phenomenon as observed in 2013
  - Prior to March 2018
    - Business Register (BR) revisions change work locations
    - CEN prevents backcasting of NZTD MBs that differ from CEN
    - New admin addresses create new residential address spells
  - On average, all these effects result in shorter commutes, implying improved (res & work) location data
- Future revisions to time series statistics should be less substantial unless major revisions to the BR occur, or a large new residential address source is added to the IDI