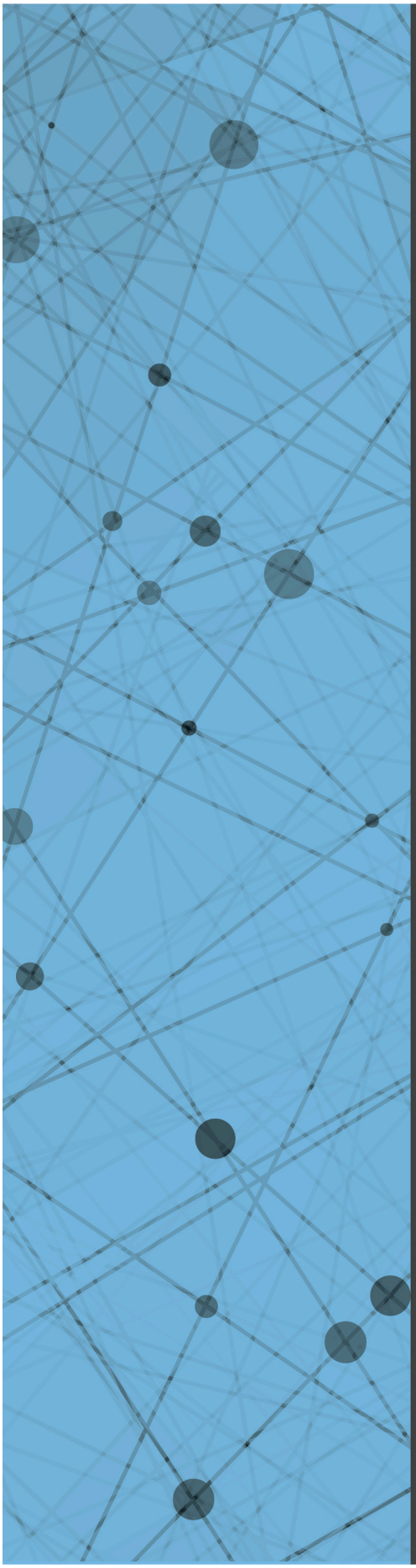
Motu



Measuring commute patterns over time: Using administrative data to identify where employees live and work

Richard Fabling and David C. Maré

Motu Working Paper 20-05

Motu Economic and Public Policy Research

July 2020

Document information

Author contact details

Richard Fabling
Independent Researcher
richard.fabling@xtra.co.nz

David C. Maré
Motu Economic and Public Policy Research
dave.mare@motu.org.nz

Acknowledgements

We gratefully acknowledge funding from the New Zealand Transport Agency (NZTA), and valuable feedback from Ian Binnie and Ernie Albuquerque.

Disclaimer

The results in this paper are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI), managed by Stats NZ. The opinions, findings, recommendations, and conclusions expressed in this paper are those of the authors, not Stats NZ, NZTA or Motu. Access to the anonymised data used in this study was provided by Stats NZ under the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this paper have been confidentialised to protect these groups from identification and to keep their data safe. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy Impact Assessment for the Integrated Data Infrastructure available from www.stats.govt.nz. The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes. Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. 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.

Note: Figures 25 and 26 have been updated since the original release of the paper.

Motu Economic and Public Policy Research

PO Box 24390 info@motu.org.nz +64 4 9394250
Wellington www.motu.org.nz
New Zealand

 ISSN 1176-2667 (Print), ISSN 1177-9047 (Online).

Abstract

We use administrative and survey data in the Integrated Data Infrastructure (IDI) to allocate workers to job locations (plants), which enables the production of over a decade of commute distance population statistics for New Zealand employees. We find that average commute distance (meshblock centroid-to-centroid) fell from 2005 to 2009 before rising again through to 2018 (the final analysis year), with most regions displaying this general temporal pattern. Census 2013 place of residence and work is used to test our methodology against the alternative of using pre-existing plant allocations from the Linked Employer-Employee Data (LEED) production system. For a consistent set of individuals, our estimate of the commute distance distribution closely matches the corresponding distribution in Census. In contrast, LEED-based estimates tend to significantly overestimate commute distances, including radically overestimating the likelihood of inter-island commuting. Our more plausible results are primarily due to re-engineering the job allocation process, as opposed to exploiting better administrative data, though we make marginal improvements to residential address identification through a new prioritisation method, allowing us to use a broader set of residential address sources than available in LEED.

JEL codes

R40; R41; M21

Keywords

commuting patterns; linked employer-employee data (LEED); Integrated Data Infrastructure (IDI); administrative data

Summary haiku

How distant is work?
Admin data will show us
(with a little help)

Acronyms used in the paper

ACC	Accident Compensation Corporation	MB	Meshblock
BR	Business Register	MSD	Ministry of Social Development
EMS	Employer Monthly Schedule	NHI	National Health Index
ENT	Enterprise Number	NZ	New Zealand
ESP	Education Service Payroll	NZTA	New Zealand Transport Agency
F-M	Fabling-Maré	O-D	Origin-Destination
FTE	Full-Time Equivalent	PBN	Permanent Business Number
IDI	Integrated Data Infrastructure	PENT	Permanent Enterprise
IR	Inland Revenue	PHO	Primary Health Organisation
LBD	Longitudinal Business Database	SA1	Statistical Area One
LBF	Longitudinal Business Frame	TA	Territorial Authority

LEED	Linked Employer-Employee Data	UA	Urban Area
LMA	Labour Market Area		

Foreword

The transport system is the primary means by which people gain physical access to the opportunities (e.g. jobs, goods, services, activities and destinations) they need for their livelihoods and wellbeing. Access to employment is a key opportunity for many people – and the location of employment is often co-located near other social services (such as retail and social services).

Practitioners involved in transport planning, land-use planning (district, regional plans), and system performance monitoring need information about where people live, where they work and, ideally the means of commuting to work. Existing data on the experience of people using the transport system to reach their place of employment is either irregular (for example, drawing upon the Census) or sample-based (for example, drawing upon transport surveys). An opportunity exists to use de-personalised administrative records to link together employees' residential addresses and their locations of employment.

Waka Kotahi's (NZ Transport Agency) interest in exploring this opportunity arises from a renewed focus on transport-accessibility as a Government priority for the land transport system. One of the four strategic priorities of the 2018 Government Policy Statement on land transport (GPS) is "a land transport system that provides access to economic and social opportunities". This includes an increased focus on "transport and land use planning that improves access by reducing the need to travel long distances to access opportunities like employment, education and recreation" (page 8 of the 2018 GPS). One way of measuring progress in this area includes "how many people can access major areas of activity within a reasonable timeframe" (page 26 of the 2018 GPS). The need for focusing on access to work is also articulated in the Ministry of Transport's Transport Outcomes Framework which states that "Inclusive Access is one of five key Transport Outcomes … which involves enabling all people to participate in society through access to social and economic opportunities, such as work, education, and healthcare".

Ian Binnie
New Zealand Transport Agency
Performance Measures Specialist

1 Motivation

We identify individual-level commute patterns from administrative data in the Integrated Data Infrastructure (IDI) to enable a range of summary statistics to be produced for performance monitoring, planning, evaluation and research purposes. These data have the advantage over other sources of covering almost the entire population of paid employees, and spanning over a decade on a consistently measured basis. Both residential and work locations are identified at the meshblock (MB) level, which corresponds to approximately a city block in dense urban areas, enabling very detailed geographic breakdowns in commuting patterns over time.

The code developed as part of this research enables the ongoing production of updated data on the IDI, working within a suite of already-developed datasets focussed on employment outcomes and firm performance.[1] The integration of the data into the IDI and Longitudinal Business Database (LBD) allow the data user to identify sub-populations of interest, including personal characteristics such as sex, age and ethnicity, as well as business characteristics such as industry and measured productivity.

A major obstacle to producing high-quality commuting statistics is the job address information available in the IDI, since Employer Monthly Schedule (EMS) filing is generally done at the firm level. The current methodology for determining job location relies on an accurate Business Register (BR) of plant locations and, in the case of workers in multi-location firms, the Linked Employer-Employee Data (LEED) processing system, which provides an existing technology to allocate workers to plants so that regional LEED statistics can be produced.[2] The job allocation process is further complicated by the existence of employers who file collective EMS returns for multiple enterprises, which we dub "joint-filers." In the case of joint-filers, the pool of potential plants an employee may work at includes all plants of grouped employers. Thus, the allocation of a worker to a specific plant also allocates that worker to a particular firm (and a particular set of co-workers).

Stats NZ (2019) explains the LEED allocation process as follows:

"For enterprises with complex structures, including multiple ge-

[1] To help other users make use of the commuting tables, a full set of data codebooks in provided in an appendix.

[2] The LEED plant allocation is available as part of the EMS table in the IR schema on the IDI. Under current access rules for tax data, these firm and plant identifiers are available only to researchers at Government agencies (including universities).

ographic units, the allocation of jobs must take into account several factors. The employment count figures for each geographical unit associated with the enterprise on the LBF are used as target figures.[3] The jobs to allocate can then be divided across the geographic units using these target figures. This is done using an algorithm which minimises the travel distance between an individual's location and the employer's geographic location while aiming to keep the employment counts in proportion to the targets. A second algorithm aims to keep continuing employees at the same geographical unit."

While it makes sense to allocate workers to job locations near where they live, the LEED system does this in a way that results in a significant proportion of implausible (inter-island) commutes, too many long commutes (benchmarked against Census), and randomised allocations. This last issue arises from requiring a deterministic allocation even in the presence of multiple equally plausible candidate job locations. In practice, therefore, Stats NZ's second algorithm – which forces job locations to be persistent – is necessary because random allocations every (quarterly) processing cycle would give the false impression of substantial movement of workers between plants in a firm. In "solving" the problem of random allocation, imposed stickiness of job location creates an issue for genuine job location mobility within firms, which is suppressed, even when a residential move across the country may clearly signal a change in job location.

The main design issues with the LEED allocation system are twofold: firstly, the method prioritises allocating workers in proportion to *expected* employment at each plant even if this results in implausible commutes; and, matching based on travel distance is a binary concept that distinguishes within-Territorial Authority (TA) commutes only as being closer than anywhere else.[4]

To demonstrate the consequences of these choices, consider a firm that has one North Island and one South Island employing location (plant). An employee who lives just outside the TA where the North Island plant is located is just as likely to be assigned to that job location as someone who works at the firm and lives in the South Island (unless they live in the TA of the South Island plant). For new workers who live in the same TA as the North Island plant it is possible that that plant is not even a candidate job

[3]The Longitudinal Business Frame (LBF) is the precursor to the Business Register (BR).

[4]New Zealand has 67 Territorial Authorities (TAs), defined under the Local Government Act (2002) as a city council or district council.

location if the BR indicates that the location already has the "right number" of workers. For these reasons, researchers experienced with using EMS data tend to avoid analysis that relies on the accuracy of LEED allocated job locations (see, eg, Fabling and Maré 2015a).

In part, the LEED algorithm results from a lack of trust in the quality of the residential address information, relative to the business location and employment information from the BR.[5] If the residential address information is relatively low quality then it makes sense to discount its impact on job location choice. However, a number of recent studies have shown that administrative address information in the IDI can do a reasonable job of identifying where people live (as assessed by Census address comparison), particularly when addresses are prioritised based on a quality rule (Stats NZ 2013, 2017, 2018; Gibb and Das 2015; McLeod 2018).[6]

Our analysis confirms the low quality of the job allocation by comparing Census commute patterns to those derived from LEED job locations paired with administrative address information from the IDI. Our re-engineered approach to job allocation is built on the assumption that residential addresses are the highest quality input in the process. To that end, we clean the residential administrative address data in a way that allows us to make use of more data sources than previous methods, while maintaining a high consistency with (benchmark) Census data. These new cleaning and prioritising processes make moderate improvements over the existing Stats NZ methodology available in the IDI, particularly in identifying the residential address of very recently moved workers.

With high quality residential address data in hand, we shift the focus of the job allocation to ensuring the feasibility of commutes. At the same time, we relax the constraint of making singular job allocations when the data do not support clear-cut choices. Instead, we apply a probabilistic approach to job location that allows for the fact that an individual is more likely to work at a large plant than a small plant, given equal commute distance. We also recognise that the data do not identify feasible commutes for all individuals, and rely on imputation (reweighting of observed commutes) for these jobs.[7]

[5]The second design issue also follows from the first issue, since a realistic commute requirement excludes some job matches and reduces the flexibility of the system to exactly match expected employment patterns, or results in workers without job allocations.

[6]LEED makes use of only IR residential addresses, rather than the full set of IDI administrative data sources. Even in the absence of LEED switching to using more comprehensive administrative data, we find that IR addresses, after cleaning, are sufficiently high quality to be included in our top tier of IDI residential address sources.

[7]We ignore some issues that may contribute to implausible commutes, particularly that the

A potential downside to prioritising commute feasibility is that we no longer match *expected* job counts from the BR, and we show how substantially our allocations deviate from these expectations at the TA level. Another issue that requires special attention is the single joint-filer responsible for the Education Service Payroll (ESP). In this case, since the density of job locations (schools) is so high, the probabilistic approach to allocation results in too many cases to allow for a manageable dataset of potential commutes. Instead, we rely on imputation, bearing in mind that the LEED method alternative is to essentially pick a school at random and then hold a teacher in that location over time.[8]

We test our method against Census (2013) data in the IDI – replicating the general commute profile for workers – and then perform several simple analyses with the resulting administrative commute dataset to demonstrate the power of the data. Specifically, we show that average commute distance (MB centroid-to-centroid) fell from 2005 to 2009 before rising again through to 2018 (the final analysis year), with most regions (TAs) displaying this general temporal pattern. Focussing on the Hamilton Urban Area as an example, we then demonstrate how the integration of road network data enables additional commute statistics such as commute times, and how general trends in commute time can be decomposed into effects due to workers moving between regions and jobs.

Section 2 describes the data we use to construct the commute dataset, and the population of interest. Section 3 summarises the processes for cleaning the residential data and identifying job locations, including testing of the methodology against Census data, while section 4 provides example analyses using the final commute dataset. Section 5 summarises our findings and also highlights a number of fruitful avenues for further analysis using the data.

BR may not be a timely, accurate portrayal of physical locations of firms (particularly for smaller firms), and that employment counts for some geographical units may be updated infrequently. As Fabling and Sanderson (2016) note, the amount of survey data feeding into BR updates is declining over time, implying that quality issues of this sort are likely to be increasing over time. Particular issues may exist for the identification of firms expanding from a single to multiple locations, since single location firms are less likely to be surveyed than known multi-location firms.

[8]The issue of imposed job location stickiness in the LEED allocation is likely to be particularly problematic for the ESP, since teachers do move between schools.

2 Data

The population of workers comes from the latest version of the Fabling-Maré (F-M) labour tables, which are based on EMS data in the October 2018 IDI instance (IDI_20181020) and include monthly gross earnings from employment from April 1999 to May 2018. The main advantages of the F-M labour tables over the raw EMS data are the: removal of self-employed from the EMS; imputation of full-time equivalent (FTE) employment; integration of the labour data with the firm productivity dataset on the LBD, including the use of permanent enterprise numbers (PENTs); and, inclusion of two-way wage fixed effects estimates which measure the portable wage premium of workers (Fabling 2011; Fabling and Maré 2015a, 2015b, 2019).

Residential address information, starting in January 2000, come from a more recent IDI instance (April 2019, IDI_20190420) that includes major revisions to Stats NZ's address data processing methodology, which substantially increased the rate at which raw administrative addresses are coded to x-y coordinates and MBs. Ideally, we would update the F-M labour tables to this IDI instance, but space constraints make it hard to justify updating the labour tables more frequently than annually. Fortunately, because individual IR numbers are permanent characteristics of workers, confidentialised versions of these can be used to map between IDI instances.[9]

Business address information comes from the BR tables on the April 2019 IDI, and both residential and business addresses use 2018 meshblocks. The links between (confidentialised) employer IR numbers and firm identifiers (IR-ENT link), and between firms and plants (ENT-PBN link), are available through either the current BR instance on the IDI or from LEED, which uses previous instances of the BR and LBF. We choose to use the links taken directly from the LEED dataset, rather than the BR, so that we maintain consistency with LEED as to which firm locations are employing (active) in any month.[10]

Our main rationale for preferring the LEED link set is threefold – firstly, the BR does not keep track of joint-filer relationships and these are an important component of the LEED allocation process, particularly for the ESP

[9] Technical linking issues are discussed in more detail in the codebook appendix.

[10] Stats NZ uses the terminology of enterprise number (ENT) for firm identifiers and Permanent Business Number (PBN) for plant (physical location) identifiers. Fabling (2011) extends this terminology to include PENTs for permanent enterprise numbers, created by tracking PBN transfers across firms.

employer IR number.[11] If the BR and LEED disagree on the correct IR-ENT link, this may not be easily reconcilable with observed joint-filing patterns in LEED. Secondly, historically, the BR (LBF) has not kept track of time-variation in IR-ENT links, whereas LEED has such variation because LEED has an eighteen month revision window, outside of which historical IR-ENT relationships are permanently locked in. Using current BR links may project back IR-ENT relationships that didn't hold in earlier time periods.[12] Finally, unreported tests suggest that including non-employing plants on the BR in the feasible job location set does not appear to substantially reduce the rate of infeasible commutes in the final data, consistent with those plants being inactive at the time.

Census 2013 is used to update residential addresses in the final commute dataset but, for testing purposes, we include only administrative addresses in the commute dataset. We then compare these results excluding Census to Census 2013, also on IDI_20190420, for the same set of linked individuals in both the administrative data and Census. For this testing we use the original Census address files, which uses 2013 MB,[13] and make a number of minor restrictions to focus on the population of interest. Specifically, from the F-M labour table we use only individuals employed in March 2013 (Census month), aged 16-79 and with non-missing sex. From the Census side, we require the usually resident address to be a MB within a TA.

Figure 1 shows the number of March 2013 workers by age in the F-M labour dataset, together with the match rate of those individuals to a usually resident Census response in the IDI (restricted to the 16-79 age range). The main point to note from the figure is that our testing sample is not completely representative of the population of interest. As other studies have shown, we have relatively low coverage of workers in their 20s. Some undercoverage derives from the Census usually resident population restriction, which excludes some temporary migrant workers. Overall, 13% of workers are not linked to a Census return with a usable residential address, with this proportion rising to over 20% for 25 year-olds. We also lose a small proportion of worker

[11] Unlike other joint-filers, the ESP employer IR number does not appear on the BR IR-ENT link table since it is not associated with a real enterprise.

[12] This potential issue is exacerbated by the fact that the IDI performs "deduplication" of IR numbers, which creates the appearance of one-to-many IR-ENT links on the BR. Applying one-to-many links in the LEED data would create additional (and incorrect) joint-filers. Testing suggests that joint-filers created by the deduplication process may exist in the IDI version of the LEED data, but that they occur so infrequently that the effect is negligible.

[13] Administrative residential address data is recoded from 2018 MB to 2013 MB using a unique mapping that follows from the fact that MBs are only ever split over time.

observations from the age restrictions we impose (0.6% of workers) and from the lack of age data in the IDI (0.1% of workers). The joint sample of March 2013 F-M labour table workers linked to Census observations is 1,563,240 individuals (table 1).

To test the job location and commute pattern methodologies, we need to make further restrictions for the Census comparison, since Census is concerned with "main" job, and EMS has all paid jobs. The simplest way to make the data comparable is to restrict to single job holders so that the concept of "main job" in Census is as consistent as possible to the administrative (EMS) job. From the Census data, we restrict to individuals whose main job is as a paid employee, who worked away from home of Census day, and where we observe the Census work address MB. Harmonising from the administrative data side, we restrict to workers in the EMS who are in a mid-spell job (ie, not a starting or ending job) with a single employer in March 2013, and who have no self-employment income in the 2013 or 2014 (March) tax years. The Census testing sample for commute patterns has 988,785 individuals (table 1).

3 Methodology

Our approach hinges on high quality residential addresses providing a sufficiently accurate locational "anchor" to then identify the correct work address(es) within a feasible commute radius. However, we begin with an address *notification* dataset that pools together various agency-specific data generation processes into a single store of address information of mixed quality. Our goal initially, therefore, is to harmonise these data streams so that they, as much as possible, take on the properties of a high quality address *change* dataset. From an address change dataset we can derive address spells that allow us to identify a best guess residential address on any date within the spell. Job locations are then drawn from the LEED and BR datasets and simple rules that prioritise short commutes allow us to select the subset of job locations that are in commute range of the residential address.

3.1 Residential address cleaning

We begin by restricting the residential address data to the population of interest, which is individuals who are ever employees in the F-M labour dataset. Residential addresses come from the `address_notification_full`

table in the IDI, except IR addresses which we take directly from the IR schema instead, so that we include addresses Stats NZ deliberately exclude in their initial pooling of the address data. Stats NZ exclude IR addresses that occur on days with unusually high numbers of notifications (dubbed "spikes"), which indicates they are unlikely to be real notifications. We return these data to the address table because we believe a subset of the observations are usable, and because we observe spikes in other address notification sources and want to apply a consistent identification and treatment method to all data sources where they occur.

Table 2 lists address sources together with the number of workers in the Census comparison who have at least one address notification from this source in the year to March 2013. Addresses are classified into two quality tiers (leftmost column) based on their potential to match the Census residential address, with match rates shown in the last three columns of the table. These columns show the probability that the same (2013) MB as Census ever appears in the administrative source over the year, that the same or an adjacent (touching) MB ever matches Census, and the difference between those two probabilities (ie, the adjacent, but never exact, probability).

Besides the address source itself, the other dimension we use to disaggregate addresses into quality tiers is whether the address can be coded to an x-y coordinate pair by Stats NZ as part of their process for determining the MB of an administrative address. While we do not have access to these coordinates for confidentiality reasons, the ability for this detailed coding to occur implies higher informational content in the administrative address. This interpretation is backed up by comparing the match rates of the first and second panel of addresses on a source-by-source basis, which shows that x-y coded addresses are more likely to match Census. Additionally, the gap between exact and adjacent matching to Census is substantially smaller for x-y coded addresses than non-coded addresses, suggesting more noise in the location identification process for the latter group.

The third panel of the table includes two address sources that have lower potential match rates to Census, regardless of x-y coding status – Ministry of Education and MSD (postal addresses). The former has very few observations, given our restriction of the test population to workers aged at least 16, while the latter's poor relative performance is due to it's inclusion of postal as well as residential addresses, as noted in the address source description.

Since the two Ministry of Health address sources – NHI and PHO – have similar measured quality we pool these into a single agency source,

which is consistent with the treatment of Inland Revenue addresses, which are generated by a number of different client lists within IR, but pooled into a single supply to Stats NZ. In contrast, we do not pool MSD addresses because of quality differences between the residential and postal address list. Instead we will promote a subset of the MSD postal addresses to tier one where they are also MSD residential addresses. Finally, the two NZTA sources – drivers licence and motor vehicle – are also kept separate because the infrequent nature of the former means that we give it a slightly different treatment in the address prioritisation process.

Each tier one address source has an adjacent-MB match rate to Census of at least 77%. When pooled together, the probability that at least one tier one address notification identifies the correct or adjacent MB is 89% (bottom panel of table). In contrast, pooled tier two addresses, have a 55% chance of including the correct or adjacent Census MB. The large majority of addresses, though, are tier one and when we prioritise addresses to determine the most likely residential address, we will use tier two addresses only for the small subset of workers who never have a tier one address.

If we relax the one year prior address requirement, only 3.1% of individuals do not ever have their Census usually resident address MB (or adjacent) in the administrative address data. Regardless of our methodology, these workers can never be assigned their "correct" address without including Census data in the address choice set – which we do in the final version of the methodology. The "correct" address may be missing because of issues with the quality of either the administrative or Census address data (though our testing assumes the latter is superior), a lower frequency of address observation than necessary for some residential movers, or because of incorrect matching of data sources on the IDI.[14]

Table 3 provides an alternative lens on the ability of the residential address data to support precise x-y coordinates. Specifically, given repeated notification of the same MB in a tier one source, it shows the probability that we observe multiple x-y coordinates for that worker living in that MB. If x-y pairs were error-free, the final column of the table could be interpreted as the proportion of workers who ever report within-MB moves to new residential addresses. However, these rates are too high for that interpretation to be likely, suggesting that there is noise in the x-y data.

[14]On the last of these, geographic information is used to link the Census to the central IDI register (the "spine") in most matching passes. Stats NZ attempt to restrict the false positive match rate to less than 2% for each dataset linked to the IDI spine. They estimate the Census 2013 false positive match rate to be approximately 0.8% (Stats NZ 2020).

The presence of noise matters because we want to construct an address change dataset. Table 3 confirms that a change in x-y coordinates is not a better indicator of address change than a change in MB is. Thus, we can focus exclusively on MB-level addresses and ignore x-y coordinates, except where their absence implies address quality issues (captured by address tier).

Table 3 also provides justification for wanting to switch from an address notification to an address change dataset. The table shows that the probability of a MB address recurring within an address source varies widely across agencies. For example, drivers licence addresses generally appear only once per individual, consistent with the low frequency of licence renewal. Conversely, Ministry of Health addresses recur 79% of the time, because repeated interactions with the health system generate address verification and/or a "pinging" of the current address. While the former provide information that an address is still current, the latter indicates only that an agency interaction occurred. Because agency systems may differ in terms of the relative frequency of these two types of data generation – verified and unverified – retaining all address notifications would potentially give more weight to data from agencies who ping unverified addresses. By collapsing each data source to address change information, we reduce the pinging issue and treat all sources equally. The downside of this approach is that we discard true agency address verifications, since they are indistinguishable from address pings without imposing assumptions based on individual agencies stated practices.

A corollary of errors in residential addresses and x-y coding is that, sometimes, addresses will be coded to the wrong side of a MB boundary. This might arise, for example, from a street address number being recorded as even when it is actually odd. Of those workers who have more than one residential MB recorded in the raw address data, an implausible 27.5% appear to have lived in two or more adjacent MBs. If we do not account for this noise, then we will identify such changes in MB as actual address changes, and addresses sources with more noise will receive greater weight in the address change dataset.

To solve this issue we merge geographically adjacent addresses (across sources) at the individual level and then choose the most connected tier one MB as representative of the connected group.[15] Table 4 shows that most

[15]The most connected MB is the one with the greatest number of adjacencies to other residential MBs the worker appears to live in. If there is a tie for most connected – as would be the case for a pair of merged adjacent MBs – we break ties with the most recently recorded MB on the expectation that address quality is likely to be improving over time.

resulting MB groups consist of two adjacent MBs. By extension, this result implies that an address may be miscoded to an adjacent MB which we don't identify as miscoded because there is insufficient variation in the notification data to pick up the alternative MB. For that reason, all testing against Census addresses treats adjacent MBs as matches unless explicitly noted in the table (as in table 2 where we show both exact and adjacent matches).

After applying the adjacency rule, we identify MSD postal addresses that are likely to be residential address as those addresses that also appear in the MSD residential addresses. For consistency with the treatment of IR and Ministry of Health addresses, we promote these addresses to a pooled tier one MSD category, leaving the remaining MSD postal addresses as a separate tier two group.

Similarly, we promote MB addresses with missing x-y to tier one if they appear as tier one on another date for that source, since address quality may improve over time. Promoted addresses, on average have a 75% chance of having any (adjacent) match to Census, which is substantially higher than the overall rate for tier two addresses of 55%, but inferior to the match rate for x-y codeable addresses of 89% (table 2, bottom panel). Since we ultimately convert these address notifications into an address change dataset that largely relies on tier one addresses, the main potential impact of promoting MSD postal and non-x-y-coded addresses is to push back in time the observed date at which a worker moves to a particular residential MB.

As already noted, Stats NZ identify spikes in the IR notifications data that are unlikely to reflect real address notification events. We address this issue for all address sources, identifying problematic spikes by focussing on the proportion of recurring addresses at a given date, rather than the total volume of addresses. Our logic is that dates where a source has a high rate of recurring addresses are likely to reflect a data dump of previously notified addresses, rather than new information. This approach picks up all the IR spikes identified by Stats NZ, additional IR address spikes, and previously unidentified spikes in Ministry of Health and MSD addresses.[16]

Aside from identifying more spikes, our method lends itself to being more selective about the addresses removed from the data. Specifically, we remove only recurring addresses that occur on high recurrence rate days, which are defined as days with a (within source) repeat MB rate of at least

In the rare case of remaining ties, the number of times the MB is observed is the final tiebreaker.

[16] ACC is excluded from this rule as the recurrence rate in the addresses is relatively high without any apparent spikes in the address notification rate.

85% (and at least ten workers with address notifications). Non-recurring addresses reported on high recurrence rate days are retained, since they may result from business-as-usual data collection.

A final residential address data quality issue relates to multiple (non-adjacent) MB notifications for the same worker on the same day in the same source. Within-source address date ties cannot be used to determine a preferred address, except where the address tier differs. However, as table 5 shows, tied addresses within the same tier do not all have the same probability of matching the Census address. Unique (non-recurring) MB address that are tied with recurring addresses are very unlikely to be good addresses, perhaps suggesting that the address date tie arises from an incorrect address being replaced by a correct address during the course of an interaction with an agency. Based on this evidence, we remove unique addresses that appear on tied dates. In the case of remaining ties (within source), we remove all tied addresses, and rely on other address notifications to determine the best address at that date.

We then remove sequential occurrences of the same MB within source to get to an address change dataset for each source. The first address in a source is subject to left-censoring so that we cannot be sure that it represents an address change. This logic applies in particular to the first day of IR data (in May 2001), which is identified as an address spike by Stats NZ. Since we know these addresses are not true notifications, but represent historical residential addresses for workers of unknown date vintage, we move this spike of data to a representative date in December 1999, so that these addresses pre-date any other tier 1 source. By doing this, we use these initial IR data as a tier one residential address only as a last resort.

Figure 2 shows the total number of tier one address change notifications by year (grey bars), together with the proportion of raw address notifications retained (black line). The higher volume in 2013 reflects the inclusion of Census 2013 addresses in the final residential address dataset. The retention rate is lowest in 2008 and 2009 due to the presence of address spikes in these years.

Table 6 reports the total loss of observations (rightmost column) by address source, and shows which data cleaning steps account for this loss. Overall, 8.5% of address notifications relate to recurring address spikes, concentrated in IR and Ministry of Health data. Sequential recurrence of addresses is common in many of the data sources, resulting in a loss of 43.2% of addresses when we move from an address notification to an address change dataset. The variation in data loss of this type shows the importance of

harmonising the data. For example, while the ACC data initially has many more notifications than the MSD data (20.7 vs 12.6 million), 80% of the ACC data is the same MB address repeating over time. In contrast, MSD data has a recurrence rate of only 14.5% which means it initially looks much more like an address change dataset. In the final address change dataset, ACC is downweighted relative to MSD with fewer than half the observations (4.1 vs 9.8 million).

The address loss from removed spikes is reduced by retaining the initial IR spike, artificially dated at December 1999. Figure 3 shows the proportion of workers in a month who have a tier one address prior to the employment month, illustrating how the retention of the initial IR spike data helps alleviate the left-censoring of the address data at January 2000 for earlier years of employment data. The arrival of new address change data naturally reduces the reliance on that spike over time. By January 2005, 97% of workers have a pre-2005 tier 1 address change notification, with only 3% of workers still relying on the initial IR spike to supply that most recent address.

3.2 Residential address prioritisation

Having a dataset of address changes with only two quality tiers means we can specify a simple and intuitive address prioritisation rule for the best guess residential MB address on any date of interest:

1. The most recent tier one address up to the date of interest[17]
2. The most recent tier one address after the date of interest
3. The closest tier two address to the date of interest

Tier two addresses form a trivial component of the final dataset (1.2%, bottom row of table 6) and, because of this prioritisation, are used only in the complete absence of tier one addresses.

Figure 4 shows how well this address prioritisation performs in matching the benchmark Census residential address, both at the MB level (solid line) and at the TA level (dashed line). The horizontal axis shows the date of the address change notification used to identify the residential address, with the solid grey bars showing the proportion of prioritised addresses that come

[17] Testing suggests that NZTA drivers licence addresses provide a good match to Census up to a year following Census date, so we allow these addresses to be priority one up to a year after the date of interest. In the case where another rank one address exists prior to the date of interest, the closest address to the current date prevails.

from each month. Following the prioritisation rules, the majority of addresses are tier one and precede Census day (ie, are priority one addresses). There is a pronounced dip in MB match rate to Census prior to Census date because workers who have more recent address change notifications are more mobile than workers with less recent changes, and more mobile workers are harder to follow in the administrative data.

Tables 7 and 8 report MB and TA match rate to Census by address source and Census TA respectively. The overall match rates are 85.3% and 95.7% at the MB and TA level respectively, with the majority of prioritised addresses coming from IR and Ministry of Health (table 7). These two sources have the lowest overall match rates though, as illustrated by the match rate dip in figure 4, the match rate for a given source is not just a function of underlying variation in data quality. Match rates also depend on the composition of the workers who interact more frequently with an agency, including the frequency with which the individual moves addresses, and interacts with and notifies the relevant agency of moves.

3.3 Comparison to IDI prioritisation

In comparison, the IDI residential address prioritisation minimises the use of ACC, IR, NZTA (drivers licence), and the residential subset of MSD postal data, by placing these address sources in a lower address tier (table 9). Not exploiting these additional data sources, coupled with the absence of data cleaning steps (other than partial IR spike removal), means that the IDI method underperforms our method by 2.6 percentage point (pp) at the MB level, and by 1.1pp at the TA level (final two columns of table 10).[18] Figure 5 plots this difference in performance by Census TA, with bubbles scaled by TA size and the diagonal dashed line representing an equal match rate between the two methods. The presence of all TAs above the line indicates our method outperforms the IDI method in every TA. Of the larger centres, the gains are particularly marked in Wellington City, which is a relatively hard TA for either method to find correct addresses for. As table 8 shows, we have a relatively poor 82.1% MB match rate for Wellington City, which is 3pp lower than the large centre average. Despite this relatively poor performance, our method outperforms the IDI method by 5.5pp (IDI match rate of 76.6%, figure 5).

[18] Comparison to the IDI method makes use of the IDI prioritised address that applied immediately preceding Census day, since the IDI address table includes Census 2013 as a tier one address source.

Figure 6 shows that the gain in performance of our method (solid line) over the IDI method (dashed line) primarily comes from improved identification of correct addresses for workers who moved to their Census address within the immediately prior year. This may be because letting more data sources contribute allows address changes to be picked up sooner, or because individuals who move frequently are more likely to have addresses in data sources that the IDI method places lower weight on than our method.

In contrast, the IDI method performs slightly better at identifying the correct residential address for workers who have been at their place of residence for four or more years, probably at least partially because the relatively low number of tier one addresses used by that method results in more address persistence. On balance, though, the performance gap is larger at short durations, and a greater proportion of workers have higher frequency than low frequency moves (figure 6, grey bars). Thus, the overall performance gain of 2.6pp (bottom row of table 10) is made up of a (weighted) 2.9pp match rate gain for workers with less than one year residence, a 0.6pp gain for workers with one to two year residence, and a 0.9pp match rate loss from worse performance for workers with four or more years residence.

Figure 7 looks at the number of residential address changes observed during the period between the first and last ever employment month of each worker. The first two bins in the figure represent the case where zero address changes occur within this maximal employment period, and distinguish between the case where the unchanging residential address is or isn't the first (ie, left-censored) address in the address table. As expected, our method (black bars) shows more address changes than the IDI method (grey bars), reflecting the increase in the number of tier one address sources allowed to determine residential location. Specifically, the IDI method (gray bars) is 3.9pp more likely to record no address changes for an individual, while our method (black bars) is 4.1pp more likely to record more than ten address changes. Overall, our address spell data has 35% more address changes than the IDI data during the maximal employment period of workers. This may reflect the prevalence of high-frequency moves for some people, which is less well captured by the IDI method, but may also potentially reflect some undesirable flip-flopping between address locations.

To test this possibility, figure 8 shows the probability that a prioritised address change indicates a return to a previous residence (MB) in our method (solid black line) and the IDI method (solid grey line). The dashed black line presents an alternative view of the IDI method recurrence rate where we treat adjacent MBs as being the same location, as is done in our methodology.

Even adjusting for adjacency our method shows a higher recurrence rate of MB addresses, though both methods produce implausibly high rates.

Recall though, that the IDI imposes this stability through the prioritisation mechanism. This is likely to hold recent movers at incorrect addresses, as evidenced by figure 6. So, while we allow flip-flopping between addresses, on average our method provides a better point-in-time estimate of the correct address. Finally, in support of our method, the spikes observed in the IDI recurrence rates reflect repeat address spikes that our method identifies and removes, but which are missed by the IDI method.

3.4 Testing job locations

Our ability to test job location allocations is hampered by a lower *potential* match rate between BR plant locations and Census work locations than we found between Census and administrative data residential addresses. In particular, we cannot be certain that matching issues are not at least in part due to issues with Census data quality or conceptual differences between what is being measured in Census compared to the BR. Key candidates for the difficulty in comparing the two types of work address data include:

- the business location data on the Business Register is incomplete (ie, missing plant locations);[19]
- misidentification on the BR/LEED of which business locations are active/employing, including delays in identifying new plants;
- failure of our imposed restrictions to identify a common job between the EMS and Census reporting;
- data quality issues with Census-reported work location, including work addresses that are hard to code to MB;
- workers who worked away from a business location on Census day (eg, on-site construction workers); and
- incorrect linking of individual tax and Census records in the IDI.

Given the relatively high potential match rate between administrative and Census residential addresses, and the low estimated false positive match rate

[19] These data are updated mainly through targeted surveys of firms who are known or likely to have multiple locations (eg, large businesses or businesses whose activities span multiple industries). Stats NZ have been doing less direct data collection over time that could fill gaps in knowledge of the geography of firm activity (Fabling and Sanderson 2016).

for the Census to the IDI spine, the last of these potential reasons seems unlikely to be relevant.[20] Testing indicates that timing on the BR helps with reconciling differences between the two sources, with the match rate to Census increasing if we broaden the search for BR active plants to include non-Census months.

However, conceptual issues are likely to still be an issue. For example, we observe individuals who work together according to EMS data, and who also report the same Census work location (MB) where that location is not ever associated with the firm on the BR. That situation, though, is not prevalent enough to rule out the possibility that Census business address inaccuracy is an important factor. Furthermore, while Census-based agreement between co-workers is consistent with missing plant locations on the BR, it is also consistent with a subset of workers having jobs that require them to work together away from the (BR-identified) office.

As an indication of the size of the matching issue, and to test the "away from the office" hypothesis, table 11 shows the probability that employing LEED plants include the Census-reported work MB, or are within 1km or 5km of the Census-reported MB. Building trades are an example with a low match rate (79.3%) between LEED and Census job locations, which might be expected to be due to site work. A further 8.3% of these employees' Census responses indicate they are working within 5km of a LEED location. However, the variation across occupations in not so strong as to suggest this is the primary explanation of the poor match rate between Census and the BR. For example, the largest occupational group, corporate managers has only a 3pp higher likelihood of LEED matching Census work location than building trades workers, and it seems reasonable to expect such managers to be more likely than builders to have desk jobs in offices identified on the BR.

Overall, 82.7% of workers have their Census-reported work location among the feasible set, and a further 9.5% are within 5km of that address (bottom row of table 11). This gives us some comfort that we can use the LEED locations to provide an at least proximate job location for most workers. Implicitly, this means we treat the physical locations of a firm as a reasonable "average" of the likely locations any worker might travel to on a given day. Because of the relatively poor match rates on business address, the job allocation method testing focuses on the distribution of estimated

[20]Furthermore, the match rate of Census business address to potential BR business address is uncorrelated with the linking pass that resulted in the Census record linking to the IDI spine. If linking accuracy were an issue, we might expect the match rate to deteriorate for linking passes that use looser matching criteria.

travel distances (MB centroid-to-centroid), rather than exact matching of job location.

3.5 Job location allocation

The fundamental issue with identifying job location is that the EMS is predominantly filed at the firm level, but work occurs at physical locations (plants). When firms have multiple physical locations, the EMS data provide no guidance on which location(s) an employee works at. This issue is exacerbated by the presence of "joint-filers," which are (confidentialised) employer IR numbers filing EMS returns on behalf of multiple firms. In such cases, not only is the job location unknown, but the firm is also unknown. In both the LEED processing system and in our method, allocation of job location then also determines the firm in the worker-firm (job) relationship for joint-filers.

Figure 9 shows the proportion of jobs that are subject to joint-filing, separately identifying the Education Services Payroll joint-filer (grey bars) from other joint-filers (black bars). The ESP accounts for approximately 4.5% of all jobs with a seasonal dip in January associated with the summer school break. The non-ESP joint-filer share of total jobs is initially around 10% before an abrupt drop in 2002, followed by a reasonably steady decline to less than 3% of jobs by 2018.

Since we cannot identify employers in the IDI, and joint-filers are used only in the LEED processing system, we cannot tell whether the decline in non-ESP joint-filers results from a real decline in the prevalence of joint-filing or is due to a reduction in the maintenance of the system. Ultimately we are stuck with what we observe, since the establishment of joint-filing relationships relies on a manual process that involves knowledge of the individual business identities and, possibly, relationships between those entities that are not captured by the Business Register.[21] If the stark change in prevalence of non-ESP joint-filing between 2002 and 2003 is not real, caution should be exercised in using data that relies on job (worker-firm) changes over this period for employees being paid by IR payers whose status changes from joint-filer to independent-filer over this period.

In the final commute dataset, approximately 10% of total FTE employ-

[21] While discussions with Stats NZ suggest the definition of joint-filer behaviour has not changed, the fact that identification of joint-filing involves manual judgement could cause some of the variation we observe in the pattern of non-ESP joint-filing.

ment is associated with joint-filing (evenly split between ESP and non-ESP). The remaining 90% of employment is associated with single filers and is evenly split between firms with a single employing location and those with multiple locations.[22] We address the job allocation process for each of these groups separately, starting with the ESP joint-filer.

We make a special exception for workers in the ESP joint-filing because the density of schools makes this a uniquely difficult job allocation problem. Figure 10 illustrates the scale of the problem by plotting the cumulative proportion of ESP workers by the number of schools within 15km (centroid-to-centroid) of the employee's residence. Approximately half of ESP workers have a 15km or less commute to over 50 schools, and 20% of workers are within 15km of at least 160 schools. Given the density of potential job locations, making an allocation in the case of ESP workers would result in either making a very low probability selection or, alternatively, maintaining a large number of very low probability potential job location links.

Instead, we make no attempt at job location allocation for this group and impute their commute distance from non-ESP workers who live in the same residential MB. Figure 11 tests the plausibility of that imputation, showing the cumulative distribution of commute distance for ESP (solid line) and non-ESP (dashed line) workers using Census data. As we might expect given the large number of proximate schools, ESP workers tend to have shorter commutes than non-ESP workers, as evidenced by the leftward shift of the solid line relative to the dashed line. For the median worker, the gap in commute distance between the two is a little over one km.

The dotted line in figure 11 shows how poorly the LEED job allocation does at picking the correct school for ESP workers, substantially overestimating the proportion of workers who have long commutes (dotted line). LEED data imply that approximately one third of ESP workers travel more than 50km to work, whereas Census data suggests this number is closer to 2%. These differences result from the loose geographic matching applied by LEED and, potentially, the imposed stickiness in work location after allocation which ignores the geographic mobility of teachers. At the extreme of this misallocation, LEED expects 8.3% of ESP workers to commute between the North and South Island for work.[23] Overall, therefore, while ESP workers are not identical to non-ESP workers in their commuting patterns, imputation

[22] A very small proportion of jobs are associated with firms where no plant locations are observed. We impute commutes for these jobs based on worker residential address.

[23] Infeasible commutes are counted in the cumulative commute profiles together with commutes of greater than 50km.

is a greatly superior approach to following the LEED allocation.

The performance of the LEED allocation improves only slightly for multi-location firms – a category that includes non-ESP joint-filers and non-joint-filers with multiple employing job locations. Figures 12-14 follow a common pattern plotting cumulative commute profiles estimated from the Census benchmark (solid line), our method (dashed line) and the "IDI method" (dotted line), which uses IDI prioritised addresses and LEED job allocations.[24] In the case of multi-location firms (figure 12), the number of legitimate feasible commute candidates is much smaller than it was with schools, but the LEED allocations still result in 28% of workers estimated to have commutes over 50km, including 7.6% of workers estimated to make inter-island commutes. In comparison, Census indicates 3.6% of multi-location firm workers have commutes over 50km, with only 0.6% having an inter-island commute.

For multi-location firms, we calibrate our method using the Census benchmark and make a probabilistic – rather than deterministic – job allocation. Subject to a feasibility constraint of commutes between TAs on the same island and at most 200km, we find the nearest potential job location and then assume workers are willing to travel to any job location up to ten kilometres further than this shortest commute. Where this process selects multiple job locations, we weight by the relative expected size of the plants (according to LEED). The dashed line in figure 12 shows the resulting (weighted) commute distribution from our method compared to the same set of individuals in Census (solid line) and the IDI method (dotted line). While we overestimate short commutes relative to Census – because not all workers travel to the closest employer location – our method is far superior to the IDI method based on the LEED job allocation.

For single location firms, the allocation decision reduces to simply testing whether the job location meets the commute feasibility test. Given, in this case, we use the same job location as LEED, we see a similar commute profile for our method and the IDI method for jobs in single location firms (figure 13). In both cases, commute distances tend to be overestimated relative to Census, though the median commute distances are similar across the three approaches.

Combining all workers except jobs associated with the ESP joint-filer, we match the Census profile closely, yielding approximately the same distri-

[24]There is no formal "IDI method" for calculating commute distances. We use this terminology because this approach uses the job allocation and prioritised address tables supplied by Stats NZ in the IDI and is, therefore, the obvious naïve approach to calculating commutes using the IDI.

bution of commute distances up to the 60th percentile worker (figure 14). This matching performance comes, in part, from the overestimation of single job location worker commutes balancing the underestimation of multi-location firm commutes. In contrast, the IDI method does poorly in aggregate because of the significant proportion of multi-location firm workers in the workforce.

Aside from performance in matching the Census commute profile, one key difference between our method and the LEED job allocation method is the emphasis placed on matching total employment at each location to the BR expectation. While we use relative plant size to weight probabilistic plant allocations when multiple plants are within close commuting proximity, we prioritise the commute-based allocation rule over matching relative plant sizes. The consequence of this reprioritisation is shown in figure 15, which captures the deviation in aggregate TA employment of our method, relative to the LEED aggregate. The vertical axis captures the total over or undercount, while the horizontal axis picks up the total absolute under/overcount. The dashed cone, therefore, shows the feasible zone where TAs may lie, with TAs close to the line having very few MBs that have counts that run opposite to the overall under/overcount trend in that region.

TAs in the Greater Wellington region are highlighted as a specific example of the trade-off that our method implies. Because we prioritise closer firm locations over more distant ones, workers in regions that feed into Wellington City – Porirua, Kapiti Coast, Lower and Upper Hutt – will appear to work in those regions if their employer has locations in both Wellington and the feeder region where they live. In aggregate then, relative to LEED expectations, these feeder regions have overestimated total employment, while Wellington City has underestimated total employment.

Of course, this presents an issue only if LEED totals provides a better representation of the real world than our allocation, which would rely on the accuracy of BR plant locations and relative sizes. As already discussed, plant locations are somewhat inconsistent with Census data, suggesting measurement error on the BR may be an issue. Figure 16 shows how our method and the LEED target method each compare to Census TA-level total employment for the job location comparison sample. The highlighted TAs confirm the same systematic over/under-estimation for our method when compared to Census, but also suggest that LEED undercounts jobs in the feeder regions (ie, has ratios less than one compared to Census). More generally, LEED target employment is not more consistent with Census than our method. Indeed, overall our method does a better job of matching Census TA-level

employment, as measured by a dissimilarity index, which captures the minimal proportion of workers who would have to be reallocated in order for a measure to produce identical results to Census.[25]

Another way to triangulate the extent of the issue with our method is to examine commute profiles at the TA level. Since the administrative data cover all workers, we can plot TA-specific commute profiles for any region. Figure 17 shows the estimated commute profile for the two largest of the Wellington City feeder regions – Lower Hutt City and Porirua City – together with Wellington City and a "combined region" commute profile for all five TAs. Both Lower Hutt and Porirua show an absence of moderate distance commutes consistent with our method underestimating the number of workers who commute into Wellington City from those regions.

While it might be natural to think that this issue could be remedied by expanding the potential commute distance multi-location firm workers are willing to travel beyond the nearest work location, such an approach does not work for at least two reasons. Firstly, our calibration of the method to Census suggests that increasing the range reduces the overall commute profile match to Census, and markedly increases the commute dataset size. Secondly, while expanding the commute range brings our estimate of Wellington City total employment closer to Census & LEED aggregates, it does this in an undesirable way. Specifically, our method does not allow for predominantly one-way commuting flows. Thus, when we relax the method to enable more commutes from, eg, Lower Hutt to Wellington we also enable more commutes from Wellington to Lower Hutt. Relaxing the commute distance constraint for multi-location firm workers, therefore, moves the total employment counts in each region closer to LEED expectations, but at the cost of much higher bi-directional traffic flows than implied by Census. Therefore, while the overall commute profile for the Wellington region looks adequate to the task of tracking median commute distances (bottom right panel of figure 17), transport planners and modellers may wish to exercise caution when using these data to look at specific flows within the region.

Figures 18 and 19 plot TA-level commute profiles for Auckland and Hamilton City – the latter of which is the subject of a more detailed analysis later in the paper. In both these regions, our method provides a better

[25] The dissimilarity index sums half the absolute difference between two TA-level employment measures divided by the total sample size. Including all TAs, the dissimilarity index for our method is 1.47%, compared to the LEED method value of 1.60%. Excluding Wellington and surrounding regions, the dissimilarity index for our method is 0.99%, compared to the LEED method value of 1.73%.

estimate of total employment (relative to Census) than LEED does. In the case of Hamilton, we also do a good job of matching the regional commute profile out beyond the 90th percentile (figure 19). For Auckland, we do well out to the 50th percentile commute, but overestimate the number of moderate distance commutes of approximately 10km to 30km (figure 18). Since the accuracy of the residential location data is no weaker for Auckland than it is for, eg, Hamilton City (table 8), the relatively poor commute profile performance for Auckland probably reflects the relatively high density of multi-location firms in the region, and the willingness of workers to make commutes beyond the closest firm location.

4 Results

4.1 Commute dataset properties

In the final commute dataset, we think it makes sense to restrict individual-level analysis to the 2005 calendar year onwards, since this limits the proportion of workers for whom we need to rely on the initial IR data spike (treated as December 1999 addresses) for their current residential address, and it also excludes the period of relatively high non-ESP joint-filer job allocation and the discontinuous transition to a lower rate of joint-filing. However, to provide complete job allocation data for the 2005 tax year for business analysis, we extend the commute table back to November 2003.[26] We do not provide commute data further back than this because of our concerns about data quality and consistency, and because of the data storage requirements needed to include more historical data. Summary statistics in this section of the paper relate to January 2005 onwards.

Under our methodology, there are three ways that commute data may be missing for an individual job: an individual has no residential addresses and, therefore, we cannot establish the feasibility of their commute to job locations; employer locations are missing either because the employer has no recorded locations or because we choose not to allocate an employer (ie, ESP workers); or no feasible commute exists. Figure 20 shows the prevalence of these three reasons over time as a proportion of total FTE employment.

[26] November 2003 is the first month required for the financial year ending October 2004 which, under LBD rules, is attributed to the 200503 dim_year_key. October balance dates are extremely rare, with most firms having March or June balance dates requiring starting month job allocations in April 2004 and June 2004, respectively.

The ESP worker group (dark grey bars) accounts for an average of 4.6% of FTE, with a seasonal drop in January. As noted earlier, this group have their commutes imputed from other workers who live in the same MB. Infeasible commutes account for an average of 3.4% of total FTE, rising somewhat from 2014 onwards. Imputation is handled the same way for this group, since we know their residential address. We considered an alternative approach of trying to identify "missing" firm locations that create feasible commutes. However, at least three factors suggest that such an approach is likely to be of marginal benefit at best: the infeasible commute group is small; some of the observed infeasibility comes from residential address inaccuracy; and it is unlikely that any method could triangulate the actual plant location below TA level.

Missing residential addresses (figure 20, light grey bars) account for a negligible proportion of missing commutes because of the coverage of the administrative address data, coupled with the decision to backcast the first residential address spell where necessary. Figure 21 shows the proportion of FTE associated with backcast addresses (black bars), and repeats the missing residential address series from figure 20 (grey bars). Without backcasting address spells, the combined total of these two series would have missing commutes. While the rate of residential address backcasting is not trivial, particularly in the first few years of data, half of the workers who have backcast addresses have those addresses backcast by only one month, with a further quarter of such workers having addresses backcast by at most a year. Given the limited duration of backcasting when it occurs, and the limited use of backcasting overall, backcasting residential addresses is unlikely to have a substantial effect on the consistency of aggregate commute statistics over time (from 2005).

The bottom panel of table 12 summarises the prevalence of the three types of missing commute by the type of EMS filer and the number of potential job locations. As might be expected, infeasible commutes become less likely as the number of candidate job locations increases – accounting for 6% of single location firm employment (L), 1.5% of non-joint-filer multi-location firm L, and 0.6% of joint-filer (non-EMS) L. While this pattern reflects the fact that many multi-location firms have plants spread across major centres in New Zealand, the relatively high rate of infeasible commutes for single location firms may also partly be due to the weakness of the BR in identifying the transition of small firms from single- to multi-location.

The top panel of table 12 shows the size of the final commute dataset in terms of table rows, job months (worker-firm pairs), and total FTE em-

ployment. A consequence of the probabilistic allocation methodology is an increased dataset size from maintaining multiple potential feasible job locations. Overall, the final commute dataset exceeds one billion rows, with an average of 3.25 rows per job month. The number of rows per job increases with the number of candidate job locations, with single location firms having a single row (by construction), non-joint-filer multi-location jobs having an average of 5.3 rows, and joint-filer (non-ESP) jobs averaging 10.6 rows.[27] Thus, while non-joint single and multi-location firms each account for 45% of total FTE employment, non-joint multi-location firms account for 70% of rows in the table.

These statistics emphasise how unlikely it is that a multi-location firm has a single obvious candidate plant location for a worker. Figure 22 demonstrates this point another way by showing the likelihood that our method picks a single unique plant location matching the LEED job allocation. For non-joint-filers (top panel), the proportion of total L with a unique location allocation under our method (dashed grey line) is between 35% and 40%. Of those, around 55% match the LEED allocation (dashed black line), implying that the proportion of non-joint-filer employment where, given multiple choices, we make the same unique allocation as LEED is around 20%. In the case of (non-ESP) joint-filers (bottom panel of figure 22), the equivalent proportion falls below 5%, which reflects both a reduced probability of there being a unique commuting candidate and a reduced probability of that candidate being the same as the LEED allocation. Overall, these probabilities highlight the arbitrary choices that LEED makes in allocating workers to plants, and explain why LEED requires a sticky allocation rule to prevent the appearance of random job mobility.

For joint-filers, the plant allocation also determines the potential firm that employs the worker. Figure 23 shows that the disagreement on job location selection between our method and the LEED method also means that it is unlikely that the LEED firm allocation is correct. On average, less than one third of (non-ESP) joint-filer L is allocated to a single firm (dashed grey line), resulting in between 20% and 30% of L receiving the same unique firm allocation under our method and the LEED method. Furthermore, this analysis excludes ESP workers, since LEED allocates these but we do not. The density of potential school locations (figure 10) is so high that the equivalent firm match rate for our method would be close to zero, reflecting

[27] Even though single location firm jobs have infeasible commutes and we do not allocate ESP workers to jobs, the commute table always contains at least one row per job so that the total (probability-weighted) FTE employment matches that in the F-M labour table.

the impossibility of pinning down a single job location for teachers without further data. Overall, the joint-filer analysis implies that we cannot be certain of the LEED firm allocation for at least 8% of total FTE employment in the F-M labour dataset from 2005 onwards, and even more earlier because of the greater presence of joint-filing in years we have excluded from the commute dataset. The implications of this finding for other work are discussed in the conclusions.

4.2 Commute timeseries

Figures 24 and 25 show average and median commute distance (MB centroid-to-centroid) over the period 2005-2018, weighted by FTE employment. In each case two series are presented, reflecting the relevant statistic for the 92% of employment with observed feasible commutes (solid line), and the same statistic additionally weighted to account for missing commutes due to infeasibility or missing work location, including the imputation of ESP worker job commutes, for the remaining 8% of employment (dashed line). The two approaches produce very similar results for the average commute, and we prefer and use the approach that weights up to the total worker population for the remainder of the paper (which focuses on average commutes).[28] Adjusting for missing commutes increases the median commute (figure 25), since the mean is above the median in most locations.

Both the average and median commute show a similar pattern over time of initially stationary or slowly declining commute distance, followed by more rapidly increasing commute distance. While it is hard to find good corroborating evidence for these trends, the general result of a downward and then upward trend in commute distance appears consistent with figure 6 of Ministry of Transport (2015), though that data series ends in 2014, covers total distance driven per driver per day, and utilises a smoothed four-year moving average which makes it hard to identify turning points. In some sense, the lack of corroborating or dissenting findings explains the value of deriving these statistics from administrative data.

One clear data artefact is present in both average and median series, and that is the impact the arrival of Census residential addresses has on measured commute distances. New Census-based address spells that start in March 2013 imply workers live closer to their jobs than immediately prior

[28] Technically, these statistics exclude jobs where the residential address MB is missing, but these account for only 0.2% of total employment (bottom row, table 12).

administrative addresses implied. We do not think this result implies the time series properties of the data are incorrect, but rather that commute distances have a tendency to be overestimated because some residential address moves are not identified from the administrative data sources.

This issue is less prevalent for median commutes, compared to average commutes, consistent with some of the false commutes being long duration. Figure 26 plots various percentiles of the weighted commute distribution, including the median, showing that the "Census effect" is stronger in absolute terms at the 75th and 90th percentile of the commute distribution, than it is at the 10th and 25th percentile of the distribution. While the scale on figures 24 and 25 make the Census-based revisions appear large, they represent only a 1.8% (5.3%) decline in median (average) estimated commute distance. In relative terms, the 90th (75th) percentile commute distance declines 4.5% (2.6%) after the introduction of Census residential addresses.

All reported percentiles show the same temporal pattern of an initially static or declining commute distance from 2005 to 2009 followed by a rise in commute distance, so that the most recent (early 2018) data represent the highest estimated commute over the entire period. Figure 27 explores whether this temporal pattern is present for the average commute distance of the five largest TAs of worker residence, plus Palmerston North City, which is the only other TA with a main university campus. While there is variation in the strength of the trends, there does appear to be a upward trend for most of the regions, particularly Hamilton City (black dashed line), which we focus on in the next section of the paper. Another feature of the TA-level statistics is the strong seasonal tertiary student effect for Palmerston North and Dunedin, which likely results from student residential addresses not updating between when they are studying and when students are working (elsewhere) over the summer break.

In figure 28 we extend the TA-level analysis to include all TAs, switch to a twelve-month moving average of the monthly commute distance to eliminate seasonality, and normalise each series to 100 in December 2005. This figure confirms the general trend down and up indicated in the nationwide average. Dispersion in this trend demonstrates the value of population time series data that support detailed geographic disaggregation.[29] The following section shows how data on a specific region – Hamilton City – can be analysed in more detail exploiting the longitudinal nature of the firm and worker identifiers.

[29]Outliers in figure 28 are identified to reinforce that each line represents a unique region.

4.3 Sample analysis – Hamilton Urban Area

We conduct three simple analyses to: demonstrate differences between time- and distance-based commute measures by linking in a road network; map spatial commute patterns; and decompose changes in commute time to uncover which types of locational change are driving trend increase in commutes.

We focus on the 2010-2017 period of commuting increase for workers who live in the Hamilton Urban Area (hereafter Hamilton), and who commute to jobs that are within 50km of the Urban Area boundary. The region incorporates Hamilton City, and surrounding semi-urban region as well as Cambridge & Te Awamutu, which are in the commuting zone for Hamilton City. We impose the 50km commute restriction to limit the size of the road network dataset that we need to construct and integrate into the IDI, since using our standard definition of a feasible commute from Hamilton (200km, same island) includes a substantial number of Auckland meshblock locations. On average over the entire period, the 50km region restrictions means that we lose 4.2% of total FTE employment of Hamilton residents (rising from 4% in 2010 and 2011 to 5% in 2016 and 2017). We focus on annual average commutes to abstract from seasonal patterns in jobs, at the cost of ignoring the most recent, but partial, 2018 year.

The available road network data is point-in-time, meaning that changes in the road network aren't captured in the time variation of commutes. Since we use an available road network from February 2015 (Beere 2016), any effect of the Cambridge Expressway – completed in December 2015 – is not captured in estimated network travel distances/times. However, the Expressway may affect residential and/or work location choices if these are influenced by commute times. The research potential of the data increases if time-varying network data become available that account for physical changes in the road network – or changes in the performance of the network (eg, peak travel times) – that are observed by individuals and firms making location choices.

Using the point-in-time road network, figure 29 compares average and median commutes measured using distance – centroid-to-centroid (solid lines) vs road network (dotted lines) – and free-flow travel time (dashed lines, right-hand scale). The leftmost figures show actual distances and times, while the rightmost figures normalise each series to 100 in 2010 to ease comparison of growth rates. The two distance-based measures produce very similar results in terms of commute growth, partly due to the fact that the road network is static, though the road network provides a more plausible (ie, higher) measure of distance. While all series show an increase over time in commutes,

the growth rate of average and median travel time is slower than commute distance implying that trips are being taken on faster roads over time (as demonstrated by the ratio of the two series, figure 30).

Figures 31 and 32 plot the spatial distribution of median commute times for Hamilton City (left panels) and the Hamilton Urban Area (right panels) in 2010 and 2017, respectively. MBs are coded to a common set of quintile commute times, so that areas that darken (lighten) over time reflect residential locations where median commute times are rising (falling). However, while informative about the spatial distribution of commute times, these graphs do not capture the changing density of residential locations and jobs, and it is the combined effect of changing commute patterns and changing density that determine the aggregate commute profile.

Another way to look as these data is to decompose the change in commute time along key decision-making margins for individual workers. For this analysis, we decompose average (rather than median) commute times as these are easier to aggregate across groups. We then focus on the decision to move into or out of the Hamilton labour market, or to move residence and/or job within the region.

Table 13 shows the results of this decomposition, dividing the population into eight distinct groups based on changes in their job and location characteristics between 2010 and 2017. The top two groups in the table are individuals who join or leave the population in 2017. These include individuals who live in Hamilton in both periods, but transition into or out of employment, and individuals who migrate into or out of the region. The remaining six groups are incumbent Hamilton workers, broken down by whether they stay at the same residential address or move, and then by whether they change job (worker-firm pair), have the same job but in a new location, or the same job and in the same location. Because firms can have multiple locations and workers can have multiple jobs, individuals in these categories can be counted in multiple groups, with each worker-job-location combination being FTE-weighted.

The growth in average commute time from 2010 to 2017 was 68 seconds (1.13 minutes) or 7.7% of the 2010 level (bottom row of table 13). To decompose how each group contributes to aggregate growth, we start by determining the average commute of each group in 2010 and 2017. For example, workers who stay in the same job and job location, but move residential address have average commutes of 12.36 minutes in 2010, increasing to 13.45 minutes in 2017 (fifth row of the table). In terms of the aggregate average commute, the change in group employment share is the other factor that

determines the contribution to aggregate commute change. For our example group, the share of total FTE fell from 11.4% in 2010 to 9.8% in 2017.

The next two columns of table 13 take the difference between the year-specific group average and the 2010 population average (14.66), and then weight this by the year-specific FTE-share for that group. By construction, the sum of these weighted contributions is zero in 2010, and equal to the overall growth in average commute in 2017 (1.13 minutes). The difference between the values for 2010 and 2017 (as a percentage of the 2010 mean) is shown in the final column, and is the overall contribution of the group to the total percentage change in commute time.

The net contribution of joiners and leavers is the largest group component, adding 3.7% to average commute times. The gross flows are large also, since joiners travel, on average, over two minutes longer on their commutes than the average worker in 2010 (16.68 vs 14.66 minutes), and because joiners make up 44% of employment in 2017. This large positive contribution is offset by leavers, since they have longer commute than non-leavers in 2010, and make up 38% of employment in 2010. Other contributions are smaller, in part because the overall FTE share of other groups is smaller. Workers who change job and move within Hamilton (third row) contribute 2% of the growth in average commute because they go from having an average commute close to the 2010 mean to a longer commute, while retaining the same overall employment share.

Given the importance of leavers and joiners to the aggregate picture, table 14 decomposes the contribution of these groups further by worker age (young, prime, old) and sex. The bottom row of the table captures the same statistics as the top two rows of table 13. In particular, the rightmost three columns show the same FTE-weighted contribution for leavers, joiners and subgroup total (labelled "net"). In addition, we add average commutes for stayers in each time period by group characteristic to establish whether an unusual average commute time is an artefact of leaving/joining, or of the sorts of workers who leave/join. For example, the middle panel of the table disaggregates by sex. Female and male leavers and joiners have higher average commutes than stayers of the same sex.

However the net contribution to aggregate commute growth is driven by males, who experience a stronger growth in commute time and make up more than half of total employment. Looking at the bottom panel in table 14 it is evident that some of the evolving difference between females and males is the relative growth in average commute time for prime-aged (aged 30-50) males compared to females.

The contributions of young and prime-aged men to commute growth are similar to each other, but arise from different mechanisms. For young men, there is a large excess of joiners over leavers, with long commutes especially for joiners. In contrast, prime-aged men have a particularly large commute difference between stayers and joiners/leavers.

Having established the importance of joiners and leavers in explaining trend changes in commute times in Hamilton, we can then plot the distribution of these worker types to see where they live in the region. Figures 33 and 34 show where these joiners and leavers settle/depart from as a proportion of the base population, and in absolute number terms respectively. On both measures, central Hamilton suburbs to the west of the Waikato river seem to be relatively important locations for joiners and leavers. These last two figures use Statistical Area One (SA1) groupings, rather than MB, because of confidentiality requirements that arise when we consider population subgroups. This issue is important enough in the context of commute patterns to warrant further discussion before we summarise our findings.

4.4 Sparse networks and confidentiality

Results in this subsection explore the feasibility of releasing a full MB-MB origins-destinations dataset under IDI confidentiality rules. At present, the relevant rules are that any released person-level statistic based on tax data must relate to more than five individuals, and that any firm-level statistic must relate to at least three businesses. In analysis of origin-destination (O-D) pairs, the binding constraint is the likelihood that enough individual workers are making the same commute to be able to release the relevant count of commutes.

In addition to these release criteria, noise in the form of random-rounding (to base three) adds additional protection to released worker counts. Random rounding to a fixed base adds proportionately more noise to small cell counts, so we additionally examine the proportion of O-D pair cells with releasable, but small, counts. In particular, in cells with six or more workers, if the average FTE is less than a half then there is potential for the cell total to be less than three and, therefore, for random-rounding to round to zero, implying there are no workers in the cell. Unfortunately, because of probabilistic weighting of potential job locations, weighted FTE for a particular commute is often less than a half.[30]

[30]For example, the average number of rows per job-month in multi-location firms is over five

Table 15 shows the potential loss of O-D pairs and FTE from suppression and random-rounding to zero, either for the single latest full year (2017) or for two years pooled (2016-2017). Even pooling years, suppression would cause the loss of 95% of O-D pairs and 81% of total employment. A further 5.4% of employment is in O-D pair cells that have the potential to be random-rounded to zero.

One-way splits of the commute data are less problematic (bottom two panels of table 15). Even so, we still lose a large proportion of destination MBs, partly because of additional suppression related to firm counts. Mean commute distance by residential MB, however, is subject to minimal FTE and O-D pair loss.

Of course, suppression becomes more restrictive if sub-population analysis is required. For example, a sex-based split of the data will approximately halve commute cell size. A partial solution is aggregation to higher geographies, which is the approach we take in figures 33 and 34 by aggregating to SA1 level. A better solution is using the unconfidentialised O-D within the Datalab to produce aggregations that are safe (eg, one-way medians) and, therefore, less subject to confidentiality constraints – an approach adopted by this paper.

5 Conclusions

We use administrative and survey data in the IDI to probabilistically allocate workers to job locations, enabling the production of over a decade of commute distance population statistics for New Zealand employees. These data show that average commute distance (meshblock centroid-to-centroid) fell from 2005 to 2009 before rising again through to 2018 (the final analysis year), with most regions displaying a similar general temporal pattern. When we study these patterns more closely in the Hamilton region, we find a large contribution of labour market joiners to the increasing average commute.

Census 2013 place of residence and work is used to test our methodology against the alternative of using pre-existing plant allocations from LEED and IDI-prioritised residential addresses. For a consistent set of individuals, our estimate of the commute distance distribution closely matches the corresponding distribution in Census. In contrast, the IDI method estimates tend

(table 12) meaning that even workers who are full time in these firms will tend to have a probability-weighted FTE less than 0.2 for each potential commute.

to significantly overestimate commute distances, including radically overestimating the likelihood of inter-island commuting. Our more plausible results are primarily due to re-engineering the job allocation process, as opposed to exploiting better administrative residential data, though we make marginal improvements to residential address identification through a new prioritisation method, allowing us to use a broader set of residential address sources than available in LEED.

5.1 Future work

One goal of this research was to create a set of tools that would enable further research on commute patterns and the geography of jobs.[31] A useful forward agenda for research might include:

- Estimating commutes by employee "type" (eg, sex, age, ethnicity, education, earnings) – simple characterisation of commute distance gaps and how these have evolved over time (eg, Giménez-Nadal et al. 2020)
- Studying the relationship between commuting and the gender wage gap (eg, Petrongolo and Ronchi 2020; Farré et al. 2020)
- Event studies based on transport infrastructure investment, to assess induced behaviour change and costs/benefits by worker type (eg, skilled vs unskilled) and firm type (eg, by share of transport costs in inputs)
- Development of admin-based commute Labour Market Areas (LMAs) – over time and by employee type (eg, Papps and Newell 2002)
- Analysis of firm access to labour/human capital due to LMA changes (exogenous) and location choices of firms (endogenous), and its impact on firm performance (eg, Maré and Fabling 2012; Maré et al. 2014)
- Evidence on employee sorting into areas with/without good access to jobs (eg, number/quality of jobs within x distance)
- Investigating the impact of access to jobs on, eg, time in unemployment (eg, Andersson et al. 2018)

On the data development side it makes sense to update the data to add new sources, particularly Census 2018 (now available in the IDI).[32] At the

[31] The codebooks in the appendix are intended to facilitate further use of the data.

[32] In the latest IDI instance (IDI_20200120), other new sources provide relatively few residential addresses, coming from Housing New Zealand (tenancy) and the Department of Internal Affairs (births, deaths, marriages and civil unions).

simplest level, updating the data would provide more recent commute statistics (up to September 2019 in the latest version of the F-M labour tables). However, inclusion of a second Census in the IDI also enables additional testing on how well our method tracks address changes over time, and an assessment of whether administrative address quality is improving over time.

Another avenue for further data development would be to extend the dataset to include self-employed, or at least the subset of self-employed with employees. Self-employed are a sizeable proportion of total employment (around 20% according to Fabling 2018), and are likely to have different commute patterns from individuals who work as employees. The main concern with including working proprietors (WPs) in the current paper is the possibility that WPs may appear to work from home because the BR records their business address as their home address if that is the postal address for contact with IR. Conversely, many WPs may actually work from home, and including them in the population of interest would tend to reduce the estimated average/median commute. A first feasibility step for expanding the population would be to compare work from home patterns in the Census and administrative data for the self-employed.

5.2 Implications for other work

Because of the general importance of worker-firm links and job location information in IDI-related labour research, the findings in this paper have a number of implications for other work.

The inaccuracy of the LEED linking impacts the quality of the labour and productivity datasets. Since these data have been designed with an expectation that the plant-level allocation of workers is not usable, most of the potential issues relate to the incorrect allocation of firms for joint-filers. For each of the current labour-productivity technologies, the primary issues are:

- PENT (Fabling 2011): Repairs to enterprise numbers are based on employee tracking at the plant level. Random allocation of workers to plants in LEED may result in under-identification of continuing plants, though imposed stickiness of job locations mitigates this issue. Ideally, employee tracking directly at the firm level would be the best way to address these issues. Since we allocate workers only to LEED employing plants, our allocation produces aggregate plant employment patterns that are consistent with firm continuity rules

- F-M labour tables (Fabling and Maré 2015a): Two-way fixed effects and other applications that rely on knowing who works in which firm are affected by misallocation for joint-filers. These issues are exacerbated if the change in joint-filer prevalence over time is due to reduced maintenance of the LEED system, since the absence of true joint-filing relationships also causes misallocation
- Firm plant size on LBD (Fabling and Maré 2015a): Plant-level aggregated FTE from the F-M labour tables facilitates analysis of the geography of business. These counts flow directly from target BR counts used by LEED, which our revised method suggests may be inaccurate
- Productivity data (Fabling and Maré 2015b, 2019) Two-way fixed effects are used as a skills proxy in the productivity dataset, so are affected by joint-filer misallocation. The restriction to the private-for-profit sector in productivity analysis removes ESP schools, halving any joint-filer issues. Cleaning steps for the productivity data remove firms where LEED gross earnings are inconsistent with annually reported wages above a threshold value. It is possible that joint-filing misallocation is partly responsible for this discrepancy[33]

Since our method prioritises residential address quality, it may be convenient in some labour market applications (eg, wage equation estimation) to simply use residential address information to control for geography.

Our work also has implications for the use of IDI residential address data to identify residential moves, and to identify households. On the former, the clear implication of our work is that using naïve changes in x-y coordinates or MB to indicate a residential move is not a good idea, because of the adjacency issue. Similarly, use of residential address x-y to identify households is potentially a bad idea unless supported by other evidence, since x-y has spurious variation for many individuals, implying that grouping individuals on the basis of a unique x-y may separate families and combine unrelated individuals.[34]

[33] It is also possible that the discrepancy derives from differences between the (time-varying) IR-ENT link in LEED, compared to the (static) current BR link, the latter of which is used to integrate tax returns into the LBD.

[34] Conceptually, co-location may not be indicative of a household, particularly at addresses that contain multiple dwellings. For example, Gath and Bycroft (2018) use a shared x-y address as a potential indicator of a household and find a significant overcount of large households (compared to Census), and that less than 50% of implied households match Census at the individual level. This lack of agreement will reflect errors in x-y (they use an earlier IDI instance that has a different address coding tool), but also the fact that high turnover addresses (eg, rental properties) can appear to have people living together

We also find that both our and the IDI method are much worse at identifying the correct address of recent (within a year) residential movers than long-term residents (see figure 6), suggesting caution should be taken when using geographic data for specific sub-populations, such as itinerant workers, or when conducting analysis that relies on the precise timing of residential address moves.

Finally, there are at least two potential implications for Stats NZ. Firstly, the IDI uses x-y coordinates as a linking variable in some linking projects (eg, Household Labour Force Survey to spine) and, in other cases, uses residential MB (eg, Census to spine). In both cases our analysis of the residential address data suggests that adjacent MB is a more appropriate matching variable. Secondly, our analysis raises questions about the accuracy of TA-level LEED statistics, including mean and median earnings. Additionally, since LEED worker accession and separation rates are measured at the plant level, these are almost certainly likely to be different in the absence of the imposed stickiness methodology. Our method provides an alternative dataset from which these statistics could be derived.

who are actually sequentially living at an address, if the administrative data do not pick up address changes in a timely manner. The IDI address prioritisation is particularly bad at identifying recent address changes.

References

Andersson, F., J. C. Haltiwanger, M. J. Kutzbach, H. O. Pollakowski, and D. H. Weinberg (2018). Job displacement and the duration of joblessness: The role of spatial mismatch. *The Review of Economics and Statistics 100*(2), 203–218.

Beere, P. (2016). Creating a road network analysis layer with travel time estimates using open-source data. Research paper, GeoHealth Laboratory, University of Canterbury.

Fabling, R. (2011). Keeping it together: Tracking firms in New Zealand's Longitudinal Business Database. Working Paper 11-01, Motu Economic and Public Policy Research.

Fabling, R. (2018). Entrepreneurial beginnings: Transitions to self-employment and the creation of jobs. Working Paper 18-12, Motu Economic and Public Policy Research.

Fabling, R. and D. C. Maré (2015a). Addressing the absence of hours information in linked employer-employee data. Working Paper 15-17, Motu Economic and Public Policy Research.

Fabling, R. and D. C. Maré (2015b). Production function estimation using New Zealand's Longitudinal Business Database. Working Paper 15-15, Motu Economic and Public Policy Research.

Fabling, R. and D. C. Maré (2019). Improved productivity measurement in New Zealand's Longitudinal Business Database. Working Paper 19-03, Motu Economic and Public Policy Research.

Fabling, R. and L. Sanderson (2016). A rough guide to New Zealand's Longitudinal Business Database (2nd edition). Working Paper 16-03, Motu Economic and Public Policy Research.

Farré, L., J. Jofre-Monseny, and J. Torrecillas (2020). Commuting time and the gender gap in labor market participation. Discussion Papers 13213, Institute of Labor Economics (IZA).

Gath, M. and C. Bycroft (2018). The potential for linked administrative data to provide household and family information. Census Transformation Papers, Statistics New Zealand.

Gibb, S. and S. Das (2015). Quality of geographic information in the Integrated Data Infrastructure. Census Transformation Papers, Statistics New Zealand.

Giménez-Nadal, J. I., J. A. Molina, and J. Velilla (2020). Trends in commuting time of European workers: A cross-country analysis. Discussion Papers 12916, Institute of Labor Economics (IZA).

Maré, D. C. and R. Fabling (2012). Productivity and local workforce composition. In R. Crescenzi and M. Percoco (Eds.), *Geography, Institu-*

tions and Regional Economic Performance, Advances in Spatial Science Series, pp. 59–76. Berlin and New York: Springer.

Maré, D. C., R. Fabling, and S. Stillman (2014). Innovation and the local workforce. *Papers in Regional Science 93*(1), 183–201.

McLeod, K. (2018). Where we come from, where we go: Describing population change in New Zealand. Analytical Paper 18/02, New Zealand Treasury.

Ministry of Transport (2015). 25 years of New Zealand travel: New Zealand household travel 1989-2014. Technical report, Ministry of Transport.

Papps, K. L. and J. O. Newell (2002). Identifying functional labour market areas in New Zealand: A reconnaissance study using travel-to-work data. Discussion Papers 443, Institute of Labor Economics (IZA).

Petrongolo, B. and M. Ronchi (2020). Gender gaps and the structure of local labor markets. Discussion Papers 13143, Institute of Labor Economics (IZA).

Stats NZ (2013). Evaluation of administrative data sources for subnational population estimates. Technical report, Statistics New Zealand.

Stats NZ (2017). Experimental population estimates from linked administrative data: 2017 release. Census Transformation Papers, Statistics New Zealand.

Stats NZ (2018). Metadata – Geospatial information in the IDI. Mimeo, Statistics New Zealand. Available on the IDI Wiki.

Stats NZ (2019). Guide to interpreting the LEED data. downloaded on 7 Feb 2019, Statistics New Zealand. http://archive.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/guide-interpreting-the-leed-data/leed-processes.aspx.

Stats NZ (2020). Integrated Data Infrastructure (IDI) refresh: Linking report for December 2019 refresh. Mimeo, Statistics New Zealand. Available on the IDI Wiki.

Tables

Table 1: Census testing sample size

		N(workers)
Employed in March 2013		1,811,928
Population loss:		
Missing age	1,218	
Younger than 16	9,561	
Older than 79	1,644	
Not in Census usually resident	236,265	
Total loss	**248,688**	
Census testing sample (residential address)		1,563,240
Population loss (Census):		
Unusable response to main job question	8,088	
Main job is not a paid employee	145,860	
Worked at home	97,896	
Census work address MB missing	174,660	
Population loss (labour table):		
Multi-job worker	59,775	
Job start or job end month	48,150	
Self-employed (2013 or 2014 tax year)	39,402	
Population loss (residential address):		
Multiple prioritised addresses	633	
Total loss	**574,464**	
Census testing sample (commute patterns)		988,776

The number of workers is derived from the Fabling-Maré labour tables. We only count Census links in the IDI that include a usually resident address with corresponding meshblock that is located within a Territorial Authority (TA). The table shows a particular sequential accounting of population loss and counts would differ for the job harmonisation if the labour table restrictions were applied first. While the testing method produces tied prioritised residential addresses (final row of population loss disaggregation), these are prevented in the final residential address methodology.

Table 2: Potential for address source to agree with Census residence, from April 2012 to March 2013

Assigned tier	Address source	N(workers)	p(any MB match to Census) Exact	Adjacent	Difference
	Coded to *x*-*y*				
1	Accident Compensation Corporation	365,985	0.754	0.768	0.014
1	Inland Revenue	576,972	0.769	0.783	0.014
1	Ministry of Health (National Health Index)	413,934	0.807	0.820	0.013
1	Ministry of Health (Primary Health Organisation)	343,113	0.824	0.837	0.013
1	Ministry of Social Development (residential)	158,874	0.793	0.805	0.013
1	NZ Transport Agency (drivers licence)	63,873	0.913	0.920	0.007
1	NZ Transport Agency (motor vehicle registration)	49,986	0.943	0.950	0.007
	Coded to MB, but not *x*-*y*				
2	Accident Compensation Corporation	6,510	0.448	0.565	0.117
2	Inland Revenue	15,228	0.400	0.497	0.097
2	Ministry of Health (National Health Index)	11,169	0.504	0.604	0.100
2	Ministry of Health (Primary Health Organisation)	9,264	0.525	0.632	0.107
2	Ministry of Social Development (residential)	2,070	0.416	0.512	0.096
2	NZ Transport Agency (motor vehicle registration)	462	0.812	0.877	0.065
	All addresses for source, regardless of *x*-*y* coding				
2	Ministry of Education	3,030	0.631	0.650	0.019
2	Ministry of Social Development (postal)	11,730	0.434	0.458	0.024
	All addresses pooled by tier				
1	Tier one addresses	1,086,699	0.879	0.889	0.011
2	Tier two addresses	52,266	0.466	0.547	0.081

Comparison to Census (2013) uses original Census dataset meshblock (MB13), not IDI recalculated meshblock (MB18). Administrative addresses are concorded from MB18 to MB13, and are taken from `address_notification_full`, except Inland Revenue addresses which are sourced directly from the IR schema to reintroduce address dates excluded by Stats NZ. An adjacent match occurs when the administrative MB or an adjacent MB, matches the Census MB or an adjacent MB. This allows for coding error in each source and implies that, in the case of an adjacent match, the coded administrative and Census MBs may be non-adjacent. The final column, labelled "Difference," is the difference between the exact and adjacent rate. Since all exact matches are also adjacent matches, the difference is correlated with the rate of coding (to MB) error in a particular source, which may be due to quality issues in the raw address data.

Table 3: Potential for residential address MB to recur with new x-y

Address source	N(workers)	p(address MB recurs)	p(recurring address has multiple x-y)
ACC	1,289,139	0.852	0.014
Inland Revenue	1,551,498	0.972	0.137
Ministry of Health (NHI)	1,526,817	0.790	0.121
Ministry of Health (PHO)	1,507,716	0.790	0.126
MSD (residential)	1,111,833	0.303	0.242
NZTA (drivers licence)	1,415,694	0.010	0.002
NZTA (motor vehicle)	980,169	0.019	0.037

Restricted to tier one observations. By construction, tier two addresses for these data sources were not coded to x-y coordinates. Since tier two addresses may point to the same MB as tier one addresses, the probability of an address recurring is underestimated and any such repetition likely reflects a variant of the same real address appearing in the raw administrative data.

Table 4: Group size distribution for residential MB adjacency groups

N(adjacent MBs in group)	N(workers)	Proportion of MB groups
2	338,322	0.862
3	48,135	0.107
4	10,161	0.022
5	2,817	0.006
6+	1,593	0.003
All groups	**378,957**	**1.000**

The "all groups" worker count is smaller than the summed sub-groups because 18% of workers with MB groups have more than one group. Consequently, the proportion column – which counts each MB group separately – reports different proportions to those derived from the worker counts.

Table 5: Potential for within-source address date ties to agree with Census

Address type	N(workers)	p(any MB match to Census)
Tier 1		
Recurring MB address	328,110	0.503
Unique MB address	109,503	0.023
Tier 2		
Recurring MB address	3,054	0.127
Unique MB address	17,394	0.017

A recurring MB address appears as a tier one address (across all tier one data sources) on a non-tied date. Conversely unique MB addresses never appear as tier one addresses on a non-tied notification date.

Table 6: Impact of data cleaning on number of tier one residential address notifications

	N(address notifications)			Loss from data cleaning step					
Address source	Raw	Clean	Tier 1 clean	Spike	Tie	Recur	Tier 2	Other	Total
ACC	20,728,824	4,134,147	4,058,355	0.000	0.001	0.800	0.004	0.000	0.804
Census 2013	2,862,132	2,861,826	2,849,094	0.000	0.000	0.000	0.004	0.000	0.005
Inland Revenue	36,243,258	16,173,123	15,817,287	0.162	0.009	0.381	0.010	0.001	0.564
Ministry of Health	38,163,663	15,149,631	14,888,163	0.105	0.020	0.474	0.007	0.004	0.610
Ministry of Social Development	12,593,730	10,144,803	9,758,067	0.000	0.010	0.145	0.031	0.039	0.225
NZTA (drivers licence)	3,636,213	3,601,371	3,601,371	0.000	0.000	0.010	0.000	0.000	0.010
NZTA (motor vehicle)	2,161,689	2,088,486	2,065,569	0.000	0.000	0.034	0.011	0.000	0.044
Ministry of Education	272,973	270,918	0	0.000	0.007	0.000	0.992	0.000	1.000
Total	**116,662,482**	**54,424,305**	**53,037,906**	**0.085**	**0.011**	**0.432**	**0.012**	**0.006**	**0.545**

The table uses the following shorthand for the methods applied to clean the administrative residential address data: "spike" – the removal of recurring addresses from data sources when a large proportion of addresses are recurring on that date; "Tie" – the removal on non-adjacent addresses reported on the same day within the same data source; "recur" – the removal of addresses that repeat the same address as last notified within the same source; "tier 2" – the demoting of addresses based either on source (MSD postal or Ministry of Education), or on an inability for the address to be coded to x-y coordinates; "other" – other minor cleaning steps. High rate of "other" for MSD arise from the deduplication of addresses when postal and residential notifications are combined following the adjacency step, consistent with postal and residential address files both being updated from the same underlying data source. By construction, all Ministry of Education notifications are assigned to tier two resulting in the tier one clean count being zero.

Table 7: Prioritised residential address match rate to Census by admin source

Address source	Proportion of prioritised addresses	Match rate to Census MB	TA
ACC	0.069	0.870	0.966
Inland Revenue	0.283	0.804	0.943
Ministry of Health	0.300	0.837	0.957
Ministry of Social Development	0.155	0.850	0.944
NZTA (drivers licence)	0.129	0.937	0.983
NZTA (motor vehicle)	0.065	0.962	0.989
Total	**1.000**	**0.853**	**0.957**

Tier two addresses from MSD (postal) and Ministry of Education address sources are excluded as they are rarely used under the prioritisation rules and small counts are subject to substantial perturbation, or suppression, under Stats NZ confidentiality rules.

Table 8: Prioritised residential address match rate by Census TA

Territorial Authority	N(workers)	Match rate to Census MB	TA
Auckland	512,226	0.854	0.983
Christchurch City	142,305	0.843	0.959
Wellington City	85,536	0.821	0.934
Hamilton City	54,879	0.844	0.943
Dunedin City	46,950	0.855	0.960
Tauranga City	39,225	0.851	0.953
Lower Hutt City	38,679	0.881	0.956
Palmerston North City	31,833	0.841	0.936
New Plymouth District	28,158	0.864	0.967
Hastings District	26,523	0.857	0.946
Whangarei District	24,054	0.861	0.964
Rotorua District	22,497	0.865	0.961
Invercargill City	22,023	0.879	0.966
Napier City	21,459	0.862	0.952
Waikato District	20,787	0.860	0.909
Total of above (N>20K)	**1,117,134**	**0.851**	**0.966**
Total (10 TAs with N∈15-20K)	174,999	0.865	0.937
Total (11 TAs with N∈10-15K)	133,926	0.847	0.940
Total (12 TAs with N∈5-10K)	81,063	0.864	0.940
Total (19 TAs with N<5K)	56,115	0.838	0.911

Territorial Authority size groups based on the total number of workers (N) in the Census comparison, not the total number of workers in the TA.

Table 9: IDI prioritised ranking of residential addresses

Address source	IDI tier	IDI rank
Census	1	1
Ministry of Social Development (residential)	1	2
Ministry of Health (Primary Health Organisation)	1	3
Ministry of Health (National Health Index)	1	4
NZ Transport Agency (motor vehicle registration)	1	5
Accident Compensation Corporation	2	1
Inland Revenue	2	2
Ministry of Social Development (postal)	2	3
Ministry of Education	2	4
NZ Transport Agency (drivers licence)	2	6

Source: IDI documentation available within the secure Datalab environment (Stats NZ 2018).

Table 10: Comparison of prioritisation methods by Census usual residence

		Match rate with Census				Gain over	
		Our method		IDI method		IDI method	
Usual residence	N(workers)	MB	TA	MB	TA	MB	TA
Census night address	1,519,134	0.857	0.960	0.831	0.949	0.026	0.011
Elsewhere in NZ	37,026	0.704	0.876	0.699	0.875	0.005	0.001
No fixed abode	261	0.287	0.552	0.276	0.540	0.011	0.011
Total	**1,556,421**	**0.854**	**0.957**	**0.828**	**0.947**	**0.026**	**0.011**

Comparison uses IDI prioritised addresses immediately preceding Census day, since the IDI address table includes Census. We restrict comparison to the common sample of workers who have a unique prioritised MB under each method, which eliminates individuals with no pre-Census address in the IDI address table, and individuals with multiple prioritised MB addresses (eg, because an individual has two sources with cleaned tier one notifications on the same day). For the latter case, since at most one such address can match Census, removal of these ties results in a slightly higher overall MB match rate than that shown in table 7. Our final methodology drops the small number of exact date MB conflicts across same-tier sources, and we rely on other dates to determine the best guess residential address.

Table 11: Potential for LEED employing work location to include Census

	N(workers)	p(LEED has Census)	p(LEED within x of Census) 1km	5km
Corporate Managers	144,735	0.823	0.865	0.922
Personal & Protective Services Workers	91,218	0.806	0.838	0.904
Other Associate Profs.	90,921	0.793	0.843	0.915
Office Clerks	80,028	0.833	0.872	0.930
Salespersons, Demonstrators & Models	57,345	0.857	0.887	0.935
Other Profs.	54,384	0.800	0.889	0.941
Labourers & Related Elementary Service Workers	49,296	0.801	0.817	0.889
Physical, Math & Engineering Science Profs.	41,985	0.791	0.859	0.930
Life Science & Health Profs.	41,304	0.911	0.931	0.965
Customer Services Clerks	38,601	0.860	0.903	0.949
Stationary Machine Operators & Assemblers	32,790	0.868	0.879	0.932
Teaching Profs.	29,559	0.880	0.905	0.948
Physical Science & Engineering Associate Profs.	29,175	0.842	0.873	0.929
Metal & Machinery Trades Workers	25,668	0.866	0.881	0.931
Building Trades Workers	22,443	0.793	0.808	0.876
Drivers & Mobile Machinery Operators	20,997	0.803	0.814	0.876
Market Oriented Agricultural & Fishery Workers	19,584	0.821	0.827	0.861
Response Unidentifiable	18,546	0.811	0.867	0.933
Life Science & Health Associate Profs.	11,214	0.854	0.878	0.937
Industrial Plant Operators	8,139	0.837	0.844	0.899
Other Craft & Related Trades Workers	7,524	0.851	0.874	0.926
Legislators & Administrators	4,920	0.818	0.871	0.926
Precision Trades Workers	4,080	0.866	0.881	0.936
Response Outside Scope/Not Stated	3,798	0.813	0.845	0.908
Building & Related Workers	3,276	0.808	0.817	0.878
Total	**931,533**	**0.827**	**0.864**	**0.922**

Potential LEED job locations include all employing plants in March 2013, not just the observed LEED allocation, nor just those MBs associated with feasible commutes. Inactive Business Register locations not receiving a LEED allocation are excluded from the potential match to Census. Education Service Payroll jobs are excluded from the analysis. Non-ESP joint-filers are included, with potential work locations including the employing MBs of every firm subject to the joint-filing. Distance between LEED and Census MB is centroid-to-centroid.

Table 12: Commute data missingness by job allocation type

		Non-joint-filer by N(locations)			Joint-filer	
	Total	Zero	One	Two or more	Non-ESP	ESP
N(Commute table rows)	1,017,356,637	93,918	150,921,213	710,915,874	141,200,481	14,225,292
N(Job months)	312,827,820	93,918	150,921,213	134,239,233	13,348,212	14,225,292
Total FTE employment (L)	249,222,993	71,856	113,090,967	112,962,426	11,627,979	11,469,738
Average rows per job month	3.252	1.000	1.000	5.296	10.578	1.000
Proportion of total rows	1.000	0.000	0.148	0.699	0.139	0.014
Proportion of total L	1.000	0.000	0.454	0.453	0.047	0.046
Proportion of within-group L with:						
Observed feasible commute	0.918	0.000	0.938	0.983	0.993	0.000
No feasible commute	0.034	0.000	0.060	0.015	0.006	0.000
Employer address missing	0.046	1.000	0.000	0.000	0.000	1.000
Residential address missing	0.002	0.018	0.002	0.001	0.001	0.001

Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Number of locations counts only LEED employing plants with non-missing meshblock. Restricted to January 2005 onwards. Commute table extends back to November 2003 to accommodate 2005 financial year analysis of business. The complete table contains 1,101,957,864 rows covering 337,650,531 job months. In the case of ESP workers, both the employer address and the firm identifier (PENT) are missing. For non-ESP jobs, if there is no feasible commute then each potential work location has a separate row in the table (with missing commute distance). If there is one or more feasible commutes, only commutes within 10km of the shortest commute are represented in the data.

Table 13: Decomposition of change in average Hamilton commute time by worker location transition type

	Average commute		FTE share		FTE-weighted contribution to deviation from 2010 average		
	2010	2017	2010	2017	2010	2017	Difference (% of 2010)
Joined Hamilton job population	N/A	16.68	0.000	0.440	0.00	0.89	6.1%
Left Hamilton job population	15.57	N/A	0.378	0.000	0.34	0.00	-2.3%
					Subtotal		**3.7%**
Changed residential location							
New job	14.58	16.37	0.166	0.168	-0.01	0.29	2.0%
Same job, changed location	14.17	14.02	0.019	0.016	-0.01	-0.01	0.0%
Same job, same location	12.36	13.45	0.114	0.098	-0.26	-0.12	1.0%
					Subtotal		**3.0%**
Same residential location							
New job	15.78	17.12	0.132	0.117	0.15	0.29	1.0%
Same job, changed location	16.12	15.31	0.021	0.016	0.03	0.01	-0.1%
Same job, same location	13.26	13.13	0.170	0.145	-0.24	-0.22	0.1%
					Subtotal		**0.9%**
Total	**14.66**	**15.78**	**1.000**	**1.000**	**0.00**	**1.13**	**7.7%**

A job is a firm (pent)-worker pair, not a specific role within a business. Individuals can be counted (on an FTE-weighted basis) as both a stayer (same residential address) and a mover (changed residential address) if they move during a reference year. Similarly individuals can contribute to multiple job states because of mid-year job changes, and because firms can have multiple locations and workers can have multiple jobs. The average commute (first two columns) for individuals who stay in the same residence and job/job location (last subgroup in table) is not identical across years because of changes in the probabilistic FTE weighting of job locations over time. The join/leave the population contribution combines a migration effect with the effect of Hamilton residents transitioning into/out of employment.

Table 14: Net contribution of leavers & joiners to change in average Hamilton commute time by worker sex & age

	Average commute				FTE share		FTE-weighted contribution to change (% of 2010)		
	2010		2017						
	Left	Stayed	Joined	Stayed	Left	Joined	Left	Joined	Net
Young (<30)	15.84	13.84	16.37	15.35	0.356	0.462	-1.1%	2.4%	1.3%
Prime (30-50)	15.72	14.10	16.49	15.17	0.358	0.396	-1.0%	2.2%	1.2%
Old (>50)	15.03	14.36	18.25	14.87	0.286	0.143	-0.3%	1.5%	1.3%
Female	14.23	13.06	15.22	13.90	0.466	0.447	0.5%	0.8%	1.3%
Male	16.73	14.99	17.87	16.08	0.534	0.553	-2.9%	5.3%	2.5%
Female									
Young (<30)	14.58	12.92	15.07	14.28	0.165	0.201	0.0%	0.3%	0.3%
Prime (30-50)	14.27	13.09	14.92	14.07	0.161	0.181	0.2%	0.1%	0.3%
Old (>50)	13.76	13.10	16.46	13.60	0.140	0.066	0.3%	0.4%	0.7%
Male									
Young (<30)	16.92	14.50	17.35	16.13	0.191	0.261	-1.1%	2.1%	1.0%
Prime (30-50)	16.91	14.98	17.82	16.06	0.197	0.214	-1.1%	2.0%	0.9%
Old (>50)	16.23	15.55	19.75	16.08	0.146	0.077	-0.6%	1.2%	0.6%
Total	**15.57**	**14.11**	**16.68**	**15.08**	**1.000**	**1.000**	**-2.3%**	**6.1%**	**3.7%**

See table 13 notes. The final column is directly comparable to the FTE-weighted subtotal contribution of joiners and leavers (3.7%). Age is measured as at the year of interest, meaning "stay" group members may change age category between 2010 and 2017. These categories are included as reference points for understanding whether variation in average commute time across groups may be due to inherent characteristics of the group, or characteristics specific to leavers/joiners.

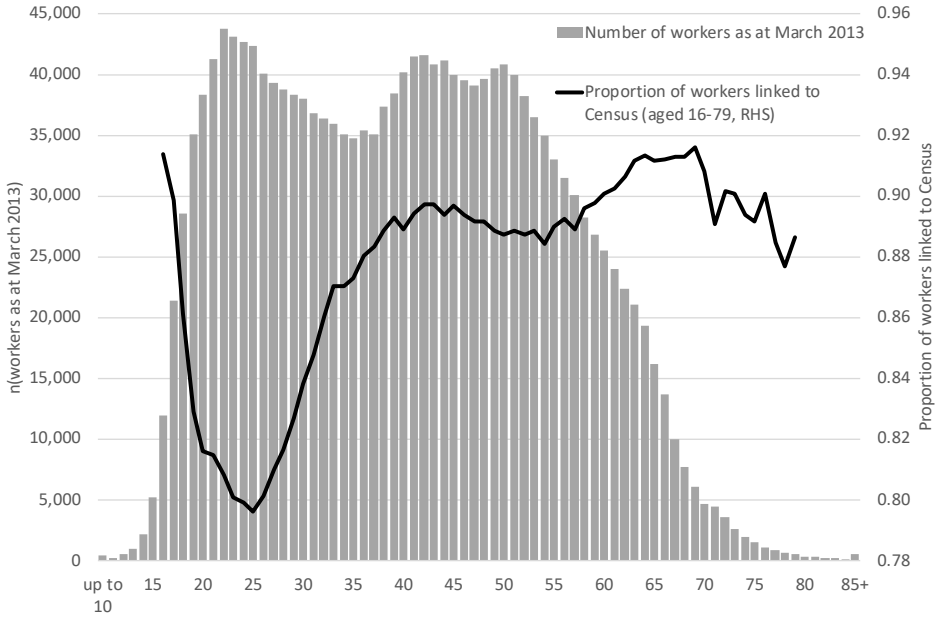
Table 15: Loss from origin-destination dataset due to confidentiality (suppression) rules

	2017		2016-2017 pooled	
	N(O-D pair)	Total FTE	N(O-D pair)	Total FTE
Origin-destination (O-D) data	4,983,678	1,740,279	6,346,155	3,421,662
and with 6+ workers	175,744	241,692	302,317	645,030
and with FTE≥3	25,474	133,533	71,772	462,366
Proportion lost (suppression)	**0.965**	**0.861**	**0.952**	**0.811**
Proportion lost (suppression + potential zero)	**0.995**	**0.923**	**0.989**	**0.865**
	N(Origin)	Total FTE	N(Origin)	Total FTE
Origin data	50,777	1,740,279	50,962	3,421,662
and with 6+ workers	47,827	1,735,386	48,572	3,415,605
Proportion lost (suppression)	**0.058**	**0.003**	**0.047**	**0.002**
	N(Dest.)	Total FTE	N(Dest.)	Total FTE
Destination data	40,630	1,740,279	41,978	3,421,662
and with 6+ workers	30,048	1,726,770	32,653	3,403,947
and with 3+ firms	20,447	1,587,633	22,726	3,158,535
Proportion lost (suppression)	**0.497**	**0.088**	**0.459**	**0.077**

Table shows the effect of applying IDI-specific confidentiality (suppression) to the hypothetical release of a full origin-destination dataset. Relevant IDI output rules are: [5.15] "Suppression under 6" for all MB-level outputs; and [5.14] "Output relating to business..." for business tax data. In addition to suppression, all counts of individuals and businesses are random-rounded to base three. Random-rounding means that FTE totals less than three have a probability of being rounded to zero ("potential zero" category) implying no labour flow between an origin-destination pair. These confidentiality rules, and others, have been applied to all outputs in this paper to protect workers and firms from identification.

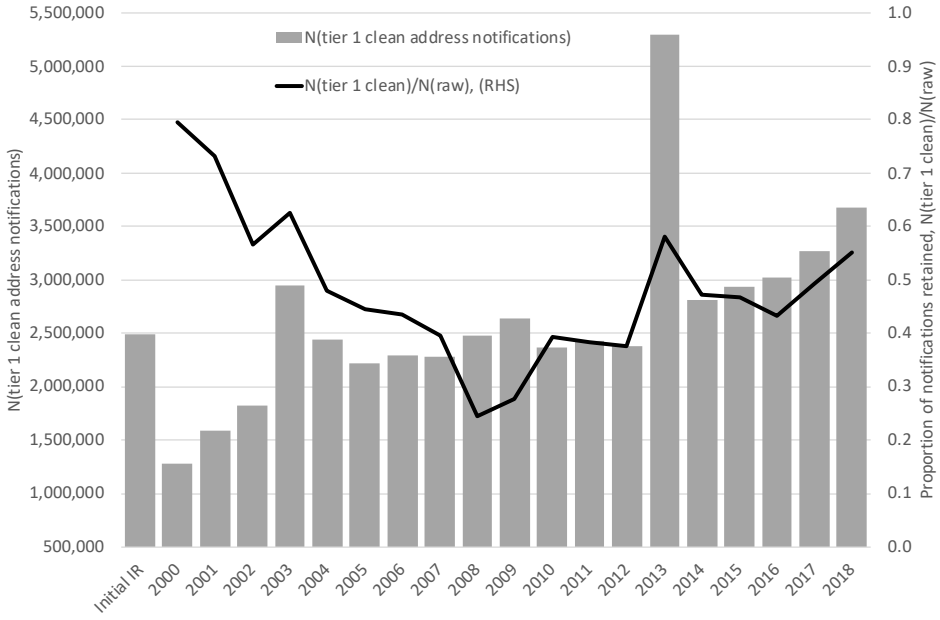
Figures

Figure 1: Census usually resident coverage of workers by age



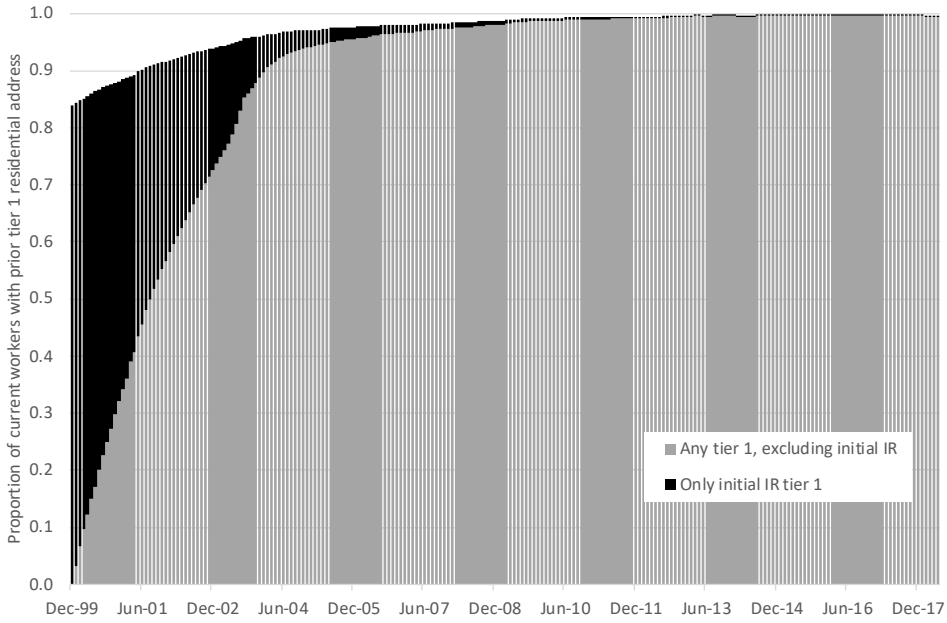
The number of workers is derived from the Fabling-Maré labour tables, using only individuals with non-missing age and sex. We only count Census links in the IDI that include a usually resident address with corresponding meshblock that is located within a Territorial Authority (TA).

Figure 2: Number of residential address change notifications by year



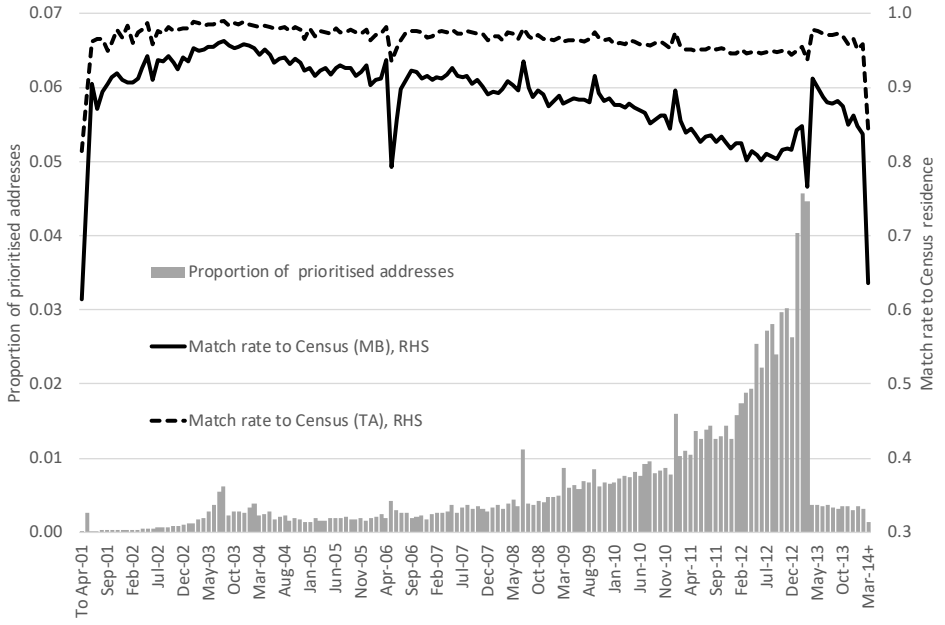
"Initial IR" refers to cleaned addresses present in the initial IR spike. Because the volume of these addresses implies that many of them relate to earlier notification dates, they are all moved to a notification date of December 1999 to allow other tier one data sources to contribute more recent address information. The retention rate for these observations is incorporated into the year of observed "notification date" (2001), since the relevant numerator to calculate a separate retention rate is unknown.

Figure 3: Proportion of current workers with prior tier 1 residential address



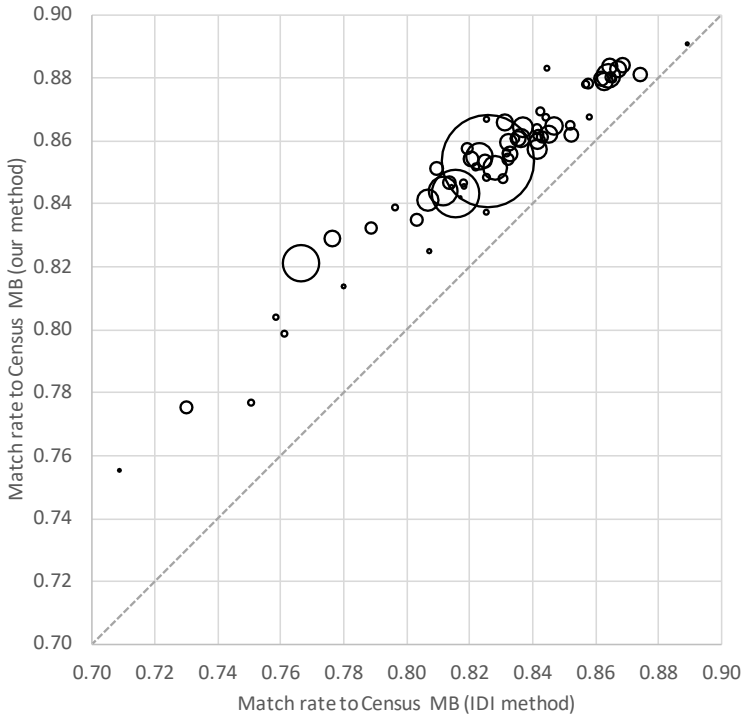
LEED jobs are attributed to the 15th of the month to establish the presence of a prior residential address. Administrative address data is available in the IDI from January 2000, and we move initial IR spike data to a notional December 1999 address, so that no employee can have a prior residential address earlier than December 1999.

Figure 4: Prioritised administrative residential address match rate to Census



For testing, the initial IR spike remains at it's recorded notification date (May 2001). As a consequence, very few tier one addresses prior to May 2001 are prioritised, and we pool these addresses for presentational purposes. Similarly, we pool addresses from April 2014 onward, as the prioritisation rules results in few addresses after March 2014 being used as the best guess address for March 2013.

Figure 5: Comparison of prioritisation methods by territorial authority



Bubble area scaled by the total number of common sample workers in the TA for the Census comparison, not the total number of workers in the TA.

Figure 6: Comparison of prioritisation methods by years at usual residence

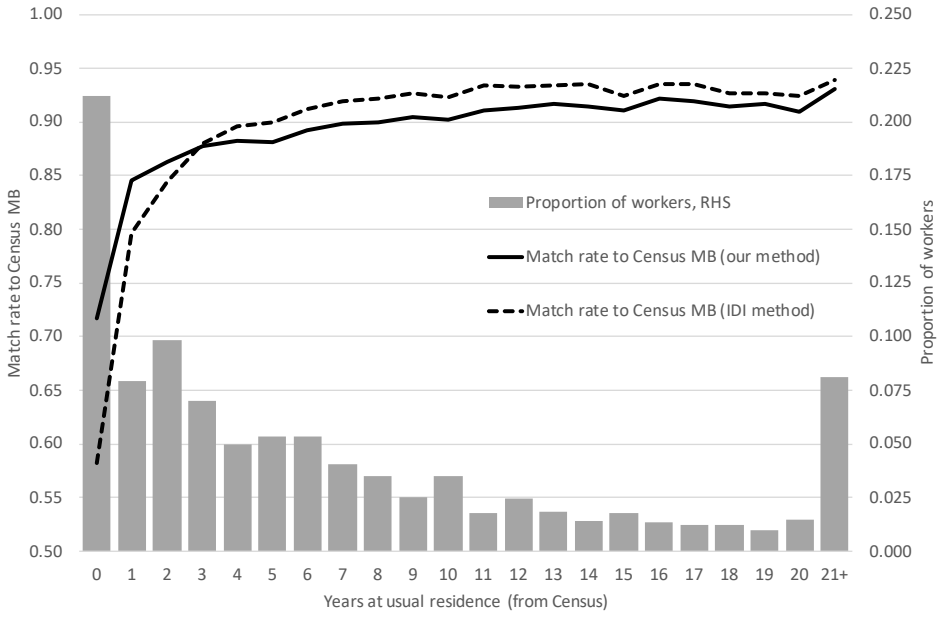
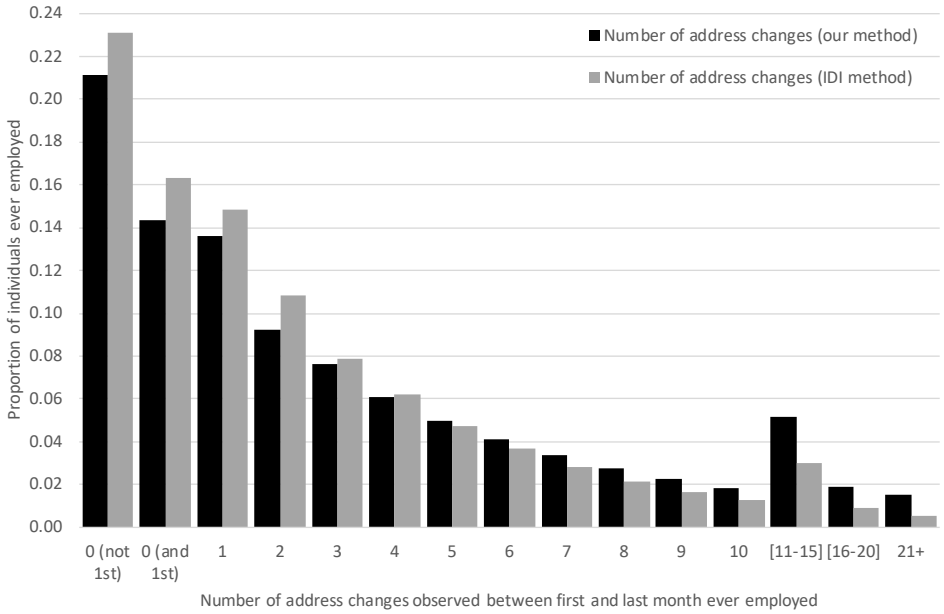


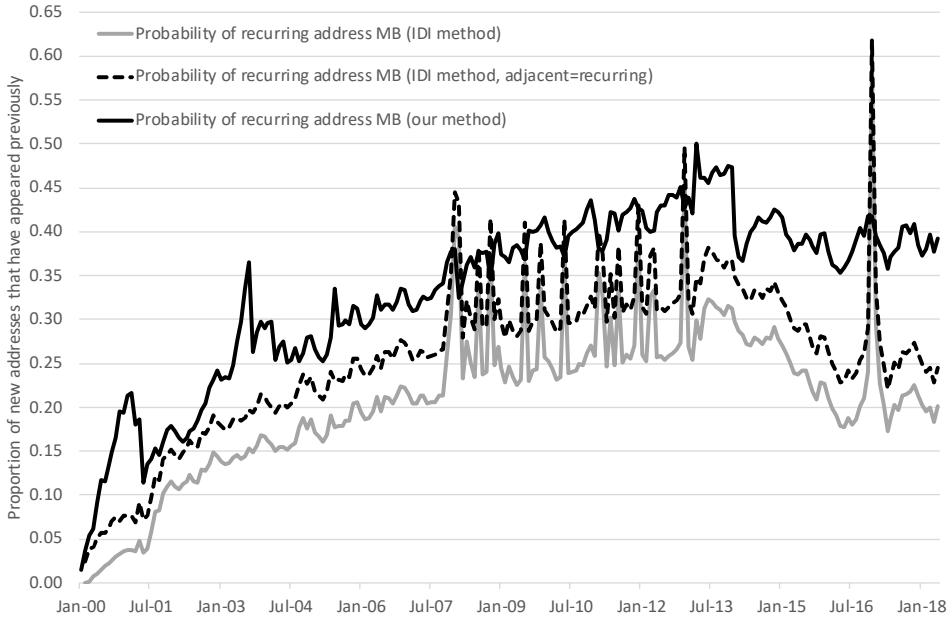
Figure excludes workers with a missing response to the Census years at usual residence question. See table 10 note for further sample restrictions.

Figure 7: Comparison of prioritisation methods by number of address changes



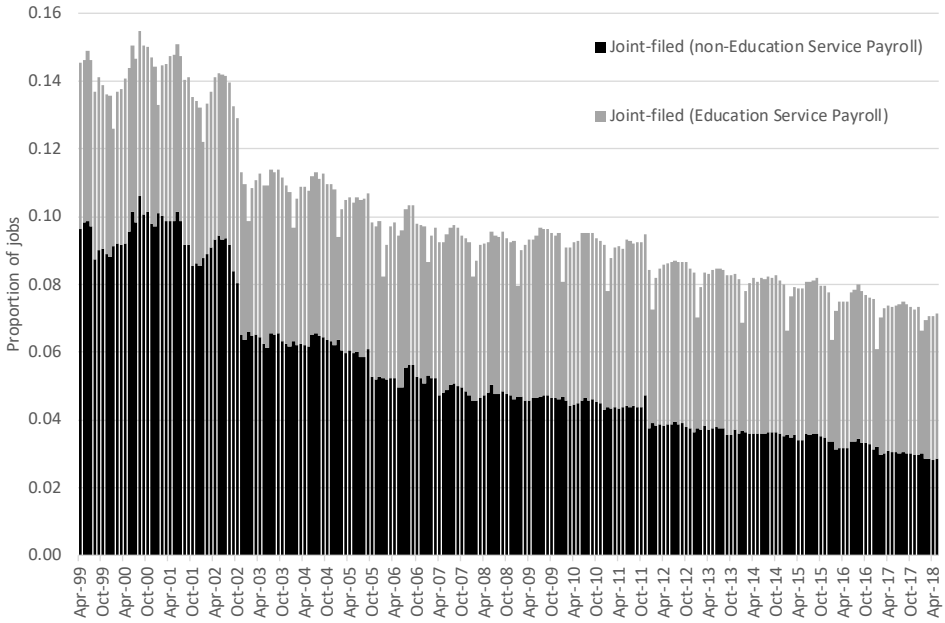
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Only address changes occurring between the first and last employment month are counted. The IDI method tracks x-y addresses so, to consistently measure address change, we ignore within-MB address changes in the IDI table. An individuals' first address table MB is not counted as an address change, since there is no observed prior state. The first two bins in the figure represent the case where zero address changes occur within the employment period, distinguished by whether the current address is or isn't the first (ie, left-censored) address in the address table.

Figure 8: Comparison of prioritisation methods by address recurrence rate



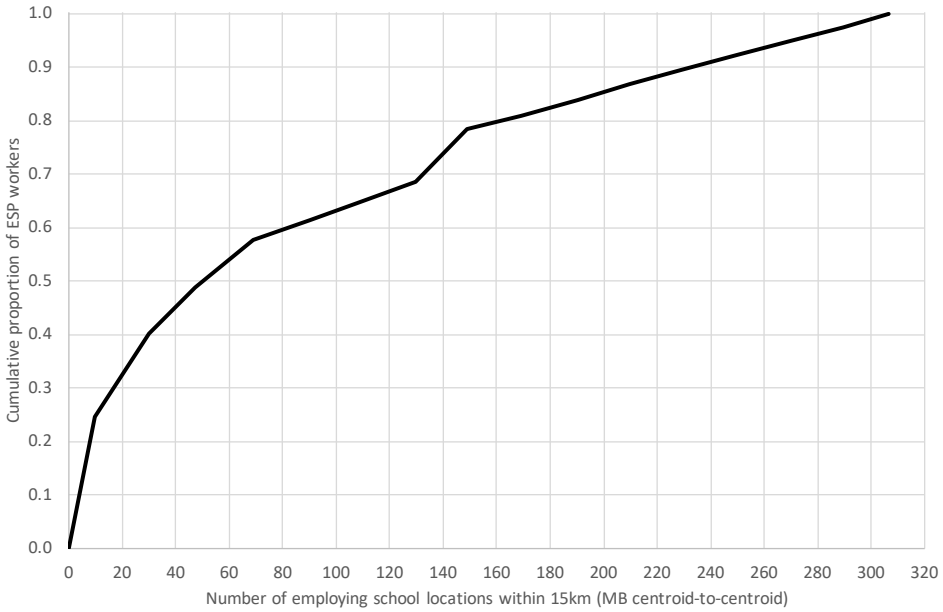
A recurring address is a MB-level address change, where the MB has previously been the active residential address. The alternative "adjacent=recurring" IDI method series (dashed line) treats adjacent MBs as the same residential address, which is likely to be true in the majority of cases. See figure 7 for other notes.

Figure 9: Proportion of jobs subject to joint-filing



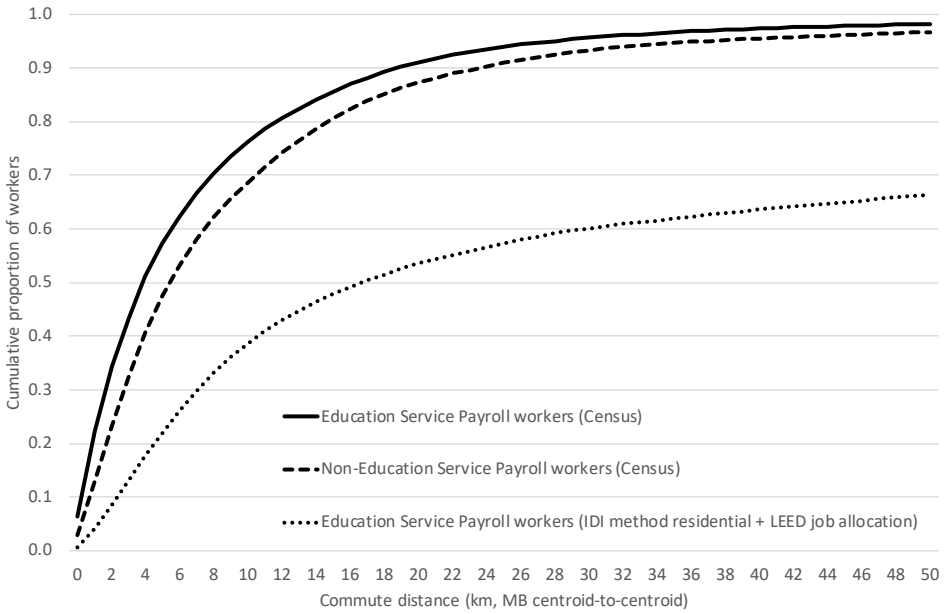
In this figure, a job is a distinct worker-IR payer pair in a month, rather than a worker-firm relationship, since joint-filing makes the specific employer (firm) unknown for some jobs.

Figure 10: Number of schools within 15km of ESP worker (March 2013)



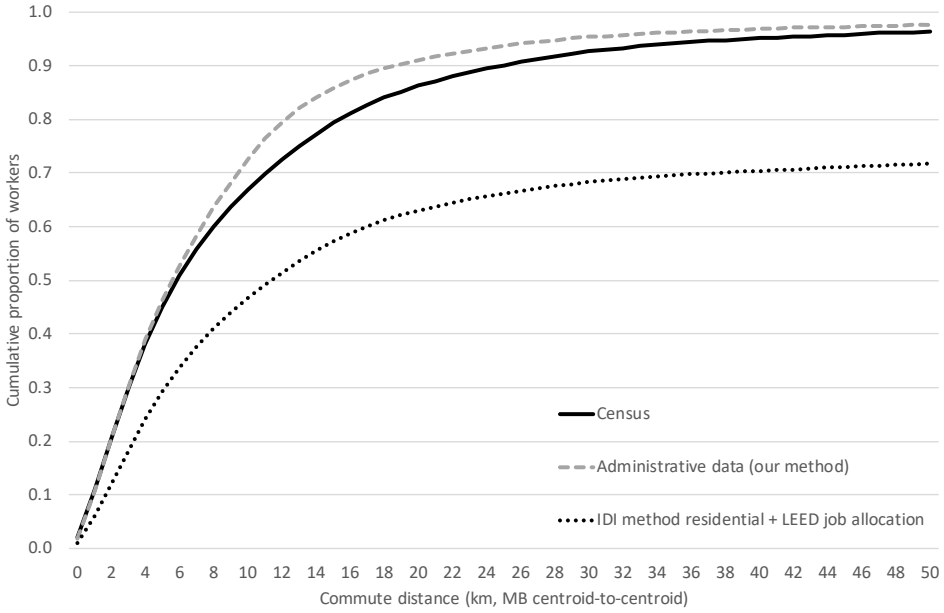
Restricted to Census employer location comparison sample, as defined in the main text. ESP workers receive earnings from the Education Service Payroll IR filer during March 2013.

Figure 11: Cumulative commute profile – ESP vs non-ESP workers



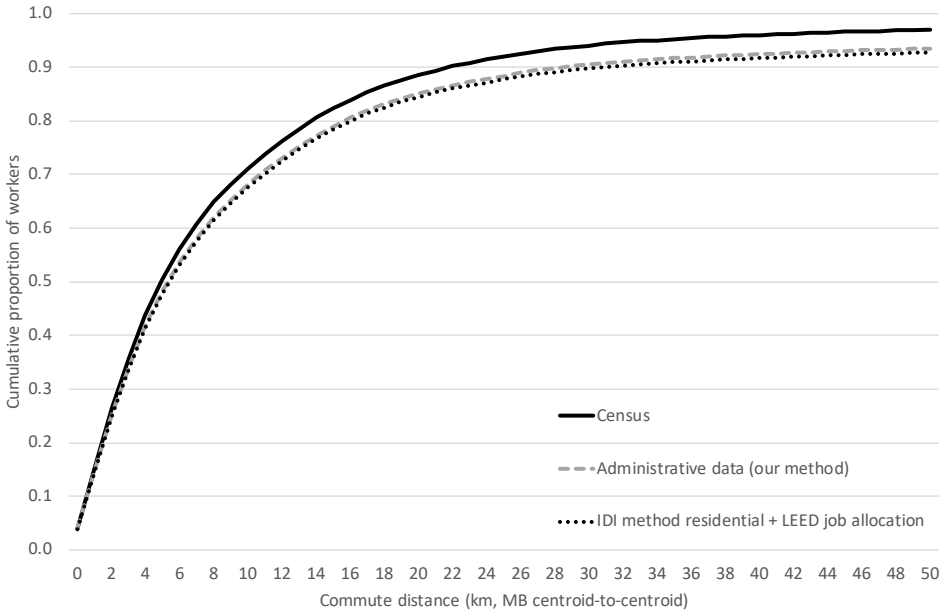
Restricted to Census employer location comparison sample, as defined in the main text. ESP workers receive earnings from the Education Service Payroll IR filer during March 2013.

Figure 12: Cumulative commute profile – workers in multi-location firms



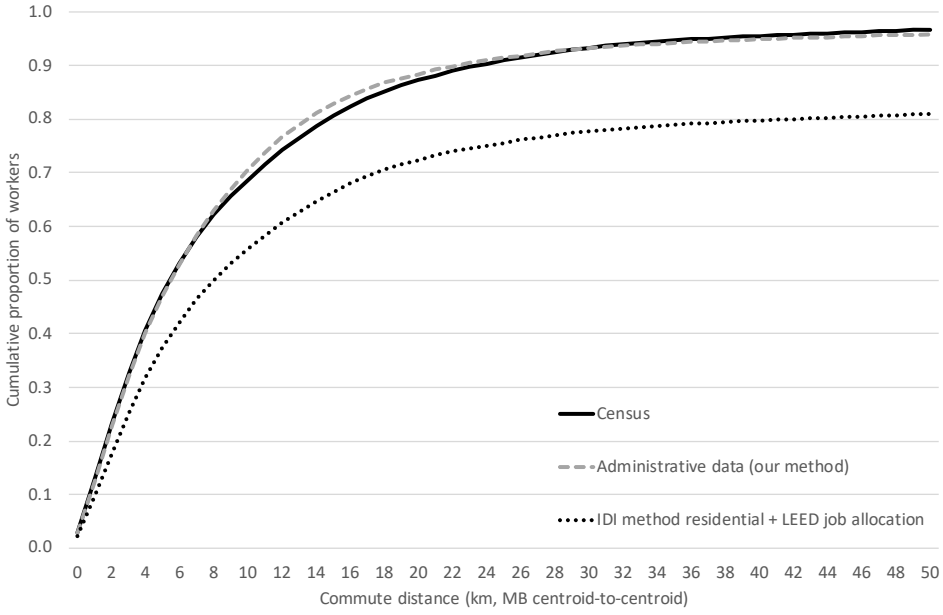
Restricted to Census employer location comparison sample, as defined in the main text. Analysis excludes ESP workers, since they do not receive a job allocation under our method. Non-ESP joint-filers are included since we allocate them to job locations in our method and, by construction, workers employed by joint-filers cannot have a single candidate location. Workers receiving a probabilistic allocation to multiple plants have weighted commutes (ie, in proportion to expected relative plant size).

Figure 13: Cumulative commute profile – workers in single-location firms



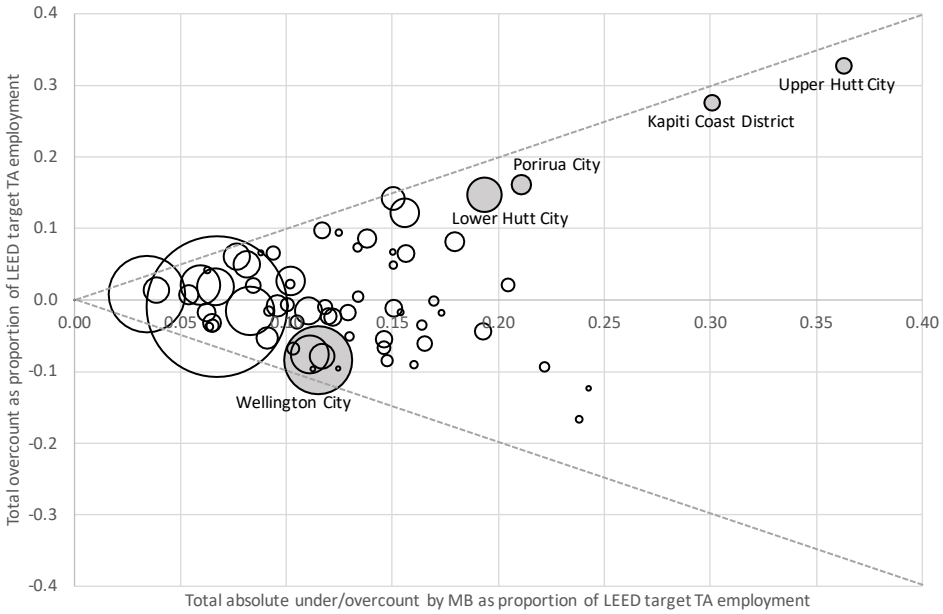
Restricted to Census employer location comparison sample, as defined in the main text. Joint-filers are excluded from this analysis since, by construction, workers employed by joint-filers cannot have a single candidate job location.

Figure 14: Cumulative commute profile – all workers



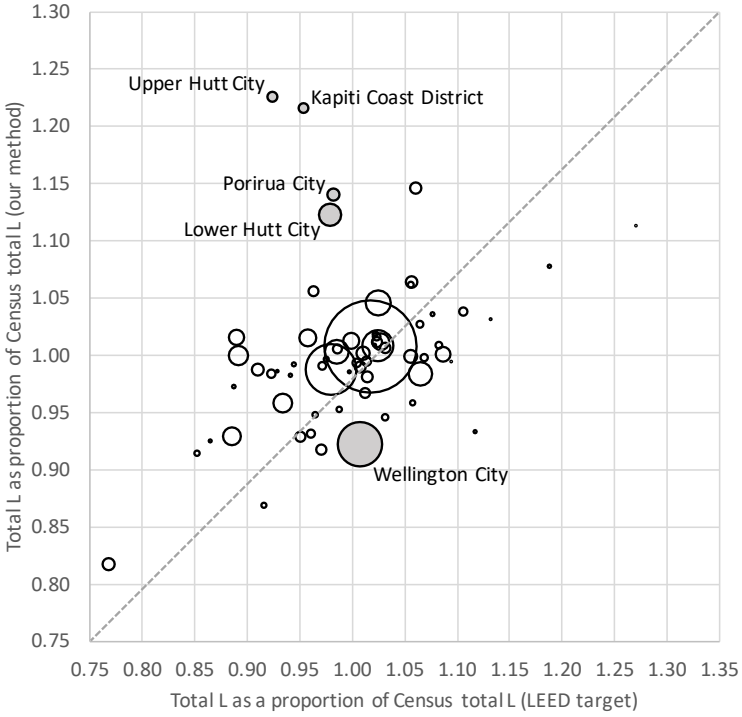
Restricted to Census employer location comparison sample, as defined in the main text. Analysis excludes ESP workers, since they do not receive a job allocation under our method. Non-ESP joint-filers are included since we allocate them to job locations in our method and, by construction, workers employed by joint-filers cannot have a single candidate location. Workers receiving a probabilistic allocation to multiple plants have weighted commutes (ie, in proportion to expected relative plant size).

Figure 15: Over- and under-count of jobs relative to LEED target allocation



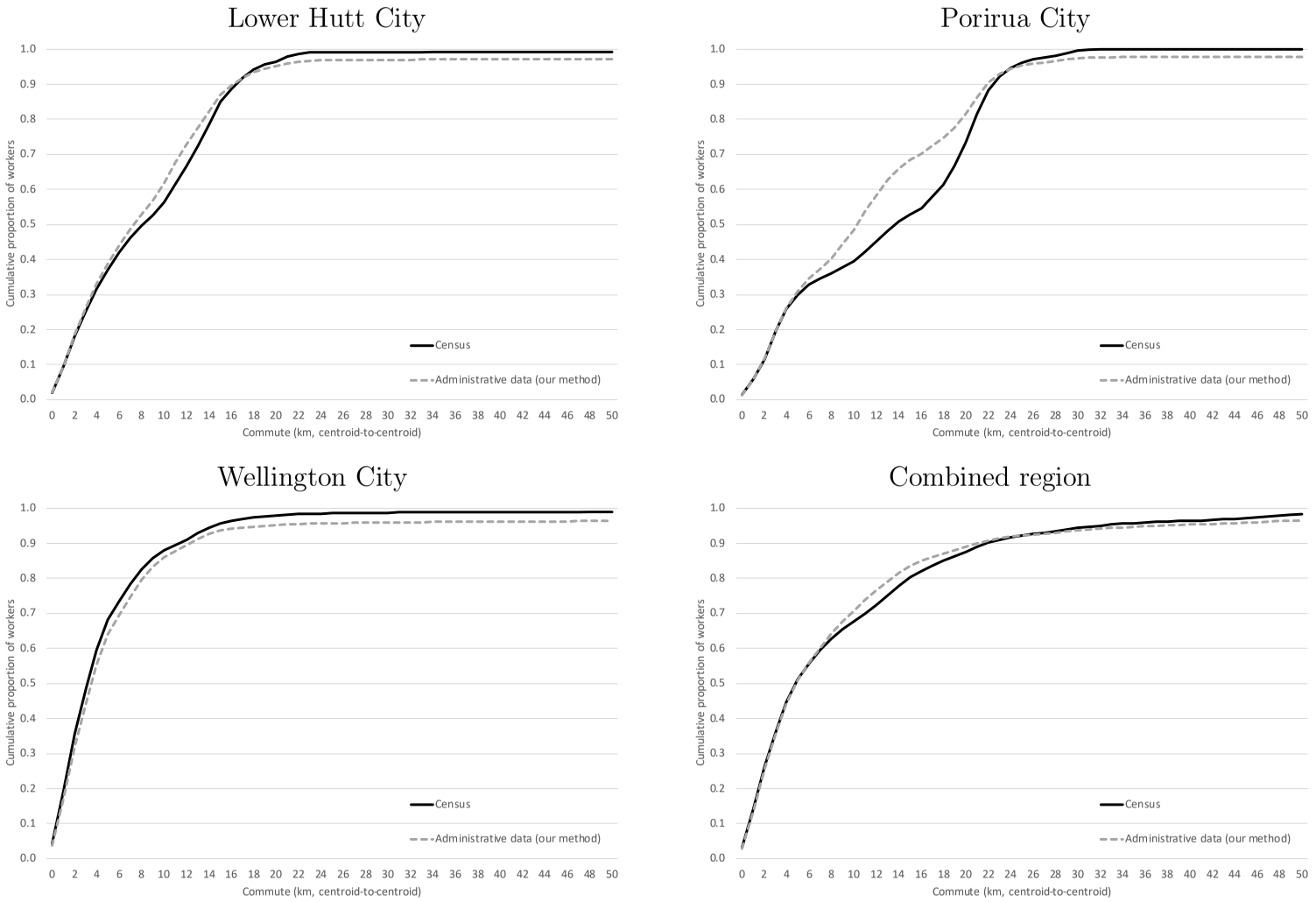
Restricted to Census employer location comparison sample, as defined in the main text. LEED target TA employment is the aggregate allocation of individuals in the sample based only on plant LEED employment shares (ie, ignoring commute feasibility). The total absolute under/overcount aggregates the absolute value of meshblock-level differences between our method and the LEED target, as a proportion of LEED target TA employment. Dashed lines indicate the cone of feasible absolute under/overcount for a given total overcount.

Figure 16: Count of jobs relative to Census – our method vs IDI method



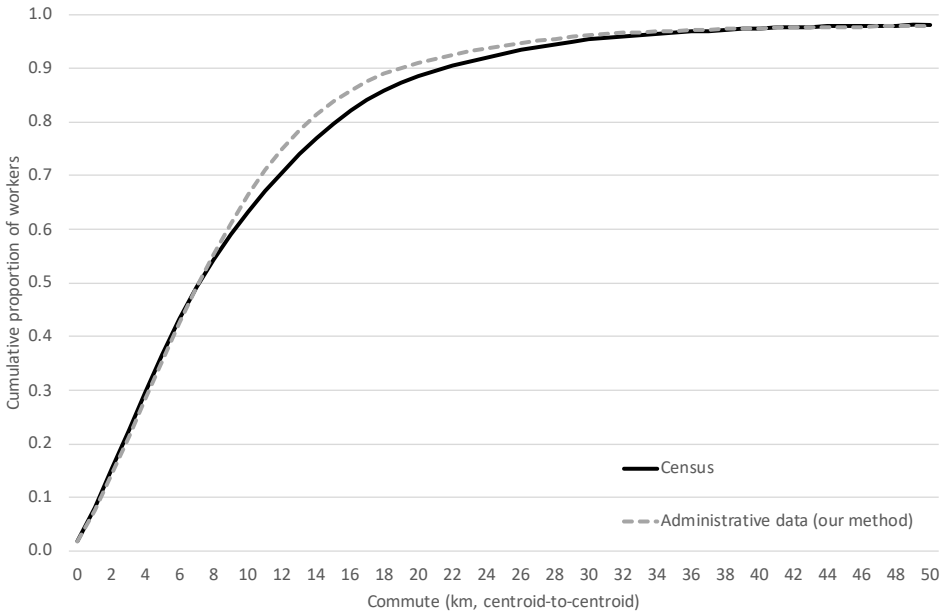
Restricted to Census employer location comparison sample, as defined in the main text. LEED target TA employment is the aggregate allocation of individuals in the sample based only on plant LEED employment shares (ie, ignoring commute feasibility).

Figure 17: Cumulative commute profile – Wellington City and surrounding region residents



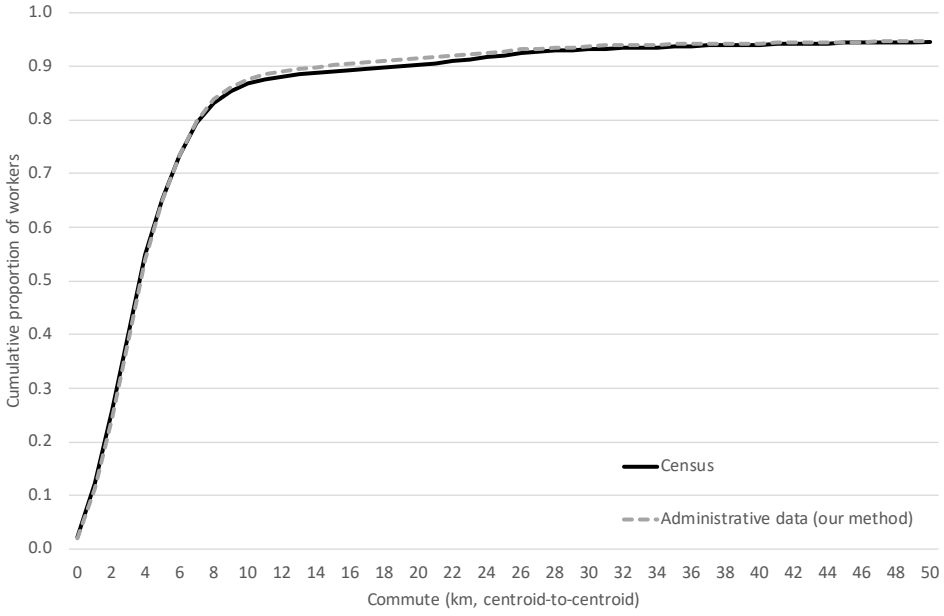
Restricted to Census employer location comparison sample, as defined in the main text. TA of residence taken from Census data to maintain a common sample for the comparison. Analysis excludes ESP workers, since they do not receive a job allocation under our method. Workers receiving a probabilistic allocation to multiple plants have weighted commutes (ie, in proportion to expected relative plant size). The combined region is the highlighted TAs in figure 15: Upper and Lower Hutt; Kapiti Coast; Porirua; and Wellington City.

Figure 18: Cumulative commute profile – Auckland residents



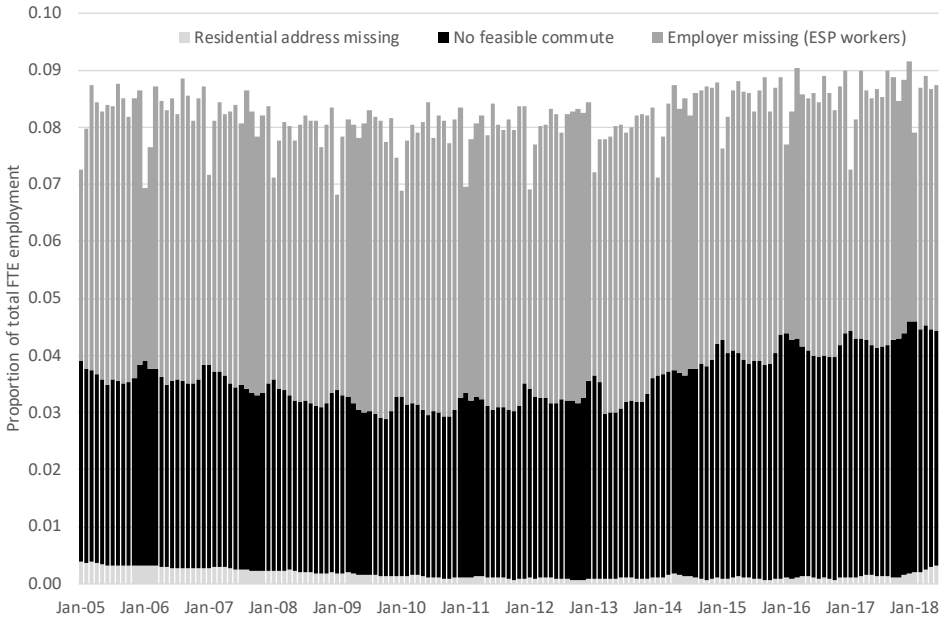
See figure 17 for notes.

Figure 19: Cumulative commute profile – Hamilton residents



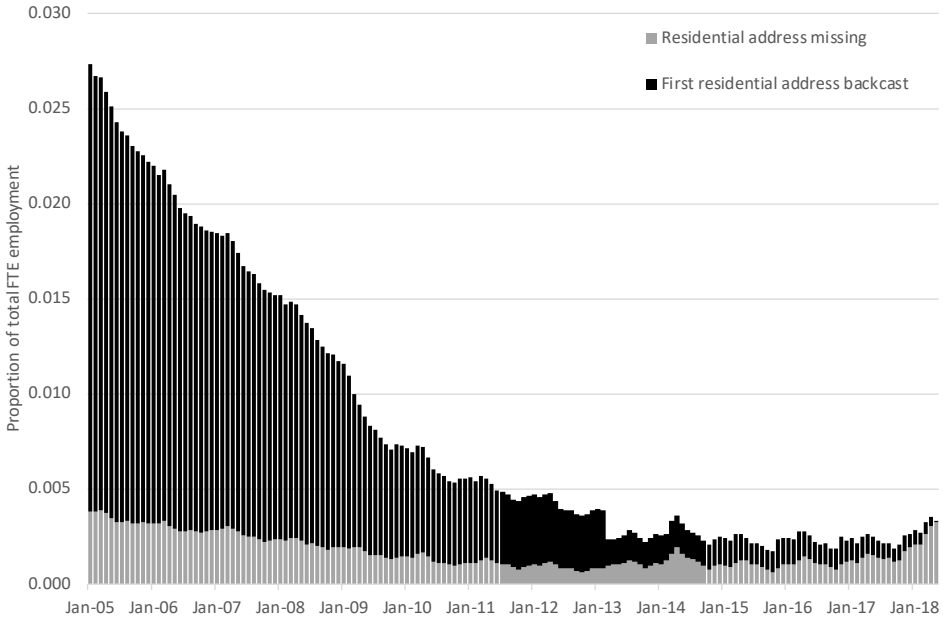
See figure 17 for notes.

Figure 20: Proportion of total FTE employment with missing commute data



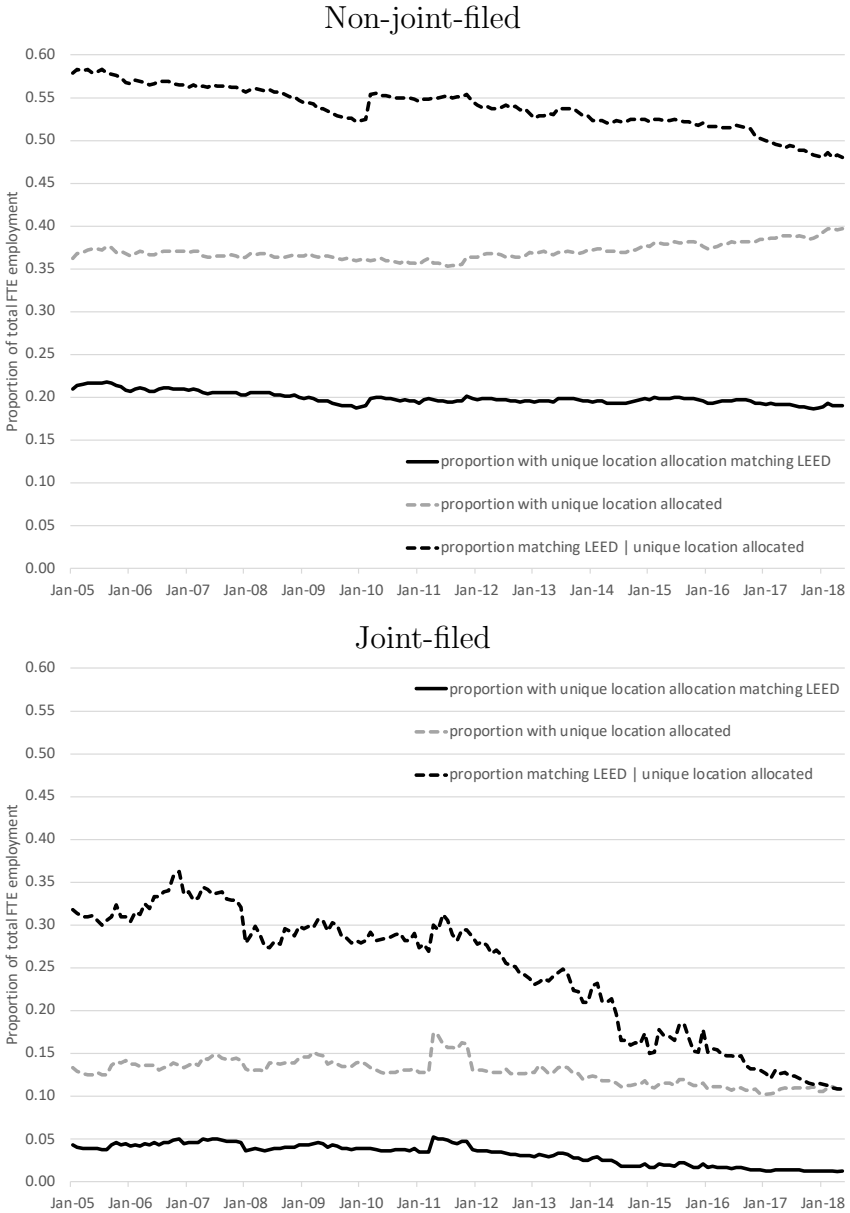
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Jobs where both the residential address and employer are missing are attributed to the residential address missing category to avoid double-counting.

Figure 21: Proportion of total FTE employment with backcast residential address



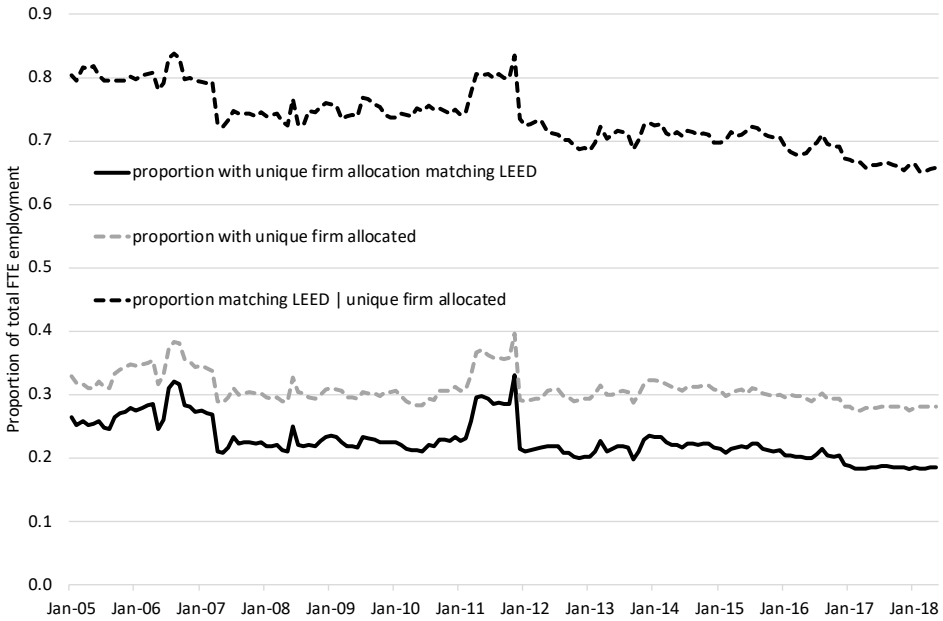
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance.

Figure 22: Unique plant allocation match to LEED for multi-location firms



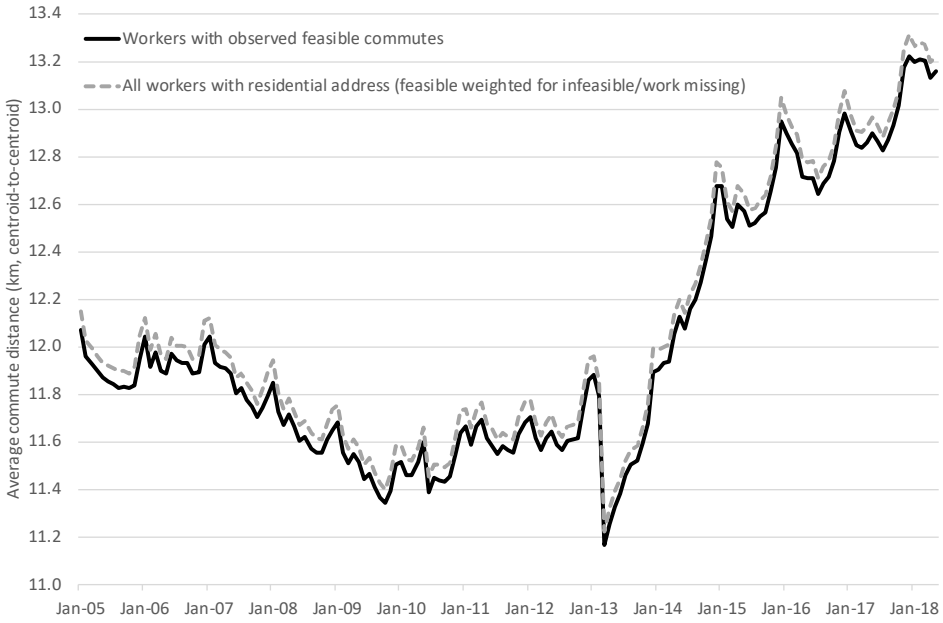
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Restricted to jobs with at least one observed feasible commute, and excluding ESP workers.

Figure 23: Unique firm allocation match to LEED for joint-filers



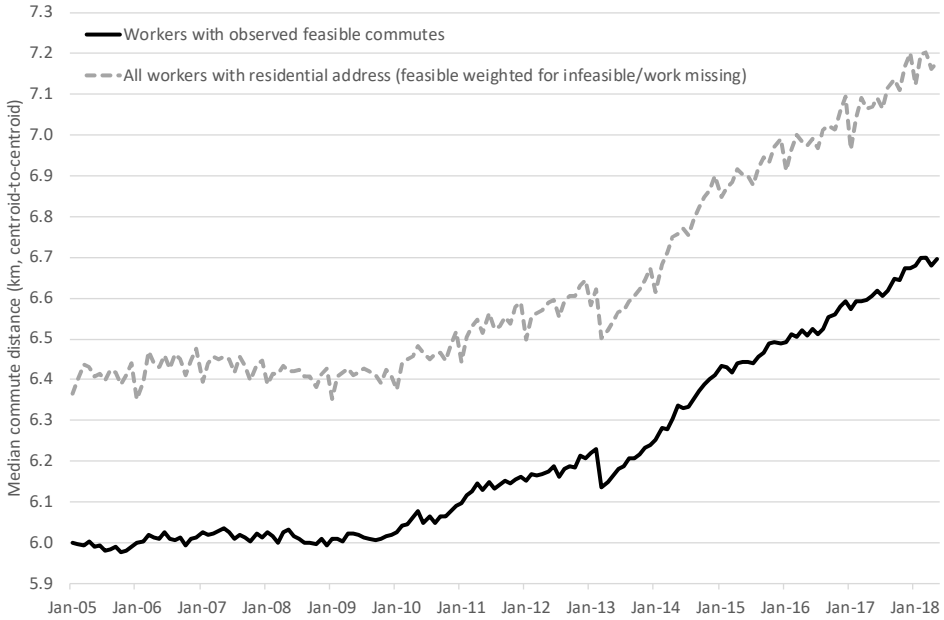
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Restricted to joint-filer jobs with at least one observed feasible commute, and excluding ESP workers.

Figure 24: Average commute distance over time



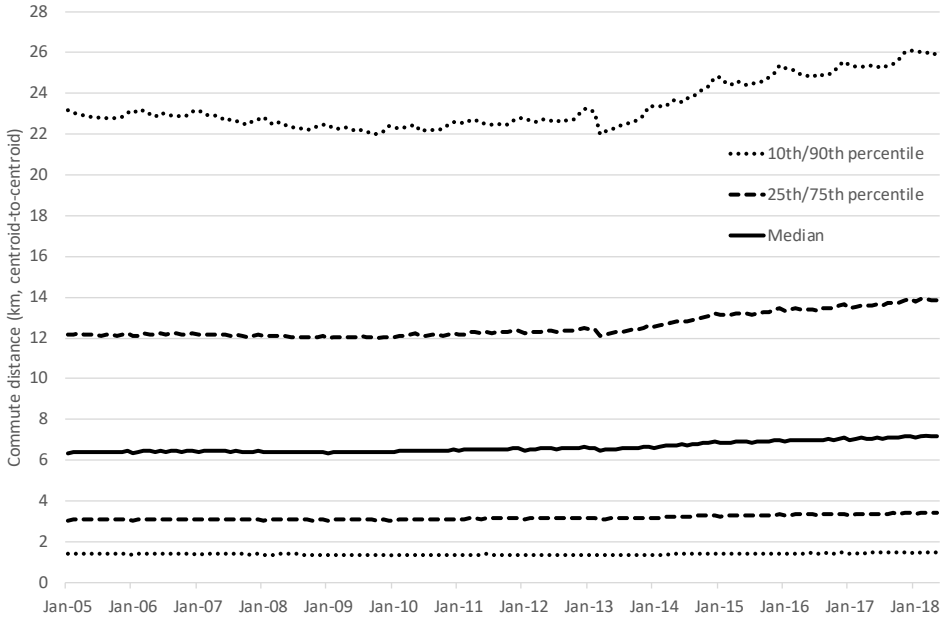
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Jobs are FTE-weighted. The all workers series is additionally weighted to account for missing commutes from the residential meshblock due to infeasibility or missing work location information. Workers with missing residential addresses make up a negligible share of total FTE (table 12) and are ignored.

Figure 25: Median commute distance over time



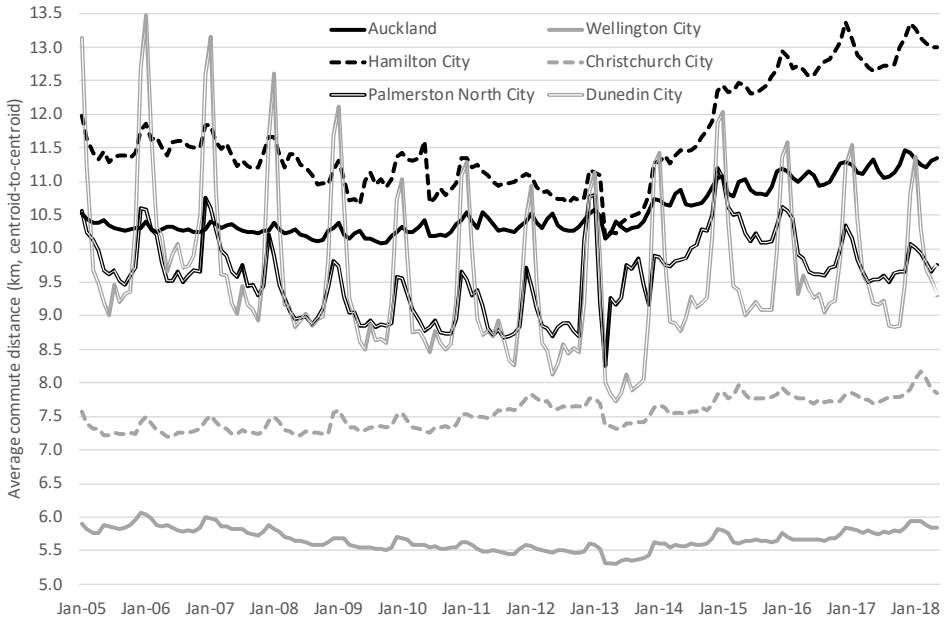
See figure 24 notes. Note: This figure has been updated since the original release of the paper.

Figure 26: Commute distance percentiles over time



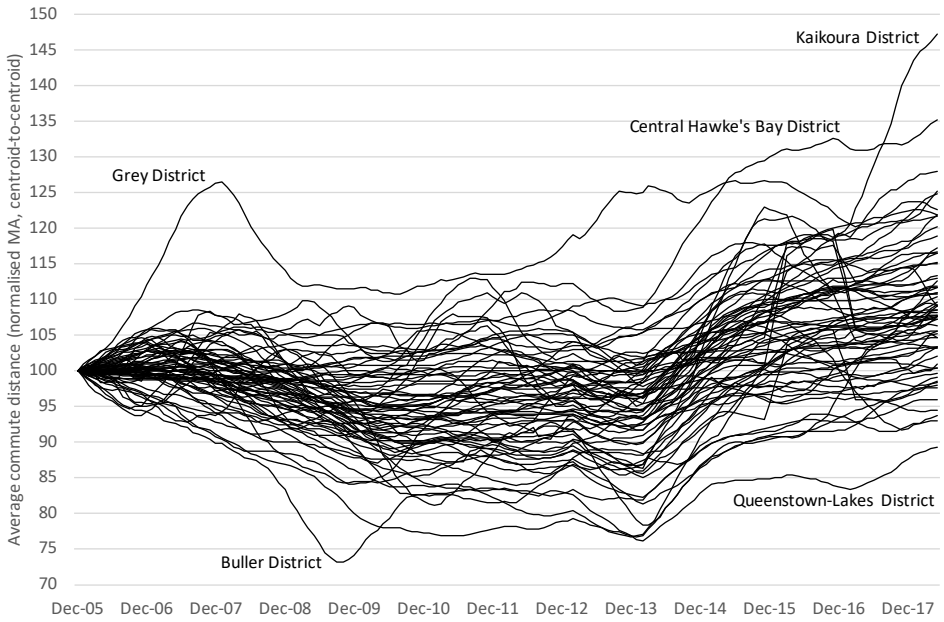
Based on final dataset, which includes all employees in the 20181020 Fabling-Maré labour table and uses MB18 addresses from the 20190420 IDI instance. Jobs are FTE-weighted and weighted to account for missing commutes from the residential meshblock due to infeasibility or missing work location information. Workers with missing residential addresses make up a negligible share of total FTE (table 12) and are ignored. Median is same as "all workers" (dashed) series in figure 25. Note: This figure has been updated since the original release of the paper.

Figure 27: Average commute distance over time – selected cities



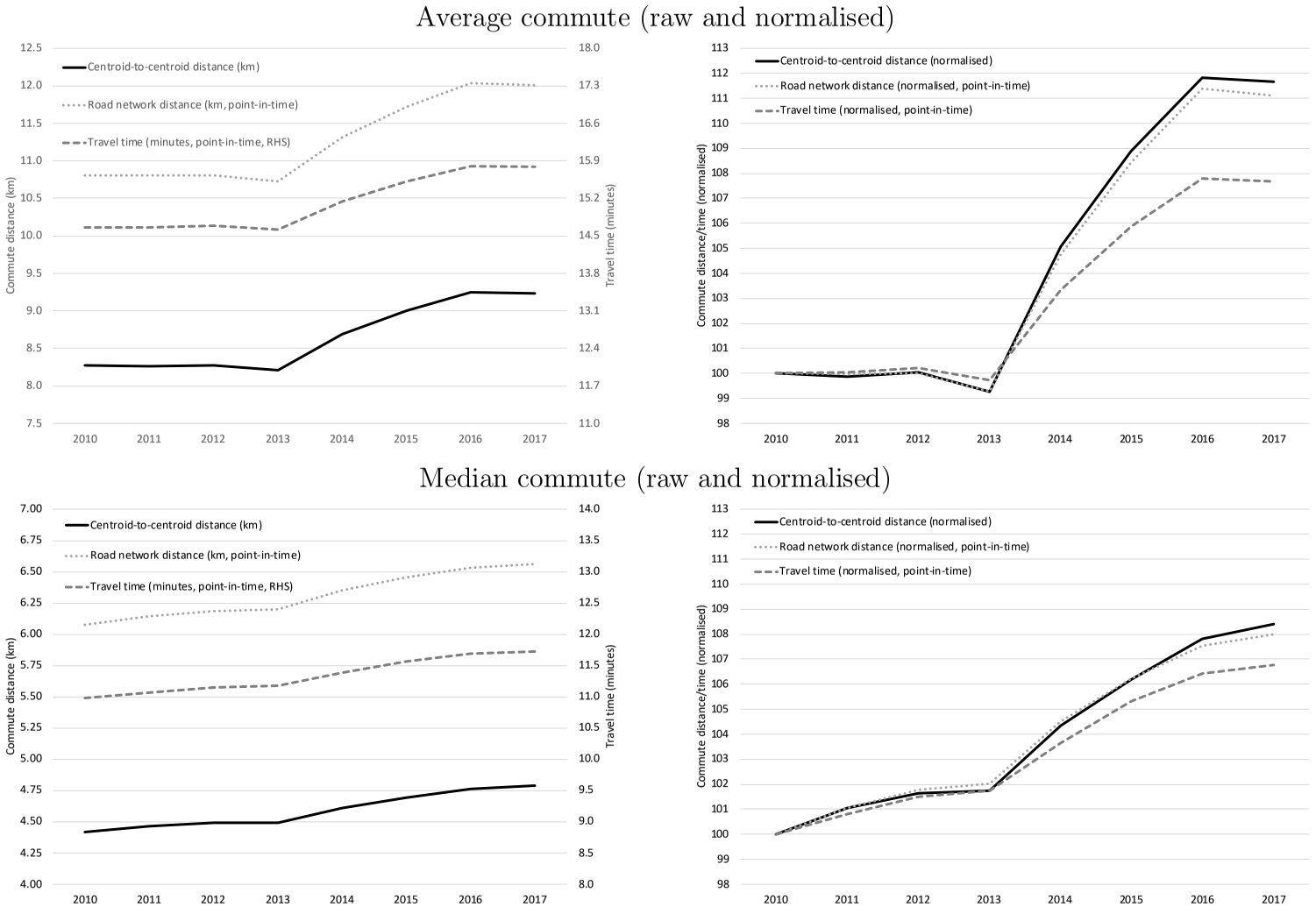
See figure 26 notes.

Figure 28: Normalised annual moving average commute distance by TA



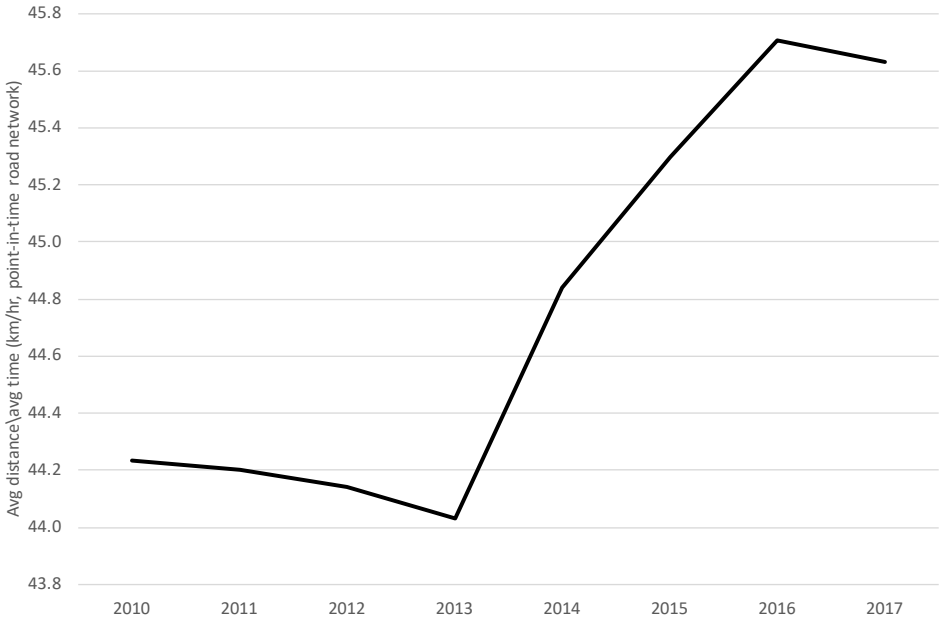
See figure 26 notes. Each of the 67 lines reflect a TA-specific twelve month moving average commute distance, normalised to 100 in the year ending December 2005.

Figure 29: Comparison of commute metrics for Hamilton UA – centroid-to-centroid vs road network vs travel time



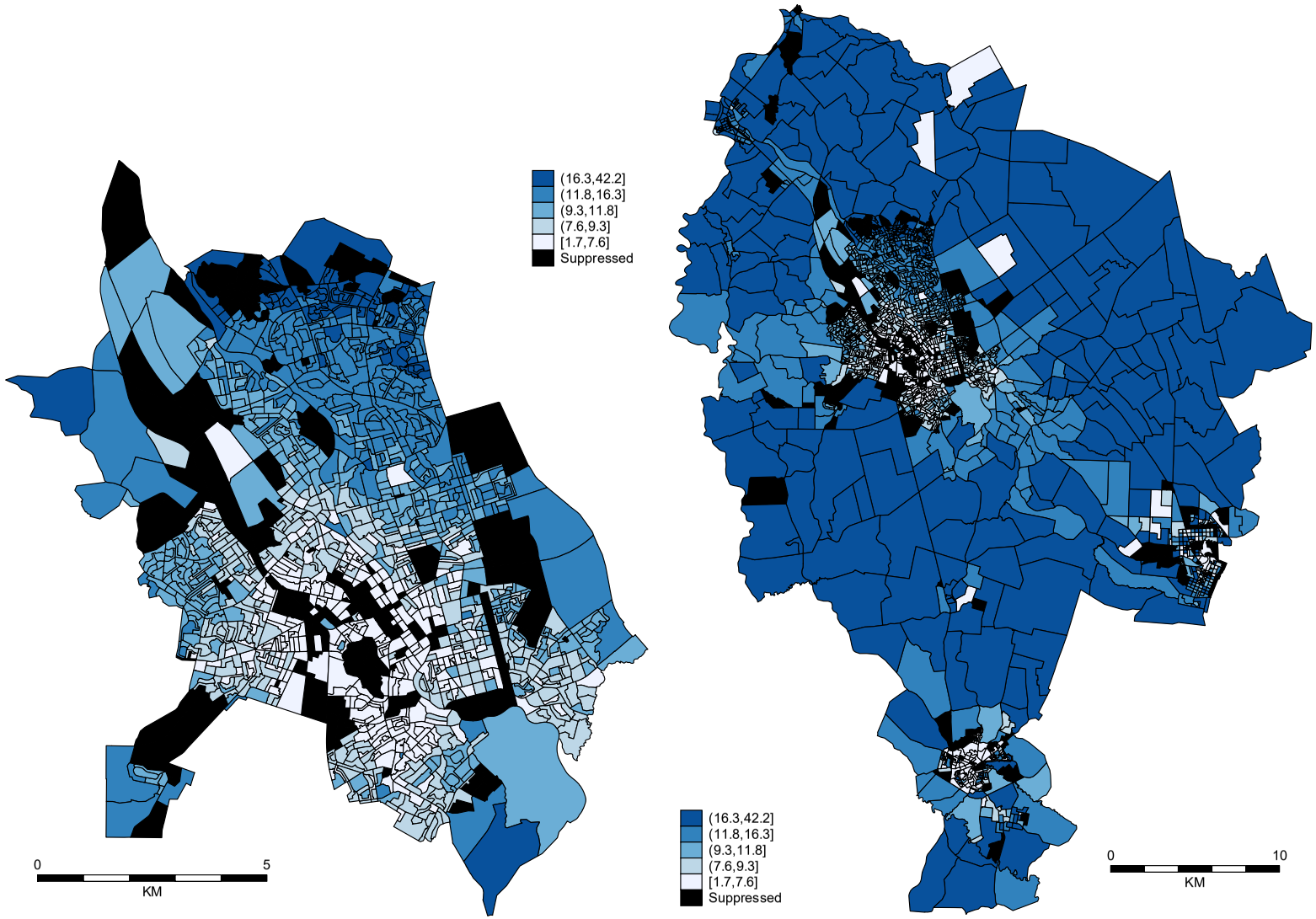
Based on final dataset, restricted to workers who live in Hamilton and who have jobs within 50km of the Hamilton Urban Area boundary. Road network distance is point-in-time as discussed in the main text. Travel time is free-flow time based on the same point-in-time road network. Left panel graphs show travel time on secondary axis, but scaled so that growth rates are comparable across distance and time measures. Right panel of figure shows normalised commute distance/time for greater ease of comparison of these growth rates, and with the same scale used for average and median.

Figure 30: Ratio of average distance over average time – Hamilton UA



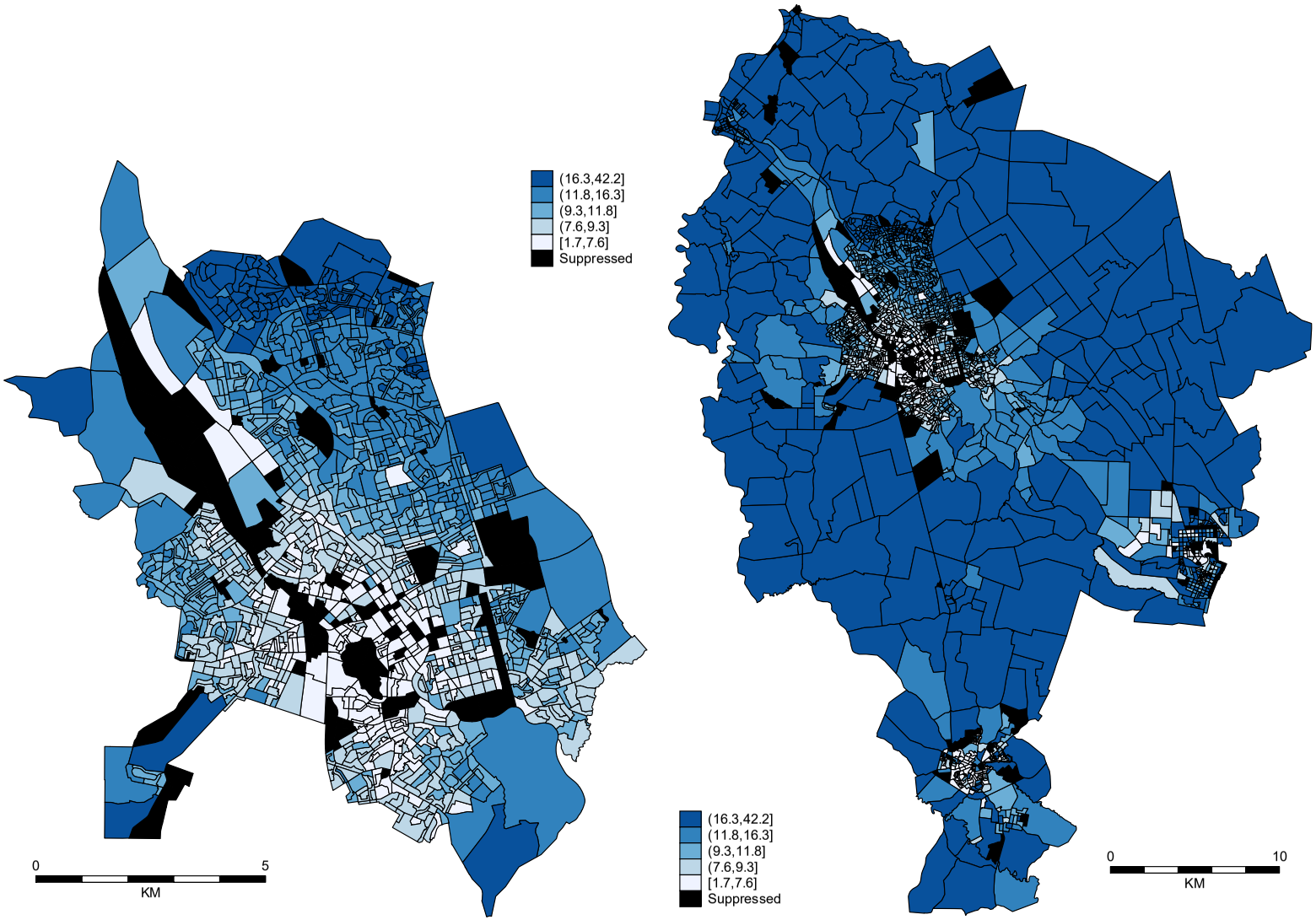
See figure 29 notes. Figure represents the ratio of average distance over average travel time (the dotted and dashed lines in the top left panel of figure 29, respectively), which is not equivalent to calculating the (FTE-weighted) average trip speed.

Figure 31: Median commute time (point-in-time road network) – Hamilton City and Urban Area, 2010



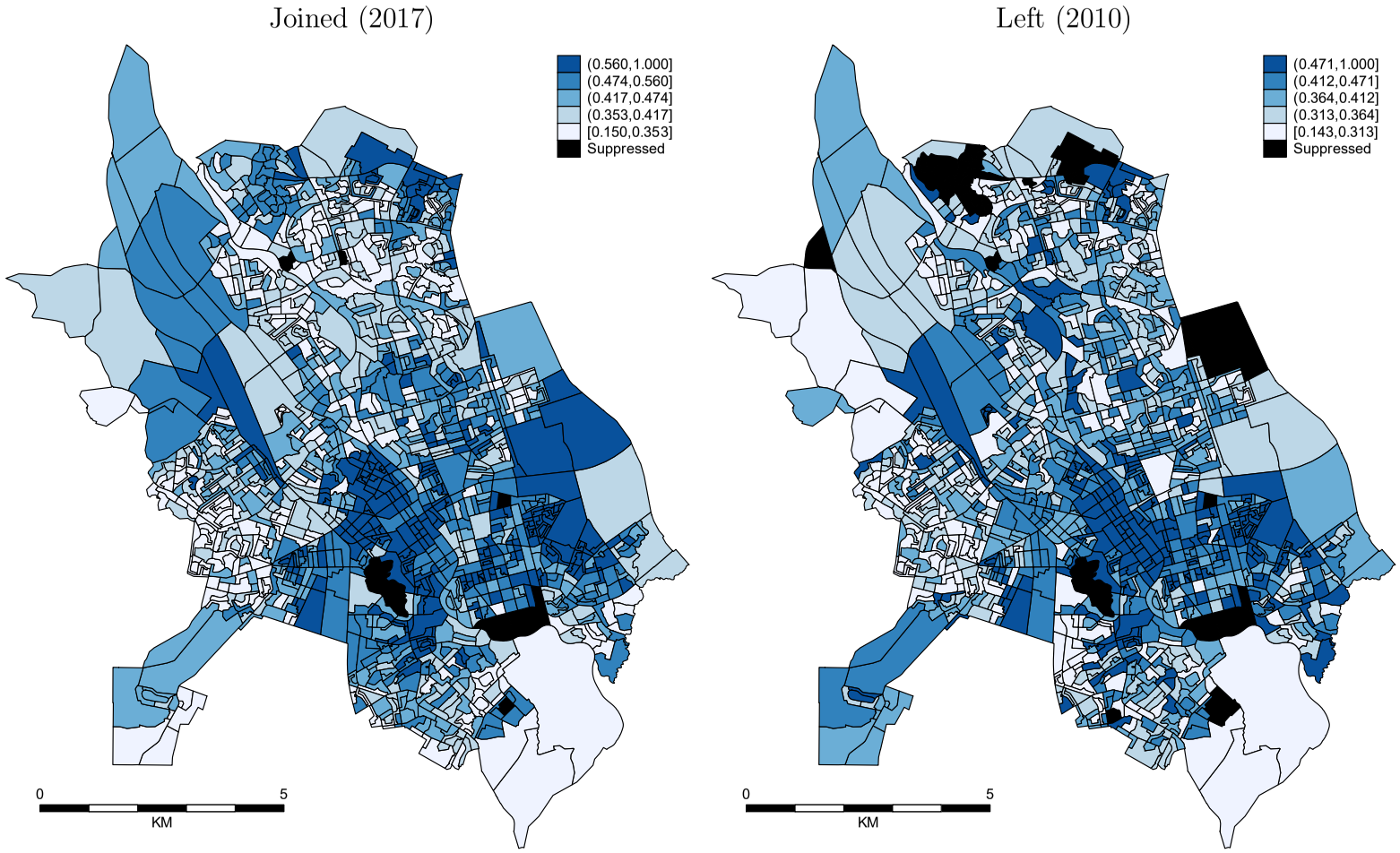
Population of interest is workers who live in Hamilton and who have jobs within 50km of the Hamilton Urban Area boundary. The Urban Area incorporates Hamilton City, and surrounding semi-urban region as well as Cambridge & Te Awamutu. The leftmost panel is zoomed in on Hamilton City, while the rightmost panel captures the entire Urban Area. Free-flow travel times estimated from a point-in-time road network from February 2015 (Beere 2016).

Figure 32: Median commute time (point-in-time road network) – Hamilton City and Urban Area, 2017



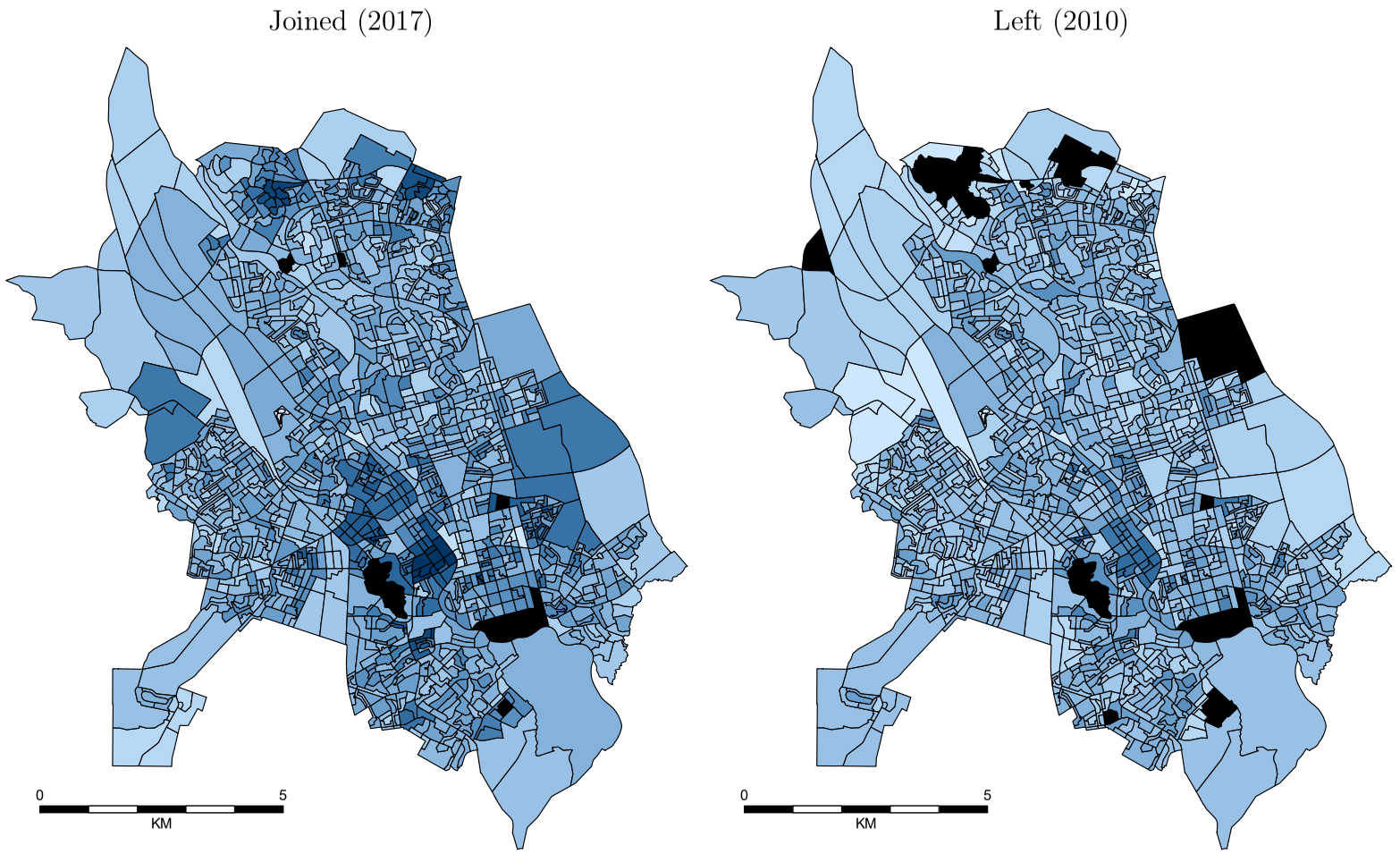
See figure 31 for notes.

Figure 33: Proportion of residents who join or leave population – Hamilton City



See figure 31 for notes. Resolution of graph is SA1, not MB, for confidentiality reasons.

Figure 34: Number of residents who join or leave population – Hamilton City



See figure 31 for notes. Resolution of graph is SA1, not MB, for confidentiality reasons.

A. Codebooks for IDI Sandpit tables

Current tables are under `IDI_Sandpit.[DL-MAA2018-55]`, restricting access to researchers on the specific Datalab agreement, and exploit two IDI instances (indicated by table name). Person identifiers (`snz_uid`s) do not relate to the same individual across instances. The IDI labour dataset includes the necessary concordance table for linking employees across IDI instances (`[clean_read_IR].[snz_uid_20181020_to_20190420_RFabling]`). Population is ever-employed individuals (according to `IDI_20181020` labour tables).

Table A.1: address_clean_IDI_20190420

Key	Variable	Type	NULL	Description
Y	snz_uid	int	N	IDI person id (instance-specific)
Y	add_source	varchar(4)	N	Address data source†
	add_tier	tinyint	N	Residential address tier (1 or 2)
Y	add_date	date	N	Date of address notification
Y	mb	char(7)	N	Residential address meshblock‡
	first_add	tinyint	N	Indicator variable set to 1 if first address in sequence (0 otherwise)
	seq	int	N	Chronological sequence number of address

†Address data source variable follows IDI naming convention: ACCC; CEN; IRDA; MOES; MOH; MSD; MSDP; NZTM; NZTD. MOH is pooled Ministry of Health (labelled MOHP & MOHN in IDI). MSD is residential Ministry of Social Development (MSDR) pooled with postal MSD (MSDP) promoted to tier one. Remaining (tier 2) MSD addresses retain the original MSDP labelling. ‡MB vintage for `IDI_20190420` is MB2018 V1.00.

Table A.2: address_spell_IDI_20190420

Key	Variable	Type	NULL	Description
Y	snz_uid	int	N	IDI person id (instance-specific)
Y	start_add_date	date	N	Start date of address spell
	end_add_date	date	Y	End date of address spell†
	mb	char(7)	N	Residential address meshblock
	add_tier	tinyint	N	Residential address tier (1 or 2)
Y	seq	int	N	Chronological sequence number of address spell

†Last spell in sequence has a null end date.

Table A.3: address_spell_emp_mth_IDI_20190420

Key	Variable	Type	NULL	Description
Y	snz_uid	int	N	IDI person id (instance-specific)
	start_month	int	N	Start month (as at 15th of month) of address spell, represented as YYYYMM†
	end_month	int	N	End month (as at 15th of month) of address spell, represented as YYYYMM‡
	mb	char(7)	N	Residential address meshblock
	add_tier	tinyint	N	Residential address tier (1 or 2)
Y	seq	int	N	Chronological sequence number of address spell

Table restricted to address spells overlapping the period between an individuals' first and last employment month. †First spell start is based on the date of the first observed address (ie, no backcasting of spell start date). ‡Final spell end month is set equal to the last month in the address table (currently 201902) if meshblock is last observed address.

Table A.4: ems_payer_pent_mth_IDI_20181020

Key	Variable	Type	NULL	Description
Y	snz_employer_ird_uid	int	N	Employer IR number (confidentialised)
Y	pent	char(10)	N	Employer id (permanent enterprise)
Y	dim_month_key	int	N	Employment month, represented as YYYYMM

One-to-many employer IR to firm relationships (from LEED EMS) are "joint-filers."

Table A.5: joint_ems_payer_emp_mth_IDI_20181020

Key	Variable	Type	NULL	Description
Y	snz_employer_ird_uid	int	N	Employer IR number (confidentialised)
Y	snz_uid	int	N	IDI person id (instance-specific)
Y	dim_month_key	int	N	Employment month, represented as YYYYMM
	gross_earn	decimal(13,2)	N	Gross earnings (from EMS)

Wage and salary (gross earnings) payments from joint-filers to employees.

Table A.6: joint_ems_pent_emp_mth_IDI_20181020

Key	Variable	Type	NULL	Description
Y	pent	char(10)	N	Employer id (permanent enterprise)
Y	snz_uid	int	N	IDI person id (instance-specific)
Y	dim_month_key	int	N	Employment month, represented as YYYYMM
Y	snz_employer_ird_uid	int	N	Employer IR number (confidentialised)
	fte	float	N	Full-time equivalent labour (from labour table)
	not_joint_filed	tinyint	N	Indicator set to 1 if employer IR to PENT relationship is one-to-one (0 otherwise)
	MinEd_payroll	tinyint	N	Indicator set to 1 if employer IR is Education Service Payroll (0 otherwise)

Payments to employees where the firm link in LEED could have been generated by a joint-filer. Table is needed because of individuals receiving non-joint-filing and joint-filing earnings involving a common set of firms, creating the potential that two distinct jobs are aggregated in the labour tables. This table identifies the components of any aggregated job that should potentially be reallocated following our joint-filer method.

Table A.7: mb2018_V1_00_adjacency

Key	Variable	Type	NULL	Description
Y	mb_s	char(7)	N	Source meshblock
Y	mb_d	char(7)	N	Destination meshblock

Meshblock pairs that share a boundary (are adjacent). Table is symmetric, and MBs are included as adjacent to themselves. A separate table (`mb13_adjacency`) is available to facilitate Census 2013 testing.

Table A.8: mb2018_V1_00_feasible_200km

Key	Variable	Type	NULL	Description
Y	mb_s	char(7)	N	Source meshblock
Y	mb_d	char(7)	N	Destination meshblock
	dist_km	float	N	Centroid-to-centroid distance (km)
	north_island	tinyint	N	Indicator set to 1 if MBs are in North Island TAs (0 otherwise)

Meshblock pairs that have a feasible commute between them, where feasible commute is defined as within 200km (centroid-to-centroid) and same (North/South) island territorial authority (TA). Table is symmetric, and `dist_km`=0 for within-MB commutes.

Table A.9: pent_pbn_mth_LEED_ec_IDI_20181020

Key	**Variable**	**Type**	**NULL**	**Description**
Y	pent	char(10)	Y	Employer id (permanent enterprise)
Y	pbn_nbr	char(10)	Y	Employer location id (PBN)
Y	dim_month_key	int	N	Employment month, represented as YYYYMM
	LEED_ec	int	N	LEED headcount employment

Used for EC-weighting the probabilistic allocations for multi-location firms.

Table A.10: commute_pair_mb2018_IDI_20181020

Variable	**Type**	**NULL**	**Description**
pent	char(10)	Y	Employer id (permanent enterprise)
snz_uid	int	N	IDI person id (instance-specific)
dim_month_key	int	N	Employment month, represented as YYYYMM
pbn_nbr	char(10)	Y	Employer location id (PBN)
fte	float	N	Full-time equivalent labour (from labour table)
mb_wk	char(7)	Y	Work (PBN) address meshblock
mb_res	char(7)	Y	Residential address meshblock
dist_km_feasible	float	Y	Centroid-to-centroid distance (km), if feasible[†] (NULL otherwise)
EC_weight	float	N	Probability weight for commute, based on PBN employment counts[‡]
add_tier	tinyint	Y	Residential address tier (1 or 2; NULL if residential address unknown)
res_imputed	tinyint	Y	Indicator set to 1 if residential address backcast from first observed address (0 otherwise; NULL if residential address unknown)
wk_nonjoint_no_loc	tinyint	N	Indicator set to 1 if employer IR to PENT relationship is one-to-one & no feasible commutes from current residential address (0 otherwise)
wk_nonjoint_one_loc	tinyint	N	Indicator set to 1 if employer IR to PENT relationship is one-to-one & single feasible commutes from current residential address (0 otherwise)
wk_nonjoint_multi_loc	tinyint	N	Indicator set to 1 if employer IR to PENT relationship is one-to-one & multiple feasible commutes from current residential address (0 otherwise)
wk_joint_nonMinEd	tinyint	N	Indicator set to 1 if employer IR to PENT relationship is one-to-many and not Education Service Payroll (0 otherwise)
wk_joint_MinEd	tinyint	N	Indicator set to 1 if employer IR is Education Service Payroll (0 otherwise)

Table has no primary key since employer and location can be NULL (in the case of ESP workers). For non-ESP jobs, if there is no feasible commute then each potential work location has a separate row in the table (with `dist_km_feasible` NULL). If there is at least one feasible commute, only commutes within 10km of the shortest commute are represented in the data. [†]Feasible commute defined as within 200km and same (North/South) island territorial authority (TA). [‡]Sum of `EC_weight*fte`=monthly total FTE employment of individual.