UNIVERSITY OF LEEDS

An Exploration into the Uses of Mobile Phone Data for Providing Spatio-Temporal Updates on Commuting Flows in Leeds and Surrounding Areas

(Image: McMillan, 2019)

Ellie Marfleet

BA Geography

201097431

2020

ACKNOWLEDGEMENTS

Firstly, I would like to thank Huq for their provision of resources during this research project – with particular thanks to Alexander Fairfax (CMO) and Steve Halsall (Red Tiger Talent) for their support, without whom this project would not have been feasible.

I would also like to thank my dissertation supervisor (and personal tutor) Dr Roger Beecham for his guidance and support throughout my time at Leeds. Finally, I would like to thank Annabel Whipp in addition to my friends and family for their unwavering support throughout the entirety of my degree.

Table of Contents

ABSTRACT

The rise of *big*-data and administrative datasets offer new methods of gaining population-level statistics that may be utilised following the termination of the England and Wales Census in 2021. Studying the evolving spatiotemporal distribution of mobile devices in the network offers the possibility of inducing these population-level statistics with little or no involvement from device users, and can be provided at more timely intervals compared to the elongated ten-year timeframe between censuses. Based on a dataset of 65,054 anonymised geo-data events (GDEs) from 2019 provided by private sector company *huq*, the potential uses of mobile phone data (MPD) in the analysis of commuting patterns will be explored through comparative techniques to reciprocal 2011 census-derived commuting statistics. This study will use GIS and spatial and statistical analysis to assess the potential of MPD in replacing or updating census statistics during the intercensal period by using two workplace areas as destinations in Leeds (West Yorkshire), and Output Areas (OAs) as journey origins. This interaction between OA (home) and workplace (destination) form the basis of the OD flow matrices used for analysis in this study, in which the MPD refers to inferred flows and the census results refer to the realistic 2011 flows. Socio-economic variables are then analysed to outline potential reasons for over and under representations of flows by huq in certain spatial zones.

The results demonstrate how MPD offers a promising methodology in measuring the movement of large proportions of the population at a reduced cost and less time-consuming scale than the census. Proportions of students, employees within upper industries and higher levels of social grade were found to have the most significant relationship in areas where the MPD exhibited the largest level of difference to census variables.

WORD COUNT: 9, 982

LIST OF FIGURES AND TABLES

Tables

Figures

Appendix

TABLE OF ABBREVATIONS

MPD – Mobile Phone Data

TTW – Travel to Work Census Statistics (refers to the 2011 commuting variables within this study)

OD – Origin Destination pairing

MNO – Mobile Network Operator

LA – Local Authority

LAD – Local Authority District

MSOA – Middle Super Output Area

OA – Output Area

WZ – Workplace Zone

CDR – Call Detail Records

LQ – Location Quotient

ONS – Office for National Statistics

WFH – Work from Home

CHAPTER 1: INTRODUCTION

1.1. Background and Rationale

One of the key sources of information referring to commuting behaviours derives from the UK Census, in which travel behaviours are captured through respondents' answers and are then disseminated into aggregated data reflecting the spatial distribution of commuters. Interaction data is also produced by the census methodology, reflecting flows of commuters between their homes (origin) and workplace (destination).

In recent months however, the UK's National Statistician has confirmed the 2021 census is likely to be the final census ever to ever occur due to cost and time constraints surrounding its collection and the inconvenience of the ten-year intercensal period (ONS, 2014). This period causes a partial repository of invalidated results due to the scale of relocation in both home and work addresses in the UK. This has instigated a need to consult alternate sources which may yield the same richness of data but through less time consuming and costly measures (Shaw, 2020).

In the private sector, a promising administrative data source in which reciprocal variables can be inferred is mobile phone data (MPD). The ubiquitous nature of mobile phones coupled with advancements in mobile technology has provided us with a new tool of monitoring our daily whereabouts. The increasing use of global positioning systems (GPS) has supported this advancement, as more spatially precise data has been made available through the exact whereabouts of mobile devices. Therefore, it is topical to evaluate the potential uses of MPD within social science studies in preparation for the termination of the census after 2021, as MPD may provide a near real-time update on potentially outdated census statistics.

This report assesses the scale of representation existent in locational mobile phone data (MPD) through comparison to the broad coverage of results attained by the 2011 census. This is achieved through a specific analysis of commuting behaviours to two workplace zones in Leeds (West Yorkshire), using an extract of huq's MPD data (2019) and the 2011 Travel to Work (TTW) census statistics for comparison. The details of these datasets are outlined within *Chapter 3: Data.*

1.2. Research Aim and Objectives

The overarching aim of this research is to evaluate the potential use of mobile phone data to infer origin-destination (OD) travel to work behaviours in and around Leeds. This research will be focused on achieving three objectives:

Objective 1: Evaluate the Huq dataset as a replacement to the census for inferring population-level origin-destination (OD) travel behaviours.

Objective 2: Consider spatial zones exhibiting over and under-representation of OD flows by huq in relation to socio-economic variables, postulating reasons for differences.

Objective 3: Characterise the extent of the differences between the two datasets and speculate on implications if new data are to be used as a replacement for the census.

1.3. Outline

The structure of this report is based around achieving the three objectives and is presented in seven chapters. First, a literature review explains the importance of the census and introduces emergent observational sources that may be used as an alternative, or as supplementary sources to the census following its potential termination after 2021 (Henley, 2011). The literature review also characterises previous studies which have utilised observational data sources such as MPD in both commuting and other analyses and addresses the issue of biases and limitations within these datasets. The two datasets which have informed the analysis of this report are then outlined, followed by a methodology explaining, justifying and evaluating the research techniques implemented to achieve the three objectives. The analysis chapter then outlines the results of the research, and the subsequent discussion chapter outlines the significance of findings in comparison to previous studies and addresses limitations of the study and how this impact the validity of results. The report concludes by suggesting areas of future research and concluding statements restating the main findings.

CHAPTER 2: LITERATURE REVIEW

This chapter will review academic work surrounding datasets used to analyse the movements of human populations. It begins by addressing the value of the census and the results it generates, and henceforth why alternate sources are necessary to reproduce similar results. The use of observational data sources will be characterised broadly, and the biases and limitations existent within these will be addressed. Specific applications of MPD in previous studies will then be offered through studies utilising Call Detail Records (CDRs) and studies analysing mobile phone locational data. Studies utilising locational data in the analysis of commuting behaviours have not yet been largely applied in a UK context, thus the review concludes by acknowledging how this avenue of data could provide an innovative alternative to manual data collection techniques derived from the census.

As this study focuses on assessing the accuracy of MPD in relation to its proximity to reciprocal census variables, definitions of commuting and its importance in social science research will not be offered. This is due to the focus on the uses of MPD, in which commuting is merely used as a tool to demonstrate this. Justifications for using commuting as a proxy for this are however offered within the *Methodology* (p. 20).

2.1. Why is the Census Important?

The history of census conduction which has spanned over a century since its introduction in 1901 has proved invaluable for contributing towards a complete source of information regarding our current population. Although its original intent was largely prompted by a desire to assess whether the population was expanding or contracting, over time responses have been fundamental by aiding governments, health authorities and other organisations how to plan and target resources and services more effectively (ONS, 2018). Its uses span across a variety of fields; including housing, transport, education, health and social research (Henley, 2011).

Its compulsory nature distinguishes it from other household surveys, as it covers the entirety England and Wales and abides to a sole questionnaire – enabling comparability of results between respondents, but also between censuses themselves conducted every ten years in late March. However, the whole operation costs £482m every decade, and increasing issues in collection techniques and tedious amalgamation of results often cause inaccurate representations of human populations, as by the time the data has been processed, people's circumstances have altered. This limitation was supported by the missing data regarding 900,000 men in the 2001 statistics – an issue which could not be pinpointed to an exact cause (Henley, 2011).

In an era of dynamic and continuous change, which has been demonstrated by the proliferation of COVID-19 which has altered the habitual movements of human populations due to government-imposed travel restrictions, the ten-year gaps between censuses appear inappropriate, and new observational datasets which have begun to emerge in the twenty-first century have questioned the feasibility of infrequent censuses (Novak et al, 2013). As the penetration of mobile phones have now proliferated across the globe, as in the UK alone seventy-nine percent of adults owned a smartphone in 2019, the uses of these emergent data sources are abundant within the study of human movements (Boyle, 2020).

2.2. Observational Data Sources

The costly nature of the census has inspired new ideas surrounding the uses of other, often administrative datasets as substitutes or supplementary sources to enable a 'refresh' on current statistics. Francis Maude (Britain's cabinet office minister in 2011) was an advocate of this view, as he claimed we needed to find "ways of doing this which will provide better, quicker information, more frequently and cheaper" (Henley, 2011, no pagination).

Blazquez and Domenech (2017) outline the relevance of the internet, smart sensors and smartphones in the digital era; technologies which are utilised on a daily basis that contribute to these emergent datasets documenting human movement and activities. The outcome of these technologies generates an influx of digitalised data, which can help reveal social and economic behaviours. Blazon (2019) estimated daily data generation to be at 2.5 quintillion bytes – hence the common reference to 'Big Data' in the literature, a term originating from the late nineties (Cox and Ellsworth, 1997).

2.3. Biases and Limitations within Observational Data

Despite their benefits, these new data sources bring new methodological risks. Questioning and testing the validity of data/models is an integral component of quantitative geography (Wilson, 1969).

Crampton et al (2013) highlights that despite the spatio-temporal benefits to new variants of '*big-data*', regardless of how 'big' these datasets may be, they are limited in their explanatory value. Gould (1981, p.166) outlines that these large datasets are often perceived to "speak for themselves", but in an era when emergent observational data sources are becoming increasingly available, the need to choose appropriate data in which patterns can be identified is crucial, as inanimate data cannot speak for themselves.

Crampton et al (2013) use geotagging locational data derived from Twitter to evidence the limited applicability of big datasets. Authors argue that studies that are reliant on social media data are naïve in

their extrapolation methodology, as they make generalised statements about society collectively. Haklay (2012) similarly outlines how sources of big geosocial data suffer inherent biases towards outliers, as irrespective of the number of geocoded tweets that are analysed, their usage is limited as these tweets only reflect a small fraction of all tweets. Other biases are outlined by Graham (2012), such as the small subset of all internet users represented on Twitter. As less than half the world's population are internet users (Bank my cell, 2020), making statements about society broadly using only these mediums of user-generated data is questionable, as Crampton et al (2013, p.132) identifies how this data often skews towards "wealthy, more educated, more white, and more male demographic".

2.4. Commuting Analysis using New Observational Data

Studies using MPD can be generated in two ways; from CDRs (call data records) or network data. CDRs refer to user activity on the device, such as calling, texting, or using the internet on a smart phone. Data is generated by the details of the cell ID of the closest cell-tower and accompanying timestamp. Network data can be generated when the phone is dormant, as updates can be gained from the phone as the device transfers between different cell towers and GPS satellites orbiting in space; indicating movement (ONS, 2017).

2.4a. Network Data

The ONS (2017) evaluated the potential of using MPD to estimate commuting flows and the modes of transport in which journeys were undertaken using administrative data. This project contributed to the assessment of whether the government ambition to ensure censuses after 2021 could be conducted using alternate sources of data was viable (ONS, 2014). The study compared commuting estimates inferred from a sample of MPD from private sector company Citilogik (MPD in this study was derived from MNO Vodaphone UK only) with equivalent 2011 TTW data. Anonymised mobile users were aged eighteen and over due to the minimum age for starting a Vodaphone contract, and the MPD extract were collected over a four-week period between March and April 2016. The study however was limited to commuter flows starting or ending in three London LAs; Southwark, Croydon and Lambeth (ONS, 2017).

The ONS (2017) primarily used linear regression models to compare how well MPD flows represented the TTW outputs. Areas under or overrepresented by the MPD were highlighted such as Lambeth, which was accursed to infrequent working patterns such as night or shift workers, in addition to commuters on holiday, ill or absent from work at the time of the study. The four-week period in which MPD was extracted was limiting in this sense as a broad overview of the year could not be given. It was also found that intra-LA flows were significantly higher within the MPD, which was accredited to the miss-

classification of students as commuters due to similar travel activities. They found that the inclusion of full-time students over sixteen in census estimates greatly improved the comparison.

The findings of the study highlighted an overall positive correlation between LA flows for longer distance commutes, and the largest commuter flow was between Lambeth (home) and Westminster (work), which represented 13% of all commuter living/working in Lambeth. Although MPD flows underestimated the scale of commuters identified by TTW data, patterns were broadly consistent relative to the largest flows between LAs.

Sadeghinsar et al (2018) also inferred average daily OD trips between home and work using phone GPS data and compared this to US Census summary tables. MPD was generated by over fifty mobile applications that anonymously collect geodata records, records were extracted from a two-week period in August 2017 in Houston (Texas). This study used GIS to produce comparative maps visualising distributions of commuters' home location using both datasets. Authors found a high correlation between the MPD and census summaries; confirming that GPS data can help understand the spatial distribution of journeys to address traffic concentration.

2.4b. CDR Studies

Other studies have utilised CDRs to study movement, such as Lai et al's (2019) exploration into the use of MPD for inferring national migration statistics based on anonymised CDRs in Nambia between 2010 and 2014, and how the derived statistics compare to census-based migration statistics. Conclusions of this study highlight the successful uses of MPD in complimenting traditional statistics due to their more up to date and localised nature, as high Pearson's coefficients were found between the census-derived population and mobile phone users for 2011, and migration flows between the census and CDRs. Authors postulated reasoning for areas where significant differences were found, such as the Zambezi region where the census derived more migrants than the CDRs, in which residents had been displaced due to flooding in the summer of 2010. This may have caused residents to be misclassified as migrants as they had moved to alternate locations before the census conduction in escape of flooding.

Similarly, Deville et al (2014) utilise CDRs of call and text messages to assess how aggregated MPD could be used to efficiently map population distributions. Effectiveness was measured through comparison to an existing downscale census data through remote sensing and geospatial data. This study has a larger coverage in comparison to Lai et al (2019) which solely analysed call records as opposed to text communications. However, both studies usage of MPD is limited by the density of cell towers which are more prolific in urban areas, catalysing bias due to reduced representativeness in rural areas. This limitation was acknowledged by Deville et al (2014) as they found lower precision in results in low-density areas.

Novak et al (2013) explored the use of MPD in the mapping of commuting patterns in Estonia. This study also utilised passive MPD (location information stored in billing memories) of call activity. The anchor point algorithm was used to identify home and work location based on the identification of locations regularly visited during the day (work) and night (home). Following identification, an OD matrix was created. Authors argued MPD represents a useful alternative to traditional sources where data may be missing or requires inappropriate/costly collection methods.

Demissie et al (2016) similarly utilise MNO data referring to CDRs of Sonatels customers in Senegal (January 2013). They utilise the same anchor point algorithm in identification of home and work as Novak et al (2013) to identify districts where the largest number of journeys originate from – such as Dakar and Thies, and the reciprocal districts with the lowest number of commuters beginning their journey. They also analyse inter-district OD flows due to the size of districts within Senegal, which are more prolific in the most populated districts.

Alternate classification algorithms were implemented by Alexander et al (2015) who presented ways to estimate daily OD trips using data extracted from triangular mobile phone records in Boston over a two-month period during spring 2010. These anonymised results were clustered into three categories; home, work and other dependent on frequency, time of day, day of week. Authors found the size of areas used to aggregate trips was a fundamental factor effecting correlations between the sources.

2.5. Conclusions from the Literature

Potential uses of mobile devices to study human populations have been highlighted amongst the literature, although there appears an apparent gap in the research regarding the uses of application-based mobile phone GPS data, which provides a higher degree of spatial accuracy and temporal frequency than CDRs. Studies utilising CDRs are often critiqued as intra-area trips cannot be identified in the case of a single cell tower covering a large surface area (Demissie et al, 2016). GPS data however can be generated without an internet connection due to GPS satellites orbiting the earth, opposed to static cell towers.

Lack of these locational based MPD studies is especially applicable in the context of the UK, as the majority of studies outlined refer to international examples. Henceforth, the utilisation of MPD appears a promising area of research to analyse commuting patterns in UK cities.

CHAPTER 3: DATA

3.1. Huq-Derived Flow Data

To assess whether MPD could produce comparable commuting statistics, a dataset was provided by huq of 65,054 anonymised records between the calendar year (1st January to 31st December) 2019. Each record, or geo-data event (GDE) comprises of an origin; the smallest aggregate level output area (OA) was used, a destination (one of eight chosen workplace zones), a flow column (the interaction between OA and WZ), the number of residents in the origin and the number of active works in the origin. Home and work locations were identified through anchor point algorithms (Novak et al 2012; Alexander et al 2016), meaning the location the device spends the most time at during the day in a given year is their inferred work location, and the reciprocal home location is defined as the location in which the device spends the most time during the night.

Table 1: Sample of huq dataset

output_id	origin_total_residents	origin_active_workers	destination_id	OD_flow
E0000029	2	1	E33010438	1
E0000029	2	1	E33012365	0
E0000030	1	0	E33010439	0
E0000031	3	2	E33010440	2

3.2 Census Workplace Statistics

The 2011 dataset used provides estimates of the usual residents of England and Wales aged over sixteen and in employment a week before the census (reference date 27th March 2011). A usual resident of the UK is defined as anyone who on census day was in the UK and had stayed or intended to stay for twelve months or more, or who had a permanent UK address but was outside the UK but intended on returning within twelve months (ONS, 2011). Of these, their work location was obtained by the question "*In your main job, what is the address of your workplace"* – the only exclusions listed include working locations with "no fixed place", "mainly work at or from home" and "offshore installation" - see Appendix 1 for questionnaire extract (ONS, 2011, p.10). The classification of residents into an output area (OA) provides an origin for their travel route between home and work. Data regarding the 2011 outputs was extracted from table WF02EW *Location of usual residence and place of work (with 'outside UK' collapsed)* via the UK Data Service (2011) query builder.

Table 2: Sample of 2011 TTW dataset

output_id	destination_id	OD_flow
E0000029	E33010438	2
E0000029	E33012365	2
E0000030	E33010439	1
E0000031	E33010440	3

To enable comparison between the datasets, the reciprocal information provided by huq (total residents at the origin, total active workers at the origin) was downloaded through Infuse relating to the proportions of residents and active workers in different locations recorded by the 2011 census. Immediate definitional differences were evident here due to the likelihood of slightly different categorisation systems of *economically active* by both huq and the census, as the ONS (2011) define an economically active person as; a resident aged sixteen or over and in employment (as an employee or self-employed) a week before the census; not employed but seeking work and prepared to start within the next two weeks; or not employed but awaiting the start date of an obtained job. Full-time students are not included in this category.

The definitional methodology huq implement to categorise a device's user as *economically active* was unable to be acquired due to lack of communication with huq's CMO (Chief Marketing Officer) since the proliferation of COVID-19. However, it has been inferred from external publications by huq that the workplace of a device is the location that it spends the most time during the day in a given year (ensuring a separate night-time location is evident), and the primary overnight location is regarded as home. Although it is unclear how many hours a device needs to spend at the inferred work location to be classified as economically active, this was disregarded as a substantial limitation due to the aim of identifying broader patterns.

The format of the TTW data also differs to that provided by huq as all OD pairings in the TTW data possessed a flow count of at least one due to specific destination refinements implemented when downloading data (i.e. the origin referring to any OA in England, and the destination referring to any of the eight WZs). This output produced 6576 rows of data each referring to a single interaction between an origin and destination. Due to the sample of huq's data from 2019, every observation referring to one row of the data did not necessarily return a positive interaction with one of the eight WZs.

Table 3: Summary of the input datasets

Source	Timeframe	No. Observations	No. Flows
Census	27th March 2011	6576	6576
Huq	1st January 2019 to 31st December 2019	65,054	45,792

Note: N. "flows" refers to the flow between OD pairings, where a positive interaction was recorded and "N. observations" refers to the total row count.

3.3. Privacy

Ethical and legal aspects of individual privacy should be outlined due to the use of MPD not willingly supplied by individuals in the same format as the census collection. Data used in this study does not entail any unique information which could cause linkage to a particular person. The anonymised nature of the data is ensured by a unique device ID attributed to each device studied, which is not linked to any individual characteristic such as name, age or job. The spatial resolution of the data is also not specific enough to pinpoint exact locations, as the average location accuracy reported by huq is 65m (Huq, 2019). Data referring to mobile phone users analysed within this study has also been granted permission by users following the installation of mobile applications which request access to location services. Due to these factors, the privacy of individual mobile phone users is ensured in the context of the EU General Data Protection Regulation (GDPR) on the security of personal information.

CHAPTER 4: METHODOLOGY

4.1. Commuting

Uses of locational data are often exercised within studies analysing human mobility. As commuting trips account for the largest portion of journey purposes during peak hours (Polzin et al, 2015), understanding these trips is a fundamental element in managing transport demand, enhancing public transport efficiency and improving travel time forecasting (Sadeghinasr et al, 2015). Commuting henceforth was chosen as the variable which would assess the value of the MPD extract due to the ability of creating OD trip matrices using locational MPD, which can be compared to reciprocal OD matrices extracted from the census statistics.

4.2. Origins

Census-based population statistics relating to the usual residence and work location of workers are the predominant tool for analysing commuting patterns in England and Wales. The home location of workers is documented using a series of hierarchical output zones, with Output Areas (OAs) being the smallest spatial scale. OAs were produced in accordance with the 2001 census and were devised from postcode units with the intention of generating output zones with consistent population and household counts (with a 125 households target), social homogeneity and geographical compactness (Coady, 2014). Due to their value in the study of residential populations, this aggregate scale was selected for the origin of both datasets, as OAs can be scaled up to reflect larger geographic areas (there were 2543 OAs in Leeds at the time of the census which contribute to 107 larger MSOAs). A visualisation of this scaling up of aggregate areas is offered below, as the MSOA Headingley is formulated by 26 OAs.

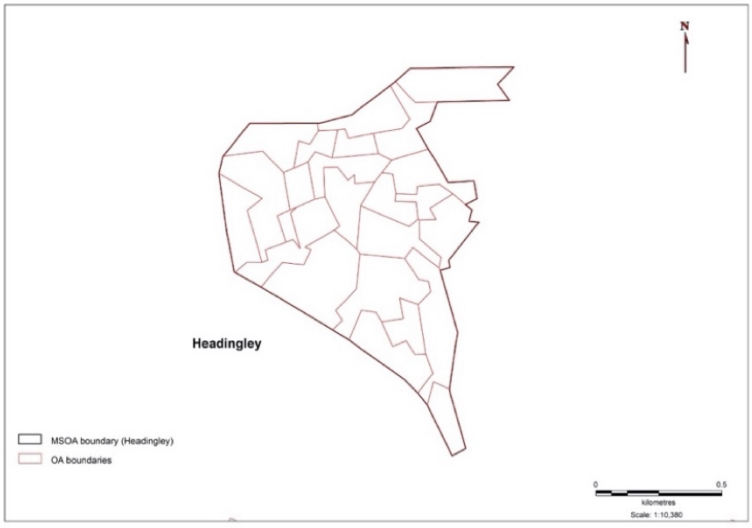

Figure 1: Example of OAs which collate to form an MSOA

Aggregation was completed through Excel using the VLOOKUP function to pair OA codes evident within both datasets to the larger MSOA and LAD scale. The OA lookup table was downloaded via the UK Data Service website (*Output Area 2011 to Lower Layer Super Output Area 2011 to Middle Layer Super Output Area 2011 to Local Authority District 2011 (E+W) Lookup*). This was conducted to give a clearer visualisation of journeys between MSOA:WZ and LAD:WZ due to the number of OAs evident within both datasets.

4.3. Destinations

Although useful for the study of residential populations and area-based geodemographics, OAs fail to adequately reflect workplace populations as many residential areas fail to meet the minimum statistical disclosure thresholds required for the attainment of workplace statistics (Mitchell, 2014). The introduction of the new geographical unit of Workplace Zones (WZs) within the 2011 Census alleviated this issue, as while OAs were designed to reflect consistent numbers of residential populations, WZs were introduced to contain consistent numbers of workers based on their work location (ONS, 2014). Using the existent OA scale of aggregation, 53,578 WZs (each with a mean worker count of 493) were produced by splitting, combining or retaining these OAs (Martin et al, 2017). The difficulty of applying residential-based aggregative levels to workplace statistics is evidenced below, as large areas in north and south Leeds possess zero WZs but are populated with residential dwellings.

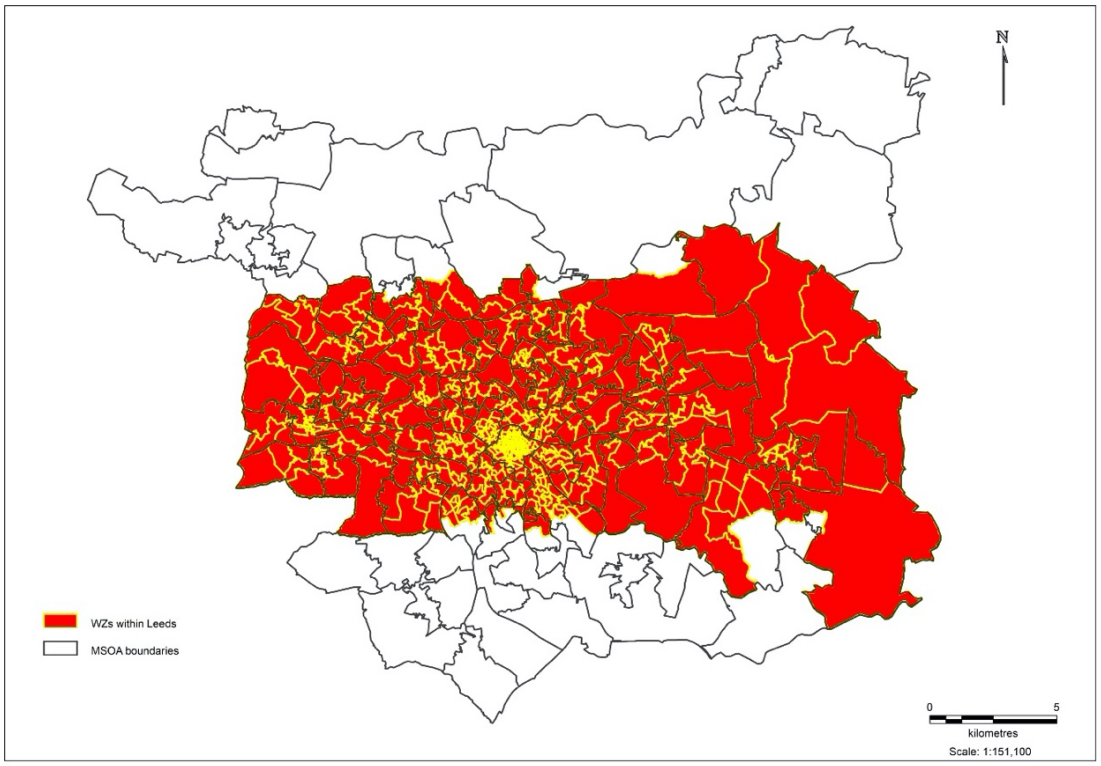

Figure 2: Location of WZs compared to MSOA aggregate level

Due to the value of these new workplace related geographical units as demonstrated by Berry et al (2016) in retail, Woods (2017) in understanding workplace accidents and London Boroughs such as Hackney Borough Council (2015) in utilising WZs and supplementary data to inform their Transport Strategy, this spatial scale will also be applied within this study. Eight individual WZs in total were chosen within Leeds, half of which collated to form 'WZ A', and half contributed towards the collation of 'WZ B'.

Table 4: Breakdown of individual WZs used to create separate zones

'Supergroup' classification	WZ zone code
WZ A	E33009667
WZ A	E33010352
WZ A	E33009660
WZ A	E33012365
WZ B	E33010438
WZ B	E33010436
WZ B	E33010440
WZ B	E33010439

The two supergroups outlined within *Table 4* represent the two destinations used for analysis within this study, reasons for their selection will now be characterised.

4.3a. WZ A, University Zone

This zone covers the majority of the University of Leeds campus, with the northern boundary at Hyde Park Road, and the southern boundary just north of the Worseley Building on Clarendon Way.

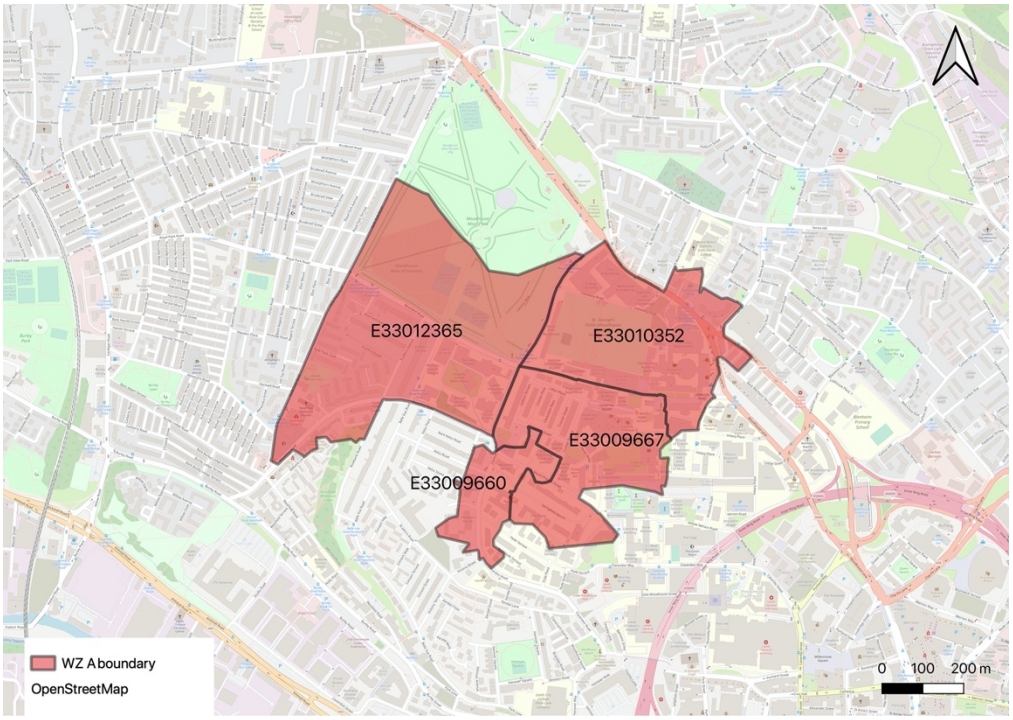

Figure 3: Location of WZ A (University) in Leeds

This zone was chosen to reflect a collective workplace zone in which limited changes in commuting patterns were expected between the 2011 census and the 2019 MPD extract due to the university's locational prominence in this area of the city since 1904 (University of Leeds, 2020). The four WZs cover the majority of the university campus and the main university address. However, the boundaries exclude the Nexus building, Leeds Dental Institute (Worseley Building), and The Edge sports facilities. This was disregarded as an implicating factor in the validity of results due to the wording of the census questionnaire "*in your main job, what is the address of your workplace*" (See Appendix 1 for the page of the census questionnaire this refers to), as regardless of the subsection university workers spend the most time during their working day, if the main University of Leeds address was given on their census questionnaire then results would not be implicated.

This however is not transferable to the validity of the MPD, as the data collection technique used to produce this dataset relies solely on locational GPS; and so certain University workers spending their working day within the Worseley Building would be excluded from the results – a limitation of this WZ.

4.3b. WZ B, Industry Zone

This zone is located in a more centralised location in Leeds (1.5km south of WZ A). WZs within this location were chosen to reflect an area which was more likely to identify increases in commutes to this area, due to the opening of various managerial firms since the conduction of the 2011 census. Examples of these include Q5 (2018), KPMG (2015), and PwC (2015) (Consultancy UK, 2015; 2019).

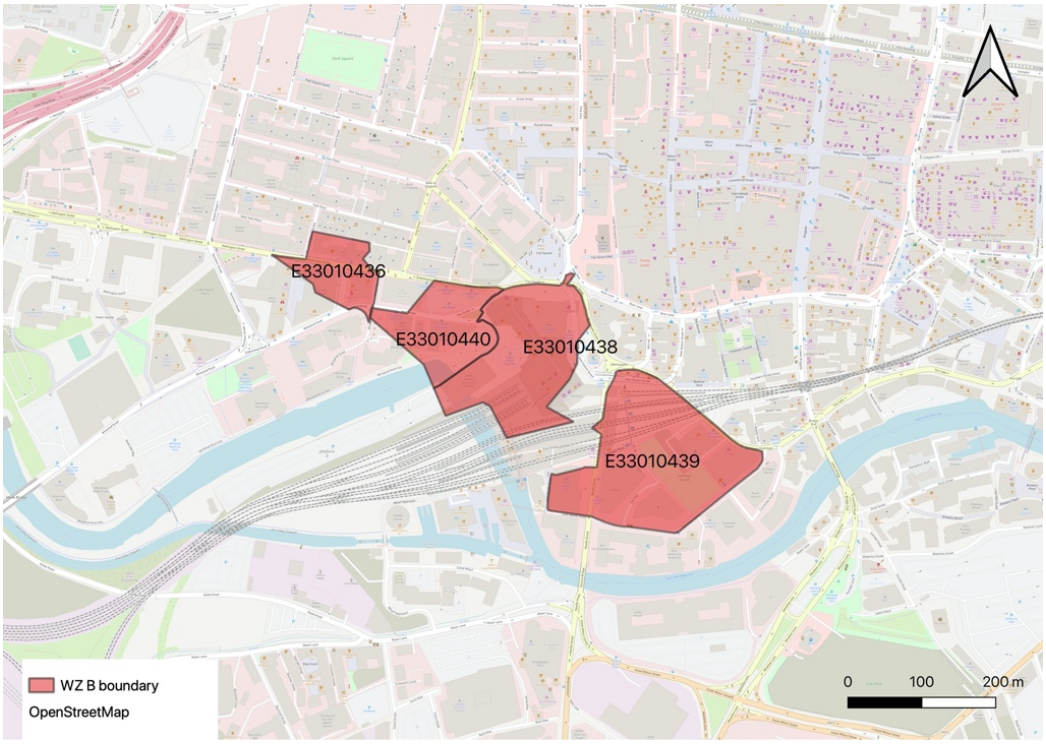

Figure 4: Location of WZ B (industry) in Leeds

The boundary (as defined within *Figure 4*) offers an alternate WZ that can be used in comparison alongside WZ A. The dual use of two supergroups (WZ A, WZ B) enhances the validity of results and ensures any patterns identified are not unique to one area and are likely to be reliable indicators of commuting patterns, reducing issues of biases and limitations within observational data studies as outlined by Crampton et al (2012) and Haklay (2013).

4.4. The Study Area

The locations of the two WZs are offered below in relation to Leeds MSOAs. MSOA was the primary aggregate scale applied within this study following aggregation from OA level. The 107 MSOAs which formulated the basis for analysis are evidenced below.

Figure 5: Location of WZ A and WZ B within Leeds MSOAs

Although Alexander et al (2015) found the scale of aggregation used within analyses can greatly affect correlations between the datasets as the aggregational unit used may disguise inner-MSOA differences, this spatial scale appeared the most suitable as although analyses at smaller spatial scales (e.g. OA) would have been computationally possible, due to the scale of OAs in Leeds alone (in 2011 there were 2543 OAs in Leeds), the analyses would have been complex with thousands of OD flows in which distinguished patterns would be hard to identify.

4.4a. West Yorkshire

The position of Leeds within the county of West Yorkshire (comprised of five metropolitan boroughs; Bradford, Calderdale, Kirklees, Leeds and Wakefield) is outlined below. This is the wider region in which the study investigates. Although the origins (i.e. the home location) of commuters were not constrained to one location and covered the entirety of England and Wales, it is unlikely that large proportions of workers undertake commutes which exceed 20 miles, as the ONS (2014) found the average distance travelled to work in England and Wales in 2011 was 15km (9.32 miles). Although the average commuting distance is rising between censuses, between 2001 and 2011 this average only increased by 1.6km – epitomising the unlikely nature of extensive average commute lengths.

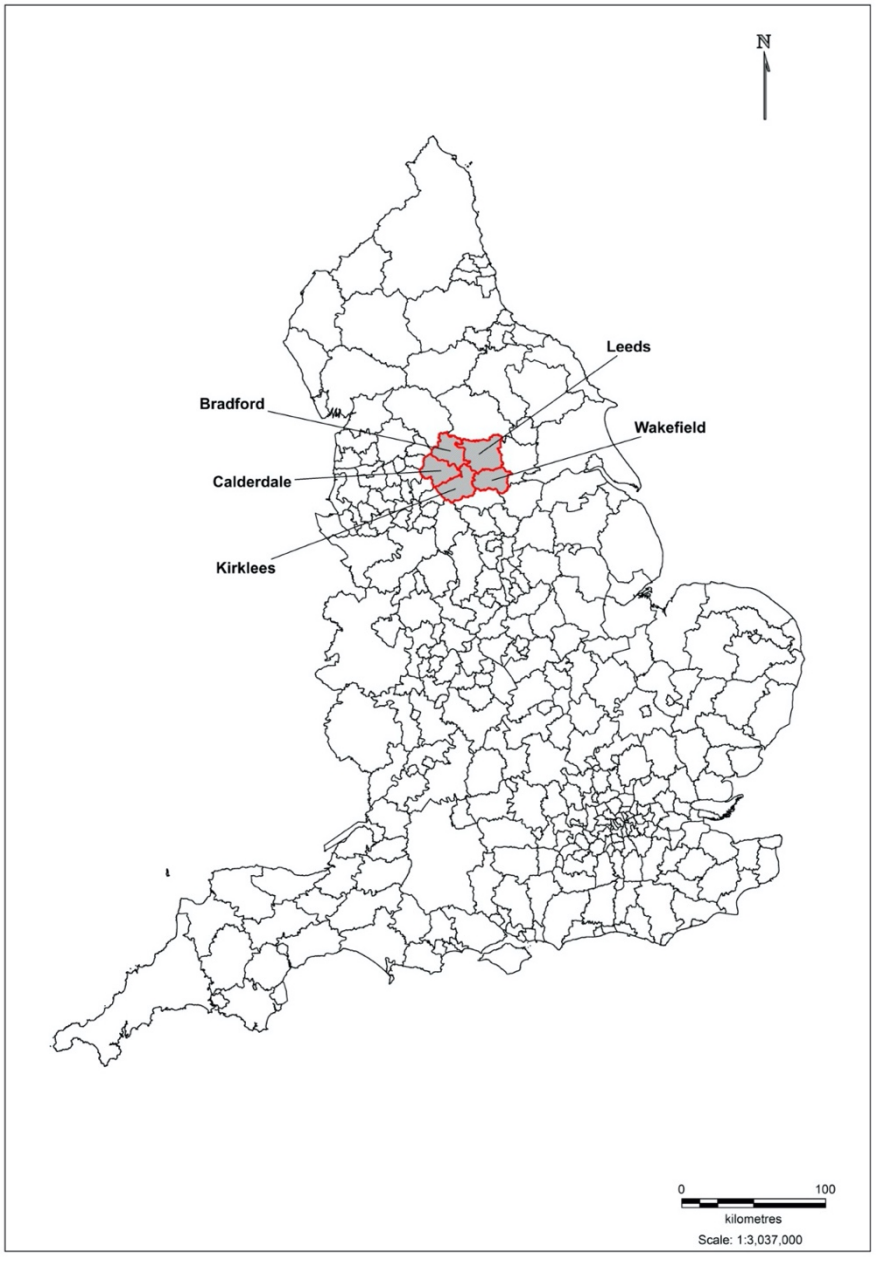

Figure 6: Location of West Yorkshire in England

4.5. Methods for Analysis and Reasons for Use

4.5a. Softwares

The GIS operating system used for all cartographic visualisations within this study was *MapInfo Pro 17.0.* The only exception to this was the creation of maps outlining the destination area of WZ A and B (*Figures 16* and *17*) due to MapInfo's restricted use of the hybrid feature on some machines complicated by COVID-19. To characterise the surrounding urban environment of the workplace zones, *QGIS3.12* was utilised for its functionality with *OpenStreetMap.* Data manipulation and aggregation was completed within *MS Excel.*

4.5b. Raw Counts

To investigate *Objective 1* (p. 11)*,* flows at raw count level were first analysed within the wider region to identify key areas underrepresented by huq due to lack of OD flows attributable to whole LADs. Underrepresentation was characterised by comparisons to census data outlining commuting flows in 2011 to assess the scale of representation offered by huq. GIS was used was to create choropleth maps highlighting LADs in which huq's data failed to derive any flows of commuters travelling towards WZ A and WZ B.

Inner Leeds was then the focus of the analyses, with tables of rankings produced to outline the top five MSOAs recorded by both the census and MPD as experiencing the largest OD flows to both WZs. The degree of similarity was then assessed based on the correlation between the ascending ranked order (for example a rank difference of -2 would demonstrate the MPD ranked that MSOA two positions lower than the census data). This supported the investigation into *Objective 1.*

As this raw count analysis fails to consider the proportions of active workers within the origin, Location Quotients (LQs) were then devised to assess concentrations of workers in home MSOAs in comparison to concentrations in the wider LAD.

4.5c. Location Quotients

LQs were used to measure the spatial concentration of workers at MSOA level in comparison to the concentration of workers at the wider LAD level. For example, a LQ of 1 reflects the same concentration as the LAD of Leeds, <1 highlight underrepresentation of the group in the home MSOA, and >1 demonstrates over representation of the group in their home MSOA in comparison to the LAD.

For both workzones and using both datasets, the following formula was applied:

$$\frac{\text{Count of workers in MSOA travelling to WZ / Total active workers in MSOA}}{\text{Count of workers in Leeds LAD travelling to WZ / Total active workers in Leeds}}$$

In the comparison of LQs, MSOAs in which huq did not derive a value for the flow column were removed, as the output was zero due to missing data. This was only in select cases:

Table 5: MSOAs removed from MSOA analysis

WZ A:	WZ B:
Wetherby West	Wetherby West
Wetherby East and Thorp Arch	Tingley West and West Ardsley
Boston Spa and Bramham	Drighlington and Gildersome West
Colton, Austhorpe and Whitkirk	
Tingley West	
Robin Hood, Lofthouse and Middleton Lane	

The removal of these MSOAs within the specific analysis of LQs was not considered a substantial limitation, as census results outlined LQs very close to 1 for all MSOAs listed, so their removal was unlikely to yield unreliable results as no significant under or over representations were evident in these areas in 2011.

Comparison of LQs was achieved through studying rank difference rather than difference between LQ outputs themselves, as often patterns were replicated by both datasets regardless of the LQ output difference. Analysing the differences between LQ outputs would be unfair due to the enhanced coverage achieved by the census due to its compulsory format, causing the LQs calculated by the MPD to appear inaccurate as the MPD reflects a smaller proportion of the population (highlighted within the *raw difference* column in *Table 6*). Difference in rankings henceforth appeared a more accurate measure of testing similarity in patterns.

Table 6: Difference between comparing raw difference and rank difference

MSOA Name	LQ (census)	LQ (huq)	Raw Difference	Rank (census)	Rank (huq)	Rank difference
University & Little Woodhouse	5.28	1.70	3.57	1	5	-4
Hyde Park Corner & Woodhouse Cliff	5.05	1.86	3.19	2	4	-2

The LQ analysis contributed towards the achievement of *Objective 1* and *Objective 3. Objective 3* was then developed further within the *Discussion* (p. 47).

4.5d. Socio-Economic Variables

Objective 2 was achieved through analysis of socio-economic data obtained via the Infuse website and refers to the 2011 census statistics. Despite the weaknesses of the census in regard to its singular collection once every ten years, its broad coverage of the nation attained through The Census Act (1920) which legislates a fine of up to one thousand pounds for anyone refusing to complete their census form outlines its value as the widest population-level based survey in the UK (ONS, 2011). Although it would have been desirable to utilise annual data regarding to socio-economic variables, no other source possesses the same degree of coverage as the census.

Data referring to all variables listed within *Table 7* were obtained at MSOA level for Leeds. Percentages of these categories within each MSOA were calculated using the total number of all usual residents within the MSOA.

Table 7: Socio-economic variables used to measure correlation

Variable	Description	Manipulation
Deprivation, classification of household	Dimensions of deprivation used to classify households are indicators based on four household characteristics. (These dimensions are outlined within *Appendix 2).*	This variable was split into two categories – 1. Not deprived in any of the four dimensions 2. Deprived by one or more of the dimensions

Age	Age is derived from the date of birth question and is the person's age at their last birthday at 27th March 2011.	Age groups studied were - 1. Ages 0-9 2. Ages 16-19 3. Ages over 65
Schoolchildren and full-time students	Derived from the age question listed above.	No manipulation.
Travel to place of work, means of	The means of travel used for the longest part, by distance, of the usual journey to work.	Modes of transport analysed included – 1. Car/van 2. Bus/coach 3. Train 4. Bike 5. On foot 6. Mainly work from home
Distance travelled to place of work	The distance in kilometres between a person's residential postcode and their workplace postcode (measured in straight line).	Average distance (km) per MSOA calculated.
Hours worked	The number of hours worked that a person aged 16 or over, in employment, the week before the census worked in their main job.	This variable was split into two categories - Full-time: 31 hours or more per week Part-time: 30 hours or less per week
Occupation	A person's occupation relates to their main job and is derived from either their job title or details of the activities involved in their job.	The top three occupation levels were combined – 1. Managers, directors and senior officials 2. Professional occupations 3. Associate professional and technical occupations
Social Grade (persons)	The socio-economic classification used within Market Research industries. Details of the differing levels of social grade are outlined within *Appendix 2.*	Two variables were extracted from this category – Social Grade AB: Higher & intermediate managerial, administrative, professional occupations Social Grade DE: Semi-skilled & unskilled manual occupations, Unemployed and lowest grade occupations
Industry	The industry in which a person aged 16 and over works relates to their main job.	Two categories were devised from this variable - Industry A (typically sporadic workplaces): Construction, transport, communication Industry B (typically uniform workplaces): Public administration, education, health; financial, real estate, professional, administrative activities
Population (usual residents)	The usual resident population at census day (27th March 2011).	No manipulation.

Note: for further details of the four measures of deprivation and Social Grade Categories see *Appendix 2.*

The CORREL function in Excel was implemented to gain correlation coefficients for these variables. This measured the relationship between the LQ rank difference and proportions of each group by MSOA. This produced an output between +1 and -1, with closeness to +1 signifying a strong positive correlation and closeness to -1 demonstrating a strong negative correlation (Rumsey, 2016). Scatter plots were then formulated in Excel of the most significant relationships between the variables for each WZ to complete the analysis for *Objective 2.*

4.6 Limitations and Alternate Methods of Analysis

4.6a. Linear Regression

Linear regression models could have been implemented as applied by the ONS (2017) and Lai et al (2019). Such models would measure how well MPD flows represent the 2011 census outputs, as the gradients of the regression lines outline the percentage of flows that are represented, such as seventy percent of flows for a given MSOA. Due to the computational constraints and lack of assistance following the outbreak of COVID-19, this analysis was unable to be conducted.

Regression analysis could have been implemented to assess the relationship between rates of over representation by huq and socio-economic factors using statistical regression in Minitab. This would measure the scale of representation in a spatial zone alongside socio-economic variables, to produce a p value distinguishing the significance of the relationship (a p value of 0.05 or below is deemed as significant, and infers the relationship is not coincidental). Regression analyses within social science studies have been widely applied in crime-based studies (Corcoran et al, 2007), health (Congdon, 2012) and transportation (Chen, 2017).

4.6b. Mobile Phone Bias

Although mobile phones appear increasingly ubiquitous, they not owned by every individual in the world. Realistic figures of the numbers of people owning a mobile phone in England is difficult to gauge, as surveys are often only conducted by residents exceeding the age of eighteen. According to research by Internet Matters, age ten is the average age in the UK for when a child obtains a smartphone (Sharma, 2020). Thus, a large proportion of the mobile phone owning market are unaccounted by official surveys/statistics. Irrespective of this, around ninety-five percent of households in the UK were recorded as owning a mobile phone in February 2020, leaving five percent of households without a mobile device (the equivalent of 1,380,000 households), (Statista, 2020).

The justification of measuring the relationship between socio-demographic factors and weak correlations to the census is henceforth relevant to assess the potential exclusion of certain groups - most significantly younger residents who are excluded from official figures and older residents who are less likely to own a mobile phone. Age categories of over 65 and 0-9 were studied as a result (it was acknowledged that these residents would now be nine years older given the nine-year duration since the census conduction).

Mobile phone bias is also relevant to mobile users, as some users access their device more frequently in a given day than other users. Although this is less of problem within this study due to the ability of location services to access the device the location when the device is dormant, this is a limitation of the uses of MPD more generally.

Further limitations of this study in relation to the significance of results is offered within the *Discussion* (p. 47).

CHAPTER 5: ANALYSIS

This chapter will outline the results using predominantly the MSOA:WZ spatial scale to outline key differences between the datasets. Raw counts of OD flows will first be analysed to identify whole LADs experiencing profound under/over-representation by huq, which will be followed by comparative location quotients (LQs) to ensure the number of active workers in the home MSOA have been acknowledged relative to the OD counts. Socio-economic factors will then be used to test correlations between MSOAs exhibiting prominence of certain groups and large differences to the LQs identified by the census.

5.1. Differences Between Raw Counts

5.1a. The Wider Region

The wider coverage gained by the census was immediately certified when visualising the raw OD flows between MSOA and WZ A, as *Figure 7* demonstrates the wide sphere of influence of home MSOAs in which commuters to WZ A derive from. Contrastingly, huq's data (*Figure 8*) is much more limited in its coverage of the region, with immediate underrepresentation of OD flows originating from *York*, *Selby*, *Barnsley*, and *Doncaster* as zero commuters were recorded as deriving from these LADs during 2019. Partial under representations of OD flows within LADs are also attributable to *Wakefield*, *Kirklees*, *Calderdale* and *Bradford*.

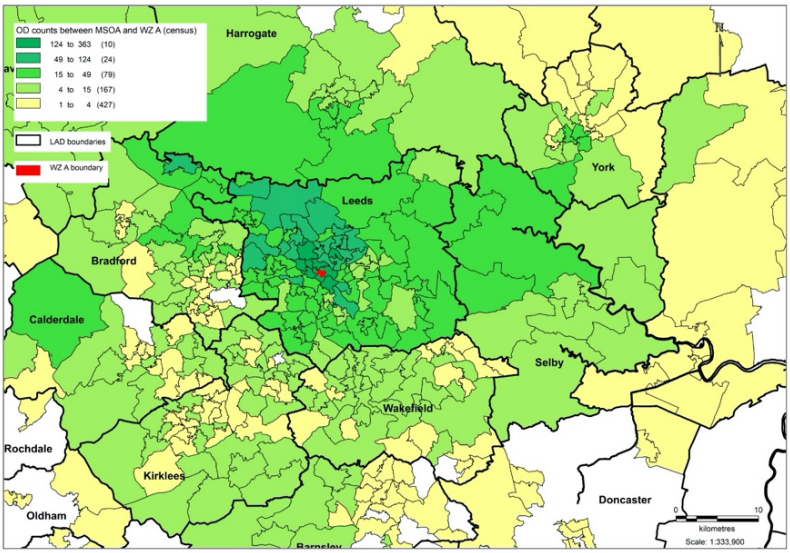

Figure 7: Commute counts from MSOA to WZ A (census)

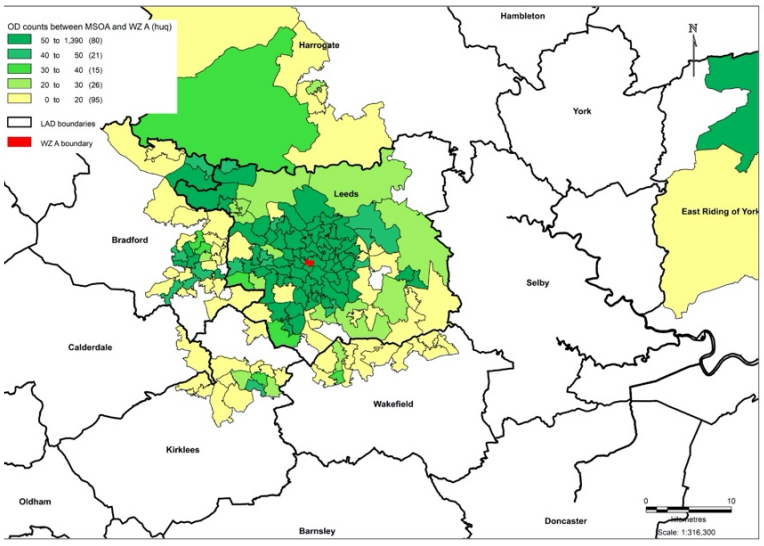

Figure 8: Commute counts from MSOA to WZ A (huq)

Table 8: Most significant areas of under-representation (WZ A)

LAD	2011 Count to WZ A	Huq 2019 count	% Difference
Selby	102	0	-100.0%
York	167	0	-100.0%
Barnsley	59	0	-100.0%
Doncaster	13	0	-100.0%
Wakefield	244	150	-38.5%
Kirklees	283	204	-27.9%
Calderdale	113	1	-99.1%
Bradford	541	881	62.8%

The only LAD exhibiting over representation of flows to WZ A by huq is *Bradford*, as other LADs are significantly lower than the 2011 census count.

General patterns in underrepresentation of certain LADs were supported by WZ B analysis, due to similar lack of OD data referring to *Barnsley* and *Doncaster.* Huq's data referring to WZ B did however provide a more robust coverage of the region. This enhanced coverage of LADs including *Selby*, *York*, and *Wakefield* is evidenced by *Figure 10*, as more MSOAs were identified as the home location of commuters to WZ B, validated by the census which also documented commuters residing in these areas.

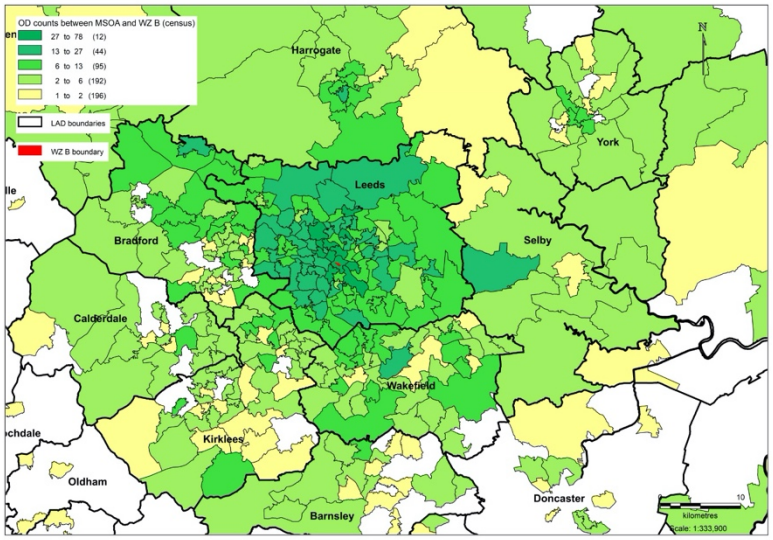

Figure 9: Commute counts from MSOA to WZ B (census)

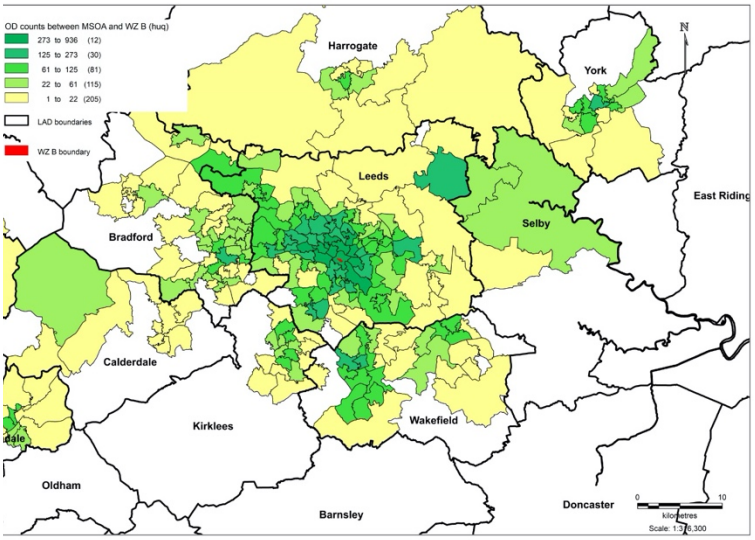

Figure 10: Commute counts from MSOA to WZ B (huq)

Table 9: Most significant areas of under-representation (WZ B)

LAD	2011 Count to WZ B	Huq 2019 count	% Difference
Selby	39	62	59.0%
York	74	723	877.0%
Barnsley	37	0	-100.0%
Doncaster	20	0	-100.0%
Wakefield	190	1522	701.1%
Kirklees	167	641	283.8%
Calderdale	55	91	65.5%
Bradford	252	1522	504.0%

MPD counts of commuters to WZ B were generally higher than WZ A (*Table 9*). For example, in *York* 723% more commuters were categorised as travelling to WZ B than to WZ A. This may reflect the changes induced by the recent opening of managerial firms bringing commuters from York into this area since 2011 as highlighted within the *Methodology.*

5.1b. Inner Leeds

Within Leeds, spatial analysis allowed comparisons between the most densely populated MSOAs with commuters to each zone according to both datasets. The most densely populated MSOA of commuters to WZ A was recorded as *Hyde Park Corner and Woodhouse Cliff* by the census, which was similarly recorded as second most densely populated MSOA by huq.

Table 10: MSOAs with highest commute counts to WZ A (2011)

MSOA code	Name	TTW count	TTW rank	Huq rank
E02006861	Hyde Park Corner & Woodhouse Cliff	363	1	2
E02006852	Far Headingley & Weetwood	322	2	15
E02002392	University & Little Woodhouse	231	3	1
E02002383	Hyde Park	201	4	4
E02002384	Woodhouse & Little London	185	5	3

Table 11: MSOAs with highest commute counts to WZ A (2019)

MSOA code	Name	Huq count	Huq rank	TTW rank
E02002392	University & Little Woodhouse	1388	1	3
E02006861	Hyde Park Corner & Woodhouse Cliff	1384	2	1
E02002384	Woodhouse & Little London	1295	3	5
E02002383	Hyde Park	1182	4	4
E02002385	Burley	908	5	7

The degree of similarity between the TTW and MPD data is characterised by *Table 10* and *Table 11*, as both datasets exhibit the same top five MSOAs with the highest frequencies of WZ A commuters (with the exclusion of *Far Headingley and Weetwood* and *Burley*). As *Far Headingley and Weetwood* reflects the MSOA which has experienced the largest decrease in commute flows between 2011 and 2019, there may be a systematic reason for this change. In general however, huq's data appears accurate due to the repeated MSOAs recorded as experiencing the largest OD counts by both the census and huq. Frequencies however should be questioned, as the ONS (2012) outline that the average population size for each MSOA in England and Wales is 7787, thus the MPD is very comprehensive for some MSOAs as OD frequencies exceed 1000 in almost all MSOAs outlined within *Table 11.*

Intra-MSOA flows are more apparent within huq's data, as the largest number of commuters derive from the MSOA *University and Little Woodhouse* in which WZ A resides. However, this is not a significant difference due to the proximity of the most densely populated MSOA of WZ A commuters recorded by the census (*Hyde Park Corner and Woodhouse* Cliff), as illustrated below.

Figure 11: Proximity of most densely populated MSOA origins in comparison to WZ A

Figure 11 identifies generally short commutes undertaken by commuters to WZ A, as all MSOAs evidenced within *Figure 11* are within 3km of WZ A.

Analysis for WZ B supported the finding that highest proportions of commuters derive from the same MSOA (or neighbouring MSOAs) in which the WZ resides, as the MSOA with the largest frequency of OD flows was *Leeds City Centre* for both the 2011 TTW data and MPD (where WZ B is located). The top five most densely populated MSOAs with commuters ending their journey at WZ B are characterised below.

Table 12: MSOAs with highest commute counts to WZ B (2011)

MSOA code	Name	TTW count	TTW Rank	Huq rank
E02006875	Leeds City Centre	78	1	1
E02002411	Holbeck	58	2	22
E02006852	Far Headingley & Weetwood	42	3	13
E02006876	Leeds Dock, Hunslet & Stourton	40	4	3
E02002344	Primley Park & Wigton Moor	38	5	319

Table 13: MSOAs with highest commute counts to WZ B (2019)

MSOA code	Name	Huq count	Huq rank	TTW rank
E02006875	Leeds City Centre	936	1	1
E02002404	East End Park & Richmond Hill	731	2	7
E02006876	Leeds Dock, Hunslet & Stourton	391	3	4
E02002385	Burley	351	4	11
E02002383	Hyde Park	322	5	36

The most noticeable difference in rankings is *Primley Park and Wigton Moor*, which has a rank difference of 314 - the most significant difference in results outlined in *Table 10, 11, 12* and *13.* Rankings in general are more different than those outlined within WZ A analysis, which may signify a more profound change between commuting patterns in 2011 and 2019 for commuters ending their route in WZ B. It may also be attributable to a misrepresentation of commuters by huq due to the location of WZ B within Leeds City Centre which attracts a diverse range of person types for other, non-work based purposes.

WZ B commuters reside in more dispersed MSOAs in comparison to WZ A and thus have longer average commuting journeys. The MSOA origins with the highest frequency of journeys to WZ B are visualised below.

Figure 12: Proximity of most densely populated MSOA origins in comparison to WZ B

To quantify the extent of these potential overrepresentations, comparative location quotients (LQs) have been calculated to incorporate the total active workers in the home MSOA to identify proportions of workers travelling to each WZ.

5.2. Location Quotients

5.2a. WZ A

Overall, the census identified that 1.78% of the total active workers in Leeds were travelling to WZ A for work. In comparison, the MPD identified that 19.14% of the active workers identified by huq travelled to WZ A for work. This considerable difference reveals an extensive overrepresentation of journeys by huq.

Census data (*Figure 13)* revealed high concentration of workers in northern MSOAs surrounding WZ A, and within the MSOA in which this WZ resides (*University and Little Woodhouse*). The MPD (*Figure 14)* reveals a more even distribution of workers across MSOAs, but still with highest concentration surrounding the MSOA of WZ A.

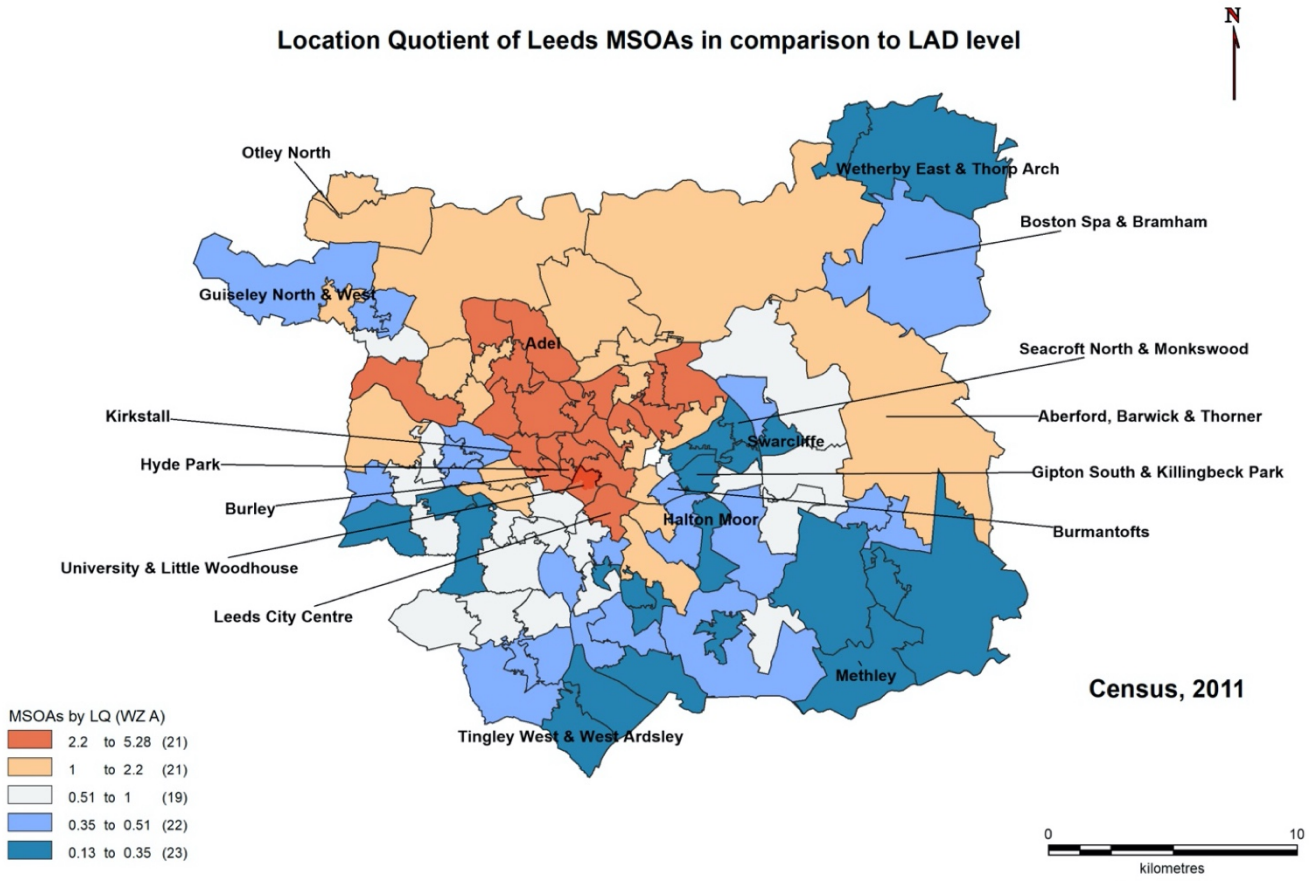

Figure 13: LQ of MSOA/LA level (census), WZ A

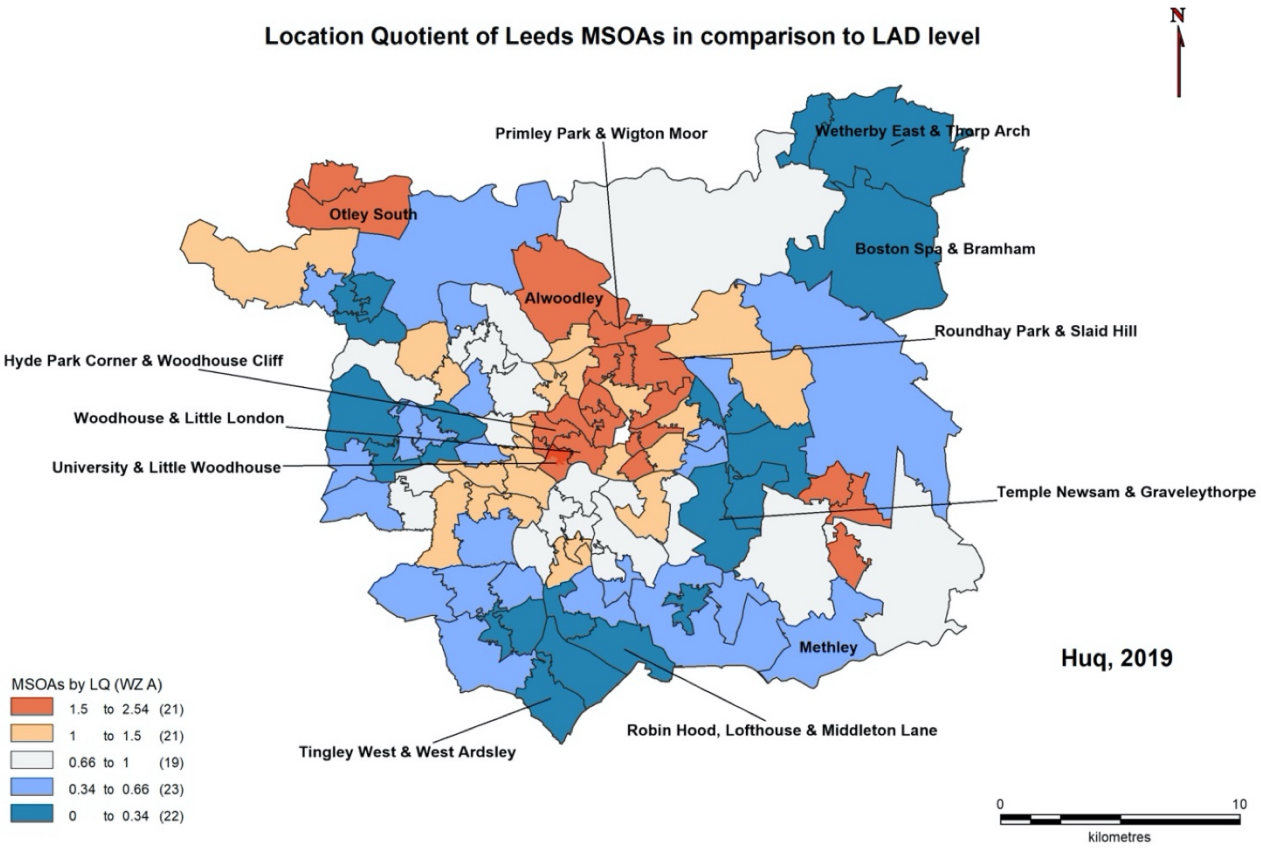

Figure 14: LQ of MSOA/LA level (huq), WZ A

Table 14: MSOAs with largest under representation of flows by huq (LQ)

MSOA	Rank (census)	Rank (huq)	Rank difference
Calverley & Farsley North	41	99	-58
Adel	10	57	-47
Cookridge & Holt Park	18	65	-47
Lawnswood & Ireland Wood	15	61	-46
Cross Gates East & Manston	55	101	-46

Table 15: MSOAs with largest over representation of flows by huq (LQ)

MSOA	Rank (census)	Rank (huq)	Rank difference
Gipton South & Killingbeck Park	101	22	79
Kippax West	88	16	72
Gipton North	83	12	71
Farnley West & Gamble Hill	91	34	57
Garforth East	74	17	57

The most significant under representations by the MPD were generally attributable to MSOAs in north west Leeds, and over representations were more common in eastern MSOAs with the exclusion of *Farnley West and Gamble Hill*.

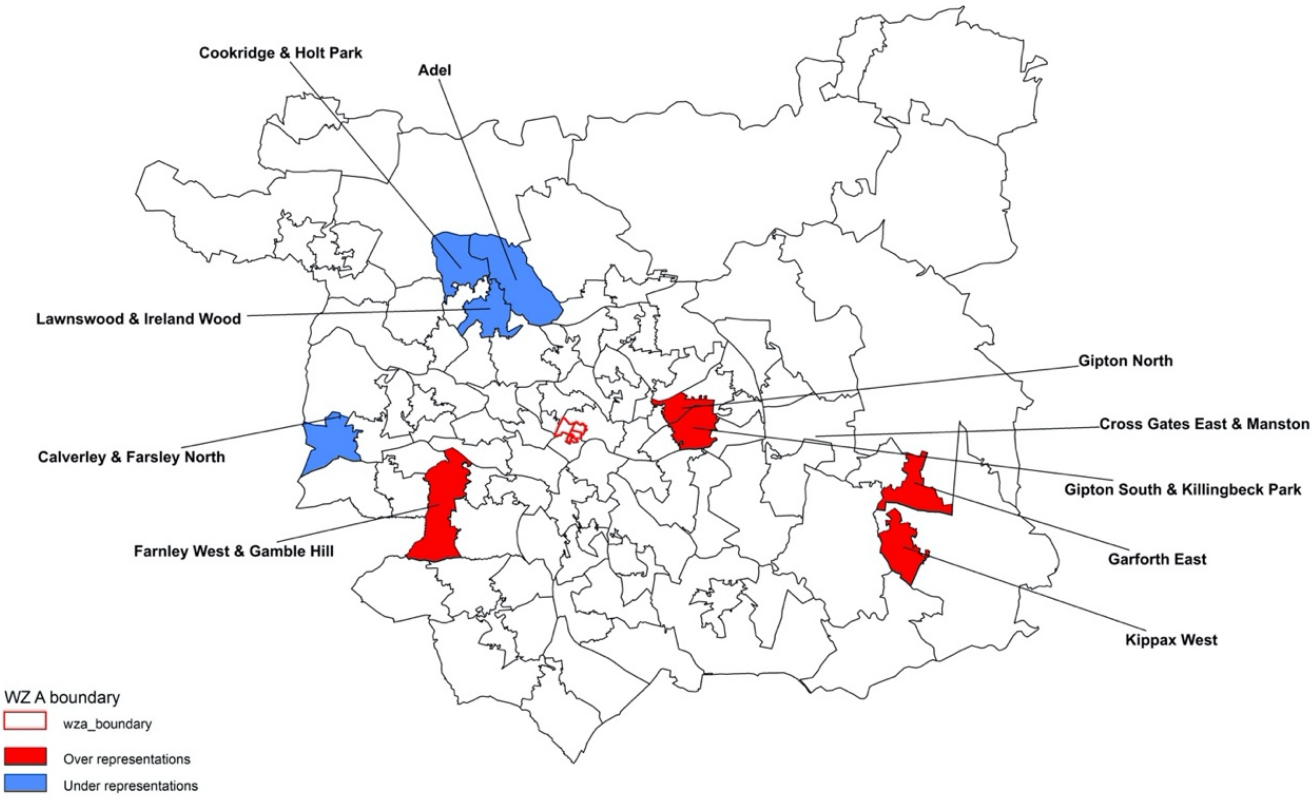

Figure 15: Distribution of MSOAs with largest under/over representation of flows by huq (WZ A)

5.2b. WZ B

For WZ B, the census identified that 0.54% of the total active workers in Leeds travelled to WZ B to work, whereas the MPD identified 12.2% of total active workers as travelling to WZ B for work.

Census LQs show general over concentration of workers in MSOAs north of the city centre (*Figure 16*), whereas MPD revealed the most densely concentrated MSOAs south of the city centre (*Figure 17*); the opposite pattern to the TTW data.

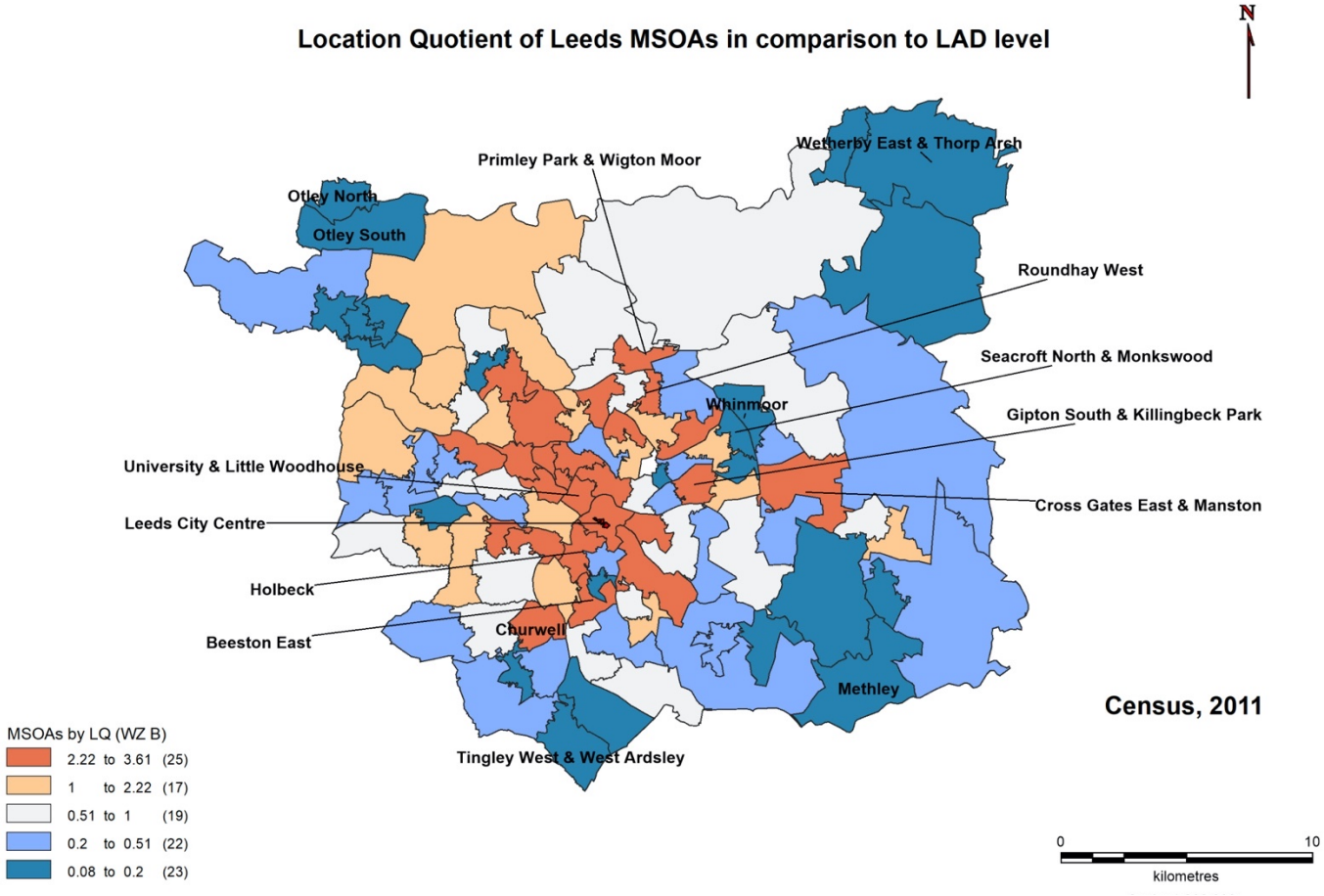

Figure 16: LQ of MSOA/LA level (census), WZ B

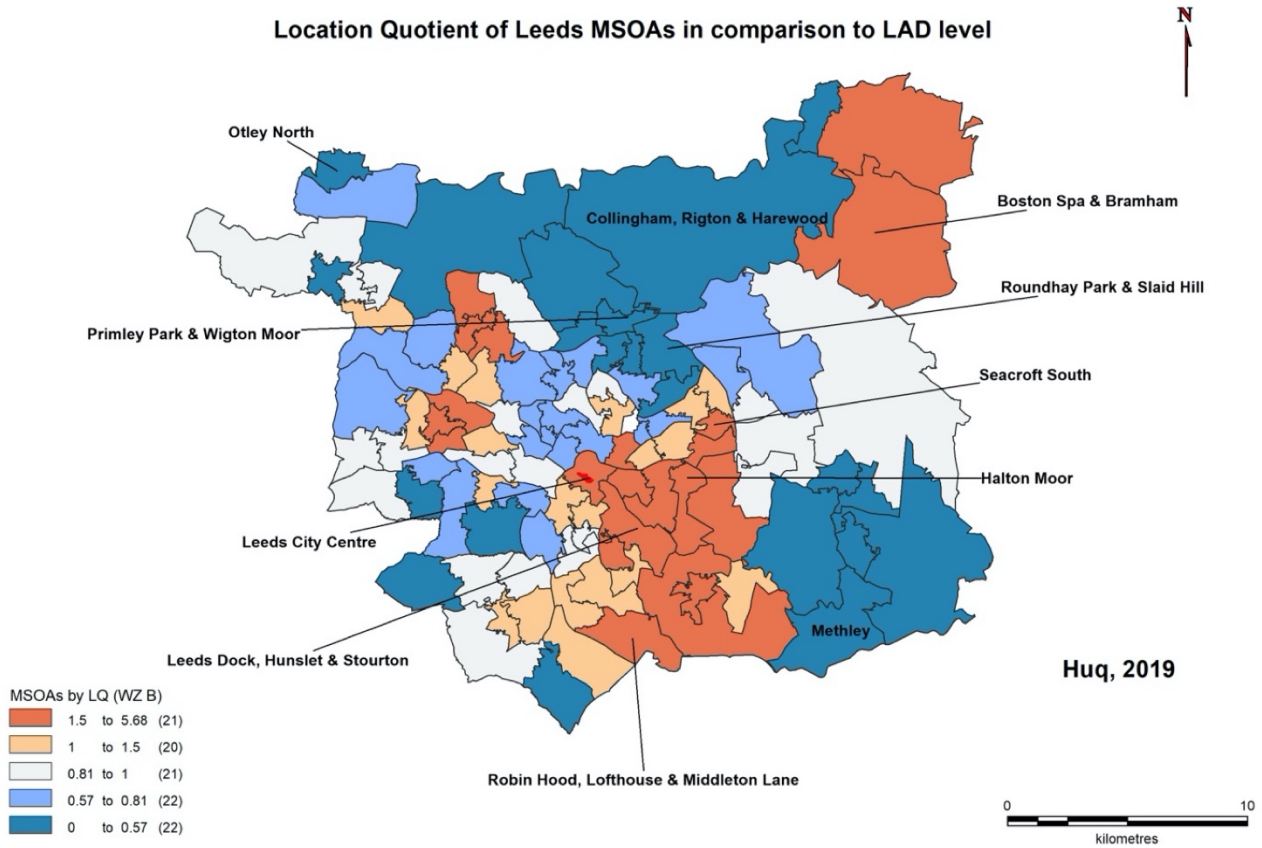

Figure 17: LQ of MSOA/LA level (huq), WZ B

Table 16: MSOAs with largest under representation of flows by huq

MSOA	Rank (census)	Rank (huq)	Rank difference
Primley Park & Wigton Moor	5	102	-97
University & Little Woodhouse	3	85	-82
Roundhay West	7	86	-79
Lady Wood & Oakwood	15	93	-78
Headingley	9	84	-75

Table 17: MSOAs with largest over representation of flows by huq

MSOA	Rank (census)	Rank (huq)	Rank difference
Wetherby East & Thorp Arch	104	16	88
Boston Spa & Bramham	89	2	87
Tinshill	95	9	86
Seacroft South	85	4	81
Rothwell Outer	80	6	74

The analysis found the MPD was over representing flows to WZ B in rural MSOAs (e.g. *Wetherby East and Thorp Arch*; *Boston Spa and Bramham*), and under representations were generally attributable to inner city MSOAs. This shows how density of areas and differentiations between urban and rural implicate the results.

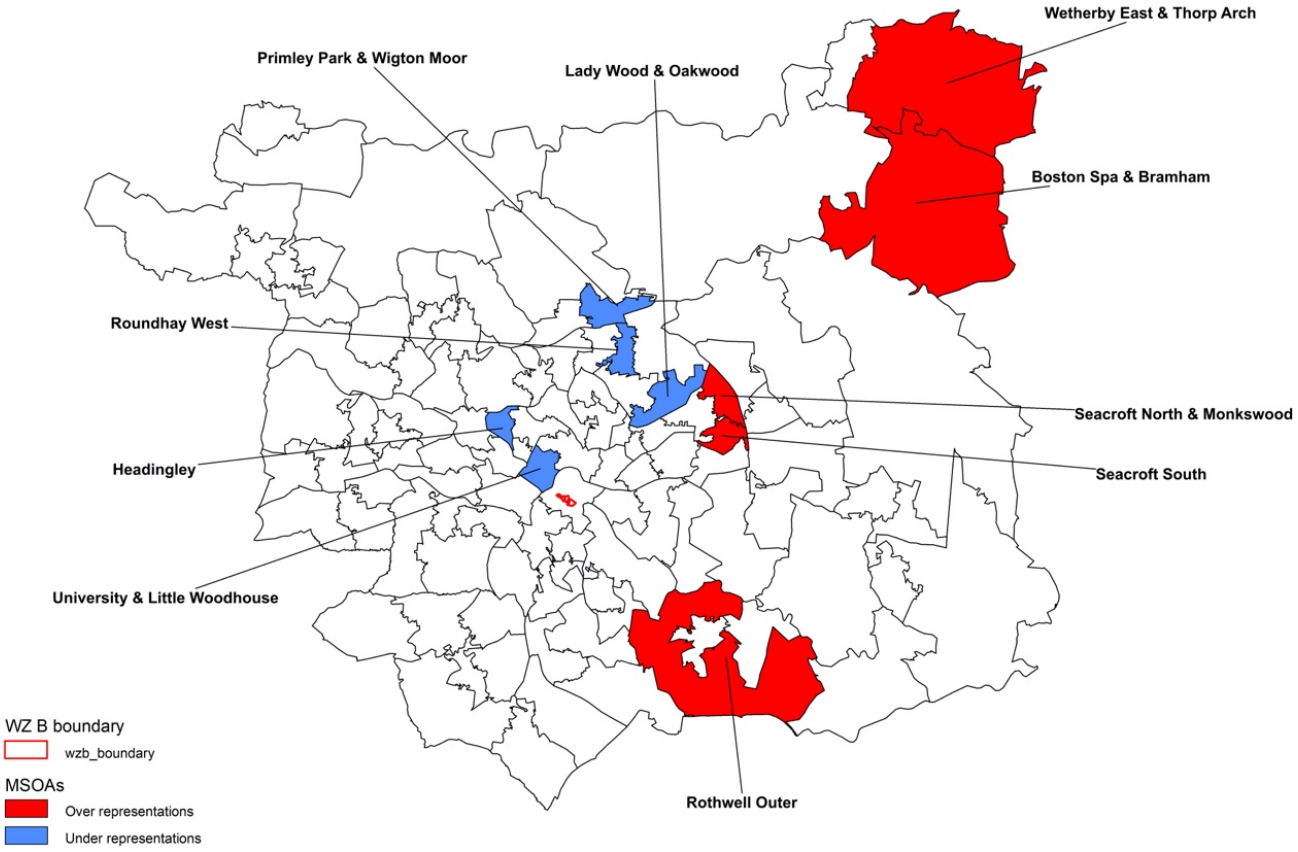

Figure 18: Distribution of MSOAs with largest under/over representation of flows by huq (WZ B)

LQs have outlined areas in which MPD frequencies are over or underrepresenting realistic commute counts derived from the census. Reasons for these differences will now be postulated using socio-demographic factors.

5.3. Socio-Demographic Factors

The most significant positive correlations for WZ A existed between Social Grade AB, levels of higher occupations and areas not deprived by any dimension.

For WZ B, the most significant positive correlations existed between larger proportions of workers in upper industries, prominence of 0-9-year olds and prominence of social grade DE. Thus, as the rank difference in LQs between the MPD and the census variables increases, as does the prominence of these socio-economic categories. These variables henceforth may contribute to the misrepresentation of flows by the MPD.

The three strongest positive and negative relationships between the variables are offered below. Other variables tested which exhibited little/no correlation are evident within *Appendix 3.*

Table 18: WZ A correlation outputs

Socio-economic variable	Correl Coefficient	% Strength of relationship (1 d.p.)
% Social Grade AB	0.427165815	42.7%
% Top 3 occupations	0.392660677	39.3%
% Not deprived in any dimension	0.392207287	39.2%
% Deprived in some dimension	-0.392207287	-39.2%
% Aged 0-9	-0.414520481	-41.5%
% Social Grade DE	-0.467804466	-46.8%

Table 19: WZ B correlation outputs

Socio-economic variable	Correl Coefficient	% Strength of relationship (1 d.p.)
% Industry A	0.451489916	45.1%
% Aged 0-9	0.377440386	37.7%
% Social Grade DE	0.347564909	34.8%
% Social Grade AB	-0.344838421	-34.5%
% Top 3 occupations	-0.375418006	-37.5%
% Full time students/pupils	-0.394929722	-39.5%

WZ A has more profound over representations of commuters from MSOAs with large proportions of social grade AB residents, and less over representations of commuters from largely social grade DE areas. This may relate to residents of higher social grade receiving larger disposable incomes, thus are more likely to own (and frequently use) a mobile device.

WZ B has the largest over representations of commuters from MSOAs with large proportions of industry A. As the amount of people working in *Industry A* rises, as does the MPDs over representation. Workers in the construction, transport and communication sector henceforth are largely accounted for.

Students have a larger impact within WZ B, as proportions of full-time students decrease, as does the scale of overrepresentation. This shows that students may be a contributing factor to large overrepresentations of commuters travelling to WZ B due to misclassifications of students as workers when travelling into Leeds City Centre.

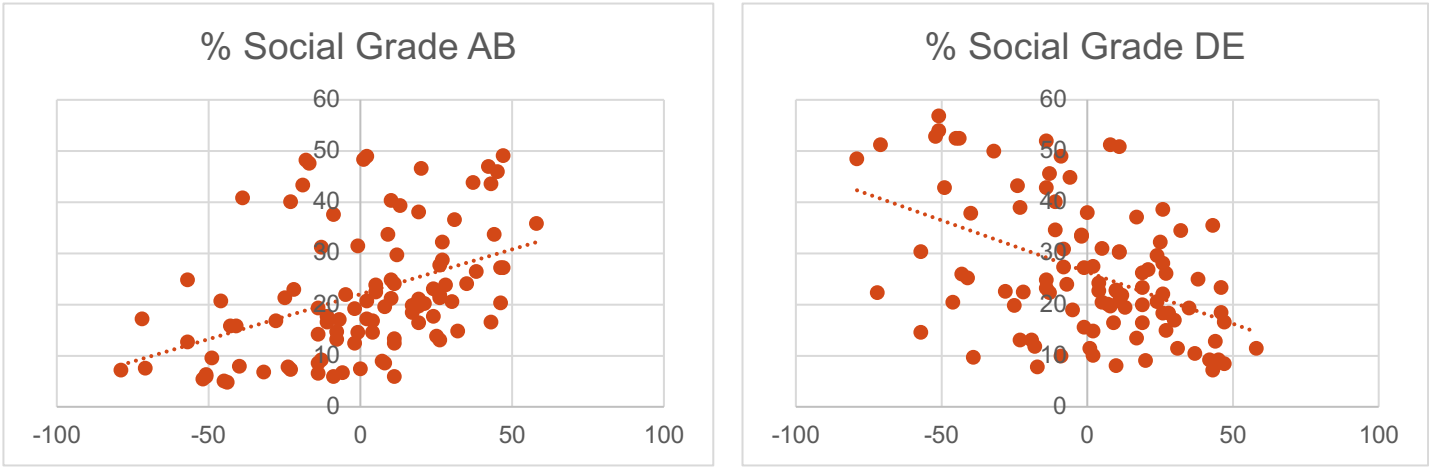

Figure 19: WZ A scatter plots of most significant correlations

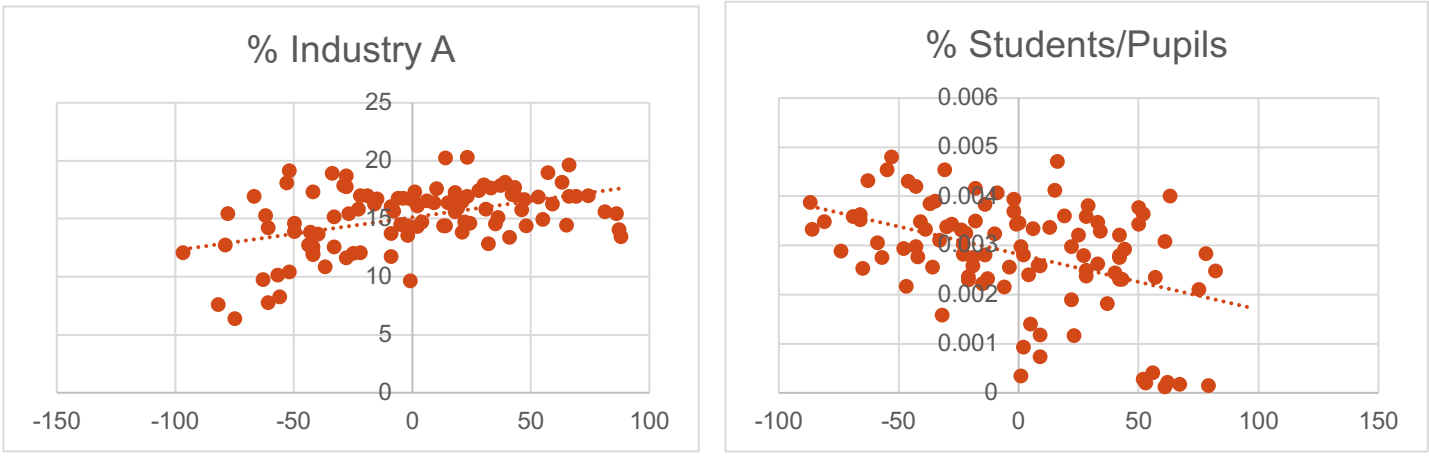

Figure 20: WZ B scatter plots of most significant correlations

Other factors, such as proportions of residents who generally work from home and method of travel to work were also assessed but no moderate correlations were found. These results are available within *Appendix 3.*

CHAPTER 6: DISCUSSION

6.1 Discussion

Location Quotient outputs revealed large proportions of intra-MSOA flows within the MPD due to the largest proportions of workers for each work zone deriving from the reciprocal MSOA the workzone resides in. This was a shared finding of the ONS (2017), who found that intra-LA flows were more significant within the MPD due to the misclassification of student journeys as commutes.

Following investigation into this socio-economic variable of student proportions in MSOAs, it was similarly found within WZ B that a lower percentage of students in the usually resident population equated to wider underrepresentations of commute counts by the MPD. This demonstrates the proportions of student mobile phone usage and their daily movements within the city centre causing misclassifications of commute journeys, when journeys are more likely to have a leisure orientated purpose. However, the strength of this correlation was only attributable to WZ B (Industry Zone), demonstrating how huq's methodology may be more rigorous in distinguishing between students and commuters in renowned student area (as WZ A refers to the University Zone).

Within the University Zone, correlations were more significant with prominence of Social Grade categories, as large proportions of Social Grade AB caused larger over representations of flows by huq. This relates to a key bias within MPD studies, as analysis is strictly limited to smartphone users. Statista (2017) found the average price of a smartphone was around four hundred and sixty pounds; a price likely to exceed budgets of residents within lower social grade categories. This bias in the predominant demographic that many MPD studies analyse relates to the finding of Crampton et al (2013), that making broad statements about society using only user-generated data is limiting as this variant of data skews towards the demographic of wealthier, educated, white males - as these people are more likely to own and frequently use smartphones.

The ONS (2017) also found MPD to have a greater coverage of commutes travelling a further distance. This study similarly identified more holistic coverage of commutes to WZ B, in which commuters on average travelled further to work than commuters to WZ A. Coverage attained within this study also appeared more succinct than that of CDR based studies (Deville et al, 2014; Lai et al, 2019) as under representations were not found in rural areas where cell towers are less prolific. Rural areas were actually overrepresented in some instances, such as *Wetherby East and Thorp Arch* and *Boston Spa and Bramham* within the WZ B analysis; illuminating the benefits of GPS data compared to CDR based studies.

The ONS (2017) also identified MSOAs containing new housing developments since 2011 to analyse whether MPD could provide more real-time updates on changing commuter flows, whereas this study

identified workplace zones in which new offices and management consultancies have arised since the 2011 census (which collated to form WZ B). This appears a promising aspect of methodology for future studies, as representation of commuters was generally more thorough for WZ B and may have represented large proportions of commuters not acknowledged by the census due to the opening of offices in the intercensal period (e.g. new commuters from York due to the 877% increase from the 2011 statistics).

6.2. Implications if MPD is to be used as a Census Replacement

As Wilson (1969) outlined the importance of questioning the validity of data and models in quantitative geography, the findings of this study will be used to address implications if MPD is to be used as a census replacement.

Classification systems need to be addressed as the methodology used by huq in defining origins (home location) and destinations (work location) should be noted. Huq's MPD may have reversed classifications of home and work due to huq's anchor point algorithm in identifying a device's home residence as the location in which it spends the most hours overnight, and the work location as the location in which the device spends the most time during the day. This methodology would incorrectly categorise nightshift labourer's work location as home, inverting commuting routes as their daytime location would be regarded as work. The scale of this limitation upon studies is important, as areas identified as having the most concentrated residential populations of commuters to different zones therefore may actually reflect zones inhabiting the end location of the commute (i.e. the destination rather than the origin). At present, the proportions of workers in which this reversed classification is attributable to cannot be accurately quantified due to the reliance on anchor point algorithms in defining home and work locations, as applied by huq, Novak et al (2012) and Alexander et al (2016). Classification methodologies were also found to be an issue by Lai et al (2019) in their study of inferring national migration statistics, as some residents were misclassified as migrants due to their temporary relocation following flooding in 2010.

Although classifications are unlikely to be infallible due to the inference involved within MPD studies, future studies may construct a more rigorous classification system by analysing the likely purpose of journeys based on the devices activities at the supposed destination (i.e. screen time spent on the device and movement to inform whether the user is likely to be asleep or at work). This would ensure greater accuracy in commuting patterns of those with less uniform working patterns. This reconfiguration of classification methodologies relates to the finding of the ONS (2017) who asserted that future research using MPD may be enhanced by studies not limited to analysing workers of standard working

hours, in attempt to include those with non-standard working patterns (e.g. nightshift, weekend, zero-hour contract workers).

6.3. Socio-Economic Limitations

The nature of statements outlining potential correlations between socio-economic factors and areas of over or under representation by huq (offered within the *Analysis*) should also be addressed. The finding that MSOAs inhabited by larger proportions of workers with sporadic workplaces generally coincide with MSOAs over-representing commutes by huq should be questioned due to Gould's (1981, p. 166) advise not to allow large datasets to "speak for themselves". Although useful in exploring connections between socio-economic factors and areas of profound difference in results to the census, the data cannot definitively represent society as this would commit an ecological fallacy by assigning the same characteristic to everyone in the area, but also broad statements about society cannot be offered due to the methodological limitations of big-datasets as outlined by Crampton et al (2013).

These broad assumptions are not universally applicable due to the inherent differences between study areas, residents and working practices. Researchers must consider the validity of their results, and how reliability could be improved by conducting reciprocal studies in different geographic locations. Only then could broader assumptions about society be offered with an enhanced degree of accuracy as results become applicable to various study areas, opposed to just one.

Additionally, socio-economic data used was collated from the 2011 census due to lack of information collected on an annual basis. Although age categories were adjusted to acknowledge the likely ages of residents now inhabiting different areas (by increasing ages by eight years to reflect the current duration of the intercensal period), reciprocal adjustments could not be made to other datasets such as MSOAs largely inhabited by workers with sporadic/uniform working practices, as no dataset provides up to date information at this spatial scale. This analysis therefore relies on the assumption that residents have remained at the same address since the census conduction (27th March 2011) – which realistically is unlikely.

6.4. Future Research

Due to the general replication of patterns by the MPD of the census results, MPD proves a promising data source for the analysis of commuting behaviours. Future research may separate students entirely from MPD algorithms by extending analysis periods to examine the differences between term-time and non-term time patterns. By disaggregating data in this way, researchers would gain a better

understanding as to whether results are exclusive or inclusive of students who are easily misclassified as commuters.

As mentioned previously, reciprocal studies are important to ensure findings are applicable to various areas and are not unique to one destination (i.e. West Yorkshire). Due to the size of WZs (as demonstrated within *Chapter 4*), this will be a time-consuming process due to amount of WZs in the UK alone. The aggregate scale of WZ is also not a universal scale, so international comparisons will be difficult to be identified as no universal aggregate scale exists for workplace data. To broaden findings and enable international comparisons, a universal aggregate scale may be necessary to compare differences between countries. A more conclusive base of research would catalyse diverse results produced by homogenous methods, enhancing the degree of validity existent within findings.

The 2021 census may also reveal changes in workplace practices caused by COVID-19, as only five percent of the UK active workforce worked from home prior to the coronavirus outbreak (ONS, 2019). Following the result of UK lockdown measures, all but key workers were constrained to their home; illuminating an aspect of future research as analysis may focus on the movement of key workers and their commuting behaviours during this period.

CHAPTER 7: CONCLUSION

7.1. Conclusion

This study has outlined data, methodology and cartographic outputs which support the promising use of mobile phone locational data in future research aiding urban planning practices. The presented methods of analyses and visualisation of commuting within UK cities are appropriate tools for reciprocal future research in other areas. The final census occurring next year in 2021 will provide a rich dataset (in both coverage and quality) that can be used to assess the representativeness of MPD offered by private sectors companies such as huq referring to the same year. Not only will this allow a more direct comparison as time frames will align, up to date socio-economic data will similarly be provided by the census which can be used to distinguish correlations between different demographic groups. This will henceforth provide a more definitive characterisation of the value of MPD based on its representativeness to realistic commutes in a given year.

The conclusions of this study have the potential to be utilised by statisticians alongside census data, as MPD provides a more real-time update on potentially outdated statistics. Findings have added to the very limited numbers of UK based studies, providing an in-depth investigation of the representativeness of MPD in analysing commuting patterns to two zones in Leeds.

7.2. Closing Statements

- Patterns were broadly similar between the 2011 census variables and 2019 MPD.
- Students cause over representations in commute flows within the MPD due to potential mis-categorisations of leisure-oriented journeys to WZ B.
- The socio-economic variable social grade had the most significant correlation with the scale of representativeness for WZ A.
- In this study, differing methods of travel to work had very weak correlations to the scale of representation by the MPD, suggesting transport mode does not cause misrepresentations of journeys by huq.
- Ensuring classification algorithms are more succinct in future studies will dramatically increase the validity of forthcoming results.

REFERENCE LIST:

Alexander, L., Jiang, S., Murga, M., Gonzalez, M. 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C.* **58**(B), pp. 240-250.

Bank my cell. 2020. *How many smartphones are in the world?* [Online]. [Accessed 20th March 2020]. Available from: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

Berry, T., Newing, A., Davies, D., Branch, K. 2016. Using workplace population statistics to understand retail store performance. *The International Review of Retail, Distribution, and Consumer Research.* **26**(4), pp. 375-395.

Blazon. 2019. *How much data do we create every day? The mind-blowing stats everyone should read.* [Online]. [Accessed 20th March 2020]. Available from: https://blazon.online/data-marketing/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Blazquez, D., Domenech, J. 2017. Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change.* **130**, pp. 99-113.

Boyle, M. 2020. *Mobile Internet Statistics.* [Online]. [Accessed 1st April 2020]. Available from: https://www.finder.com/uk/mobile-internet-statistics

Chen, Y. 2017. Transportation and quantitative analysis of socio-economic development of relations. *IOP Conference Series Earth and Environmental Science*. **100**(1).

Congdon, P. 2012. Assessing the Impact of Socioeconomic Variables on Small Area Variations in Suicide Outcomes in England. *International Journal of Environmental Research and Public Health.* **10**(1), pp. 158-177.

Corcoran, J., Higgs, G., Brunsdon, C., Ware, A., Norman, P. 2007. The use of spatial analytical techniques to explore patterns of fire incidence: A South Wales case study. *Computers, Environment and Urban Systems.* **31**(1), pp. 623-647.

Cox, M., Ellsworth, D. 1997. Managing Big Data for scientific visualization. *CM Siggraph, MRJ/NASA Ames Res. Cent*. **5**, pp. 1-17

Crampton, J.W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. Zook. 2013. "Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb." *Cartography and Geographic Information Science*. **40**(2), pp. 130–139.

Demissie, M., Antunes, F., Bento, C. Phithakkitunkoon, S., Sukhvibul, T. 2016. Inferring Origin-Destination Flows Using Mobile Phone Data: A Case Study of Senegal. *Conference Paper,* May 2016.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J. 2014. Dynamic population mapping using mobile phone data. *Proc Natl Acad Sci USA.* **111**(45), pp. 15888-15893.

Gould, P. 1981. "Letting the Data Speak for Themselves." *Annals of the Association of American Geographers*. **71**(2), pp. 166–176.

Graham, M. 2012. Big Data and the End of theory? *The Guardian*. [Online]. March 9, 2012. [Accessed 1st April 2020]. Available from: http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory

Haklay, M. 2012. "'Nobody Wants to Do Council Estates': Digital Divide, Spatial Justice and Outliers". Paper presented at the 108th Annual Meeting of the Association of American Geographers, New York, February 25, 2012.

Henley, J. 2011. Do we actually need a census? *The Guardian.* [Online]. 10th March 2011. [Accessed 24th January 2020]. Available from: https://www.theguardian.com/uk/2011/mar/10/census-2011-do-we-need-it

Huq. 2019. Geo-Behaviour Data FAQs. *FAQs for the Geo-Behavioural dataset.* Supplementary document for purchasers.

Infuse. 2011. *2011 Population Statistics.* [Online]. [Accessed 29th April 2020]. Available from: http://infusecp.mimas.ac.uk/

Lai, S., Erbach-Schoenberg, E.Z., Pezzulo, C., Ruktanonchai, N.W., Sorichetta, A., Steele, J., Li, T., Dooley, C.A., Tatem, A.J. 2019. Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications.* **5**(34), pp. 1-10.

Martin, D., Gale, C., Cockings, S., Harfoot, A. 2017. Origin-destination geodemographics for analysis of travel to work flows. *Computers, Environment and Urban Systems.* **67,** *pp 68-79.*

McMillan, I. 2019. The tale of my strange and stressful Thursday rail commute. *I News.* [Online]. 6th September 2019. [Accessed 6th May 2020]. Available from: https://inews.co.uk/opinion/commuting-rail-commuting-sheffield-manchesters-trains-awful-499061

Novak, J., Ahas, R., Aasa, A., Silm, S. 2013. Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia. *Journal of Maps.* **9**(1), pp. 10-15.

ONS. 2011. *2011 Census Glossary of Terms*. [Online]. [Accessed 24th January 2020] Available from: https://www.ons.gov.uk/file?uri=/census/2011census/2011censusdata/2011censususerguide/glossary/glossaryv1025july2017.pdf

ONS. 2014. *The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority.* [Online]. Newport: ONS. [Accessed 29th March 2020]. Available from: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation--and-recommendations/national-statisticians-recommendation.pdf

ONS. 2017. *Research Outputs: Using mobile phone data to estimate commuting flows*. [Online]. [Accessed 23rd January 2020]. Available from: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/populationcharacteristics/researchoutputsusingmobilephonedatatoestimatecommutingflows

ONS. 2019. *Coronavirus and homeworking in the UK labour market: 2019.* [Online]. [Accessed 2nd April 2020]. Available from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/coronavirusandhomeworkingintheuklabourmarket/2019

Polzin, S. and A. Pisarski. 2015. "Commuting in America 2013: The National Report on Commuting Patterns and Trends." *American Association of State Highway and Transportation Officials*. Washington, DC.

Rumsey, D. 2016. *Statistics For Dummies.* 2nd Ed. John Wiley & Sons Inc: New York.

Sadeghinsar, B., Akhavan, A., Wang, Q. 2018. Estimating Commuting Patterns from High Resolution Phone GPS Data. *Department of Civil and Environmental Engineering: Northeastern University.*

Shaw, D. 2020. UK's 2021 census could be the last, statistics chief reveals. *BBC News.* [Online]. 12th February 2020. [Accessed 30th March 2020]. Available from: https://www.bbc.co.uk/news/uk-51468919

Statista. 2017. *Smartphone average price forecast.* [Online]. [Accessed 1st May 2020]. Available from: https://www.statista.com/statistics/283334/average-smartphone-price

Statista. 2020. *UK households: ownership of mobile telephones 1996-2018*. [Online]. [Accessed 4th May 2020]. Available from: https://www.statista.com/statistics/289167/mobile-phone-penetration-in-the-uk/

UK Data Service. 2011. *Census Support: Flow Data, Flexible Query Builder.* [Online]. [Accessed 2nd January 2020]. Available from: https://wicid.ukdataservice.ac.uk/

UK Data Service. [No date]. *Census Forms.* [Online]. [Accessed 1st March 2020]. Available from: https://census.ukdataservice.ac.uk/use-data/censuses/forms.aspx

UK Geographics. 2014. *Social Grade A, B, C1, C2, D, E*. [Online]. [Accessed 29th April 2020]. Available from: https://www.ukgeographics.co.uk/blog/social-grade-a-b-c1-c2-d-e

Wilson, A. G. 1969. "The Use of Analogies in Geography." *Geographical Analysis*. **1**(3), pp. 225–33.

APPENDIX

Appendix 1 – *Census Questionnaire extract relating to address of main workplace.*

Person 1 - continued

32 Answer the remaining questions for your main job or, if not working, your last main job.

- Your main job is the job in which you usually work (worked) the most hours

33 In your main job, are (were) you:

- ☐ an employee?
- ☐ self-employed or freelance without employees?
- ☐ self-employed with employees?

34 What is (was) your full and specific job title?

- For example, PRIMARY SCHOOL TEACHER, CAR MECHANIC, DISTRICT NURSE, STRUCTURAL ENGINEER
- Do not state your grade or pay band

35 Briefly describe what you do (did) in your main job.

36 Do (did) you supervise any employees?

- Supervision involves overseeing the work of other employees on a day-to-day basis

☐ Yes ☐ No

37 At your workplace, what is (was) the main activity of your employer or business?

- For example, PRIMARY EDUCATION, REPAIRING CARS, CONTRACT CATERING, COMPUTER SERVICING
- If you are (were) a civil servant, write GOVERNMENT
- If you are (were) a local government officer, write LOCAL GOVERNMENT and give the name of your department within the local authority

38 In your main job, what is (was) the name of the organisation you work (worked) for?

- If you are (were) self-employed in your own organisation, write in the business name

☐ No organisation, for example, self-employed

39 If you had a job last week ➔ Go to **40**

If you didn't have a job last week ➔ Go to **43**

40 In your main job, what is the address of your workplace?

- If you work at or from home, on an offshore installation, or have no fixed workplace, tick one of the boxes below
- If you report to a depot, write in the depot address

Postcode

OR
- ☐ Mainly work at or from home
- ☐ Offshore installation
- ☐ No fixed place

41 How do you usually travel to work?

- Tick one box only
- Tick the box for the longest part, by distance, of your usual journey to work

- ☐ Work mainly at or from home
- ☐ Underground, metro, light rail, tram
- ☐ Train
- ☐ Bus, minibus or coach
- ☐ Taxi
- ☐ Motorcycle, scooter or moped
- ☐ Driving a car or van
- ☐ Passenger in a car or van
- ☐ Bicycle
- ☐ On foot
- ☐ Other

42 In your main job, how many hours a week (including paid and unpaid overtime) do you usually work?

- ☐ 15 or less
- ☐ 16 - 30
- ☐ 31 - 48
- ☐ 49 or more

43 There are no more questions for Person 1.

➔ Go to questions for Person 2

OR If there are no more people in this household,
➔ Go to the Visitor questions on the back page

OR If there are no visitors staying here overnight,

Appendix 2:

Four different dimensions of household deprivation.

Dimension	Description
Employment	Where any member of the household, who is not a full-time student is either unemployed or long-term sick.
Education	No person in the household has at least Level 2 education, and no person aged 16-18 is a full-time student.
Health and disability	Any person in the household has general health that is 'bad' or 'very bad' or has a long-term health problem.
Housing	The household's accommodation is either overcrowded, with an occupancy rating -1 or less, or is in a shared dwelling, or has no central heating.

Social Grade categories.

Social Grade	Description
AB	Higher & intermediate managerial, administrative, professional occupations
C1	Supervisory, clerical & junior managerial, administrative, professional occupations
C2	Skilled manual occupations
DE	Semi-skilled & unskilled manual occupations, Unemployed and lowest grade occupations

Appendix 3 – *All correlation outputs of socio-demographic analysis.*

Socio-economic variable	WZ A	Moderate Correlation (Y/N)	WZ B	Moderate correlation (Y/N)
% Aged 0-9	-0.41452	Y	0.377440386	Y
% Aged 16-19	0.036548	N	-0.272892574	N
% Aged over 65	0.071284	N	0.121676	N
% Bike	-0.27496	N	-0.268126416	N
% Bus/Coach	-0.27496	N	0.060920561	N
% Car/Van	0.200246	N	0.056556085	N
% Deprived in some dimension	-0.39221	Y	0.338112289	Y
% Industry A	-0.19857	N	0.451489916	Y
% Industry B	0.31291	Y	-0.34414425	Y
% Not deprived in any dimension	0.392207	Y	-0.338112289	Y
% Full time students/pupils	0.114674	N	-0.394929722	Y
% of workforce WFH	0.302658	Y	-0.19258512	N
% of workforce full-time	0.184703	N	0.242494589	N
% of workforce part-time	-0.1847	N	-0.242494589	N
% On foot	0.091661	N	-0.049906091	N
% Social Grade AB	0.427166	Y	-0.344838421	Y
% Social Grade DE	-0.4678	Y	0.347564909	Y
% Top 3 occupations	0.392661	Y	-0.375418006	Y
% Train	0.180373	N	-0.052987558	N
Average TTW distance (km)	0.170257	N	-0.239861039	N