

Tessellating the Space-Time Prism: Regionalisation of In-app Location Data for Privacy Protection and Data Preservation

Louise. S. Sieg

Thesis submitted in conformity with the requirements of
Doctor of Philosophy (Ph.D.)

**Department of Geography
University College London**

September 2024

Declaration

I, Louise Sieg, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Signed

Thesis Outputs

Peer Reviewed Journal Publications

2024 The regionalisation and aggregation of in-app location data to maximise information and minimise data disclosure, *Geographical Analysis*, , L. Sieg, J. Cheshire.
DOI: 10.1111/GEAN.12406

2024 Ethical challenges arising from the mapping of mobile phone location data, *The Cartographic Journal*, L. Sieg, H. Gibbs, M. Gibin, J. Cheshire
DOI - 10.1080/00087041.2024.2318055

Peer Reviewed Conference Proceedings

2023 Regionalising mobile phone data: attributing time components to optimised regions to create region classifications, *GISRUK2023*, Glasgow, UK. L. Sieg, J. Cheshire.

2022 Representing big mobile data using automated grid systems – an assessment of the quadtree method, *GISRUK2022*, Liverpool, UK. L. Sieg, J. Cheshire.

2021 Spatially aggregating mobility data – implications and challenges, *GISRUK2021*, Cardiff, UK. L. Sieg, J. Cheshire.

Other Conference Presentations

2023 A new methodology for the regionalisation and aggregation of in-app location data, *Association of American Geographers Annual Meeting 2023*, Denver, Colorado, USA. L. Sieg, J. Cheshire.

2022 Regionalisation of mobile phone datasets – A new method to conserve both privacy and granularity, *Royal Geographical Society-IBG Annual International Conference 2022*, Newcastle, UK. L. Sieg, J. Cheshire.

Acknowledgments

First of all, I would like to thank my primary supervisor, James Cheshire, for his central role in overseeing this work and for providing continued guidance and support. I am also very grateful to my secondary supervisor, Paul Longley, for his mastery and for contributing key ideas throughout. This thesis would not have happened without either of them. I would also like to thank the ESRC for providing the funding for this research.

Secondly, I would like to thank my friends and family, particularly my parents, for their enthusiasm and kindness, and for their unwavering support through the harder stretches of this thesis. Merci.

I would finally like to extend my deep gratitude to my colleagues and friends at the CDRC. I am particularly indebted to Hamish Gibbs for his remarkable patience and altruism. Thank you all for a memorable 3 years.

Abstract

Datasets containing locational information collected by mobile phones are increasingly used in social science research and mobility analysis, yet they continue to present a range of technical, financial, and ethical challenges. The risk of disclosure of personally identifiable information is a foundational concern in this context, and so to alleviate it, datasets are often aggregated to pre-defined geographic units and presented as counts of the number of mobile devices within them at a given time. The use of grids or units created by statistical agencies for the dissemination of traditional datasets - such as censuses - are common choices for this aggregation process. However, these can result in large variations in the number of devices encapsulated within each geographic unit, resulting in over-generalisation and a loss of information in some areas.

Investigating this issue is the core theme of this work, which will explore different regionalisation methodologies and their consequences before alighting on a novel method for generating spatial units tailored for mobile phone data. The central aim is to maximise the granularity of the data, whilst minimising the risks of disclosing personal information. This methodology has applications to widely available datasets and enables bespoke geographical units to be created for different contexts and timescales. The generated units are compared to established aggregates from the England and Wales Census and Ordnance Survey, and assessed through varying temporal granularities. This work seeks to demonstrate that these bespoke outputs minimise data omission caused by low counts and preserve underlying data distribution better than existing aggregation methods.

This thesis speaks to the need for data-driven and context-driven regionalisation methodologies in enabling the best use new forms of data in research. It endeavours to contribute to a better understanding and safer use of mobile phone location data in social science and promotes regionalisation as a promising solution to reconcile data granularity with disclosure for sensitive location datasets.

Impact Statement

The most notable output of this thesis is the regionalisation methodology proposed to reconcile data preservation and granularity. The development of these techniques is important as different data stakeholders and researchers seek to increasingly investigate new forms of data which may present ethical and legal risks. This research puts at the forefront of the discussion the need to produce bespoke aggregates which permit the valuable research conducted whilst protecting data users. This is especially timely as the COVID-19 pandemic saw a drastic increase in the uptake of these still relatively misunderstood datasets.

This thesis proposes that sensitive location data can be regionalised and aggregated to be used successfully across many more fields and applications than they currently are with safe dissemination methodologies. The methodological aspects of this work, as well as the core discussions surrounding data protection and granularity, have impacted a variety of audiences, through peer-reviewed journals, presentations, conversation with stakeholders and seminar teaching. Below, we outline some evidence of this impact.

Academic impact

Publications and presentations attest for this research's impact in the academic community of Geographic information Sciences (GIS). Three academic conference abstracts have been published, including as part of the annual GIS research UK and American Association of Geographers conferences, highlighting the international range of impacts. The material included in Chapter 5, which outlines the thesis' key output of the regionalisation methodology, was published in an internationally recognised peer-reviewed journal (Geographical Analysis). The ethical discussions woven throughout this work were also in part published in a peer reviewed publication, and they notably provided me with key expertise and knowledge to lead seminars for modules related to data policy and society at Master's level.

Impact outside academia

The research presented focuses on a consumer dataset provided by a data partner. This partnership enables applications of this research outside of academia, as the results obtained are also discussed in the context of commercial data dissemination. Discussions with partners at British Telecom also revealed an interest in the thesis' outputs for telecom services and

industry applications. As part of this research, I also engaged with various public policy events, such as an Office for National Statistics public conference on data acquirement for the UK Levelling Up agenda. Finally, the key methods developed in this thesis are all provided open sourced online for general use inside and outside of academic applications.

Research Paper Declaration Forms

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s) 1/2

1. 1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

(a) What is the title of the manuscript?

[The Regionalization and Aggregation of In-App Location Data to Maximize Information and Minimize Data Disclosure](#)

(b) Please include a link to or doi for the work: <https://doi.org/10.1111/gean.12406>

(c) Where was the work published? [The Ohio State University](#)

(d) Who published the work? [Geographical Analysis](#)
(<https://onlinelibrary.wiley.com/journal/15384632>)

(e) When was the work published? [07 June 2024](#)

(f) List the manuscript's authors in the order they appear on the publication: [Sieg Louise, Cheshire James.](#)

(g) Was the work peer reviewed? [Yes](#)

(h) Have you retained the copyright? [Yes](#)

(i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi

If 'No', please seek permission from the relevant publisher and check the box next to the below statement: [No](#)

☒ I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):

(a) What is the current title of the manuscript?

(b) Has the manuscript been uploaded to a preprint server 'e.g. medRxiv' ?

If 'Yes', please give a link or doi:

(c) Where is the work intended to be published?

(d) List the manuscript's authors in the intended authorship order:

(e) Stage of publication:

3. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):

[Sieg, L conceived of the presented idea and performed the analysis. Cheshire, J supervised the project and provided feedback and manuscript editing support.](#)

4. In which chapter(s) of your thesis can this material be found?

[Chapter 5: Operationalising H3: development of a bespoke regionalisation methodology](#)

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate: Date: [31/08/2024](#),

Supervisor/Senior Author signature (where appropriate):

Date: 03/09/2024

UCL Research Paper Declaration Form 2/2

1. 1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

- (a) What is the title of the manuscript?
- (b) Please include a link to or doi for the work:
- (c) Where was the work published?
- (d) Who published the work?
- (e) When was the work published?
- (f) List the manuscript's authors in the order they appear on the

publication:

- (g) Was the work peer reviewed?
- (h) Have you retained the copyright?
- (i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi
If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐ I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3):

- (a) What is the current title of the manuscript? [Ethical challenges arising from the mapping of mobile phone location data](#)
- (b) Has the manuscript been uploaded to a preprint server 'e.g. medRxiv' ? **No**

If 'Yes', please give a link or doi:

- (c) Where is the work intended to be published? [The Cartographic Journal](#)
- (d) List the manuscript's authors in the intended authorship order: [Sieg Louise, Gibbs Hamish, Gibin Maurizio, Cheshire James.](#)

(e) Stage of publication: [accepted and proofs submitted to editor. Awaiting publication.](#)

3. For multi-authored work, please give a statement of contribution

covering all authors (if single-author, please skip to section 4): [J.C, H.G and L.S conceived of the presented idea. H.G and L.S wrote the manuscript with support from M.G. All authors provided critical feedback and helped shape the final version of the manuscript.](#)

4. In which chapter(s) of your thesis can this material be found?

[Chapter 2: Literature review and Chapter 7: Discussion.](#)

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate: Date: [31/08/2024](#),

Supervisor/Senior Author signature (where appropriate):

Date: 03/09/2024

Table of Contents

DECLARATION	2
THESIS OUTPUTS.....	4
ACKNOWLEDGMENTS.....	6
ABSTRACT	8
IMPACT STATEMENT	10
RESEARCH PAPER DECLARATION FORMS.....	12
TABLE OF CONTENTS	14
LIST OF FIGURES.....	18
LIST OF TABLES.....	24
LIST OF ABBREVIATIONS.....	26
1. INTRODUCTION.....	27
1.1. AIMS	28
1.2. THESIS STRUCTURE	29
1.2.1. Chapter 2: Literature review	29
1.2.2. Chapter 3: Data Presentation and Preliminary Analysis	30
1.2.3. Chapter 4: Regionalisation - Explorations	30
1.2.4. Chapter 5: Operationalising H3 – Development of a bespoke regionalisation methodology..	31
1.2.5. Chapter 6: MTUP and times of low data – assessment of the H3-based regions	31
1.2.6. Chapter 7: Discussion.....	32
1.3. NOTES ON SOFTWARE AND CODE	32
1.4. ETHICS	32
2. LITERATURE REVIEW.....	33
2.1. CONCEPTUAL FRAMEWORK – TIME GEOGRAPHY AND MOBILITY RESEARCH	33
2.1.1. Key concepts and definitions.....	33
2.1.1.1. Mobility.....	34
2.1.1.2. Time Geography.....	34
2.1.2. The value of mobility analysis.....	35
2.1.3. Traditional methods in capturing mobility.	37
2.1.3.1. Census.....	37
2.1.3.2. Capturing population information between censuses	38
2.2. THE IMPACT OF NEW FORMS OF DATA ON MOBILITY ANALYSIS	40
2.2.1. New forms of data: provenance and definitions.....	40
2.2.1.1. New, 'Big' data:.....	40
2.2.1.2. Consumer data definition	41
2.2.1.3. Mobile phone data.....	41
2.2.2. How these new datasets reshape geographic research.....	42
2.2.2.1. Paradigm shift - new forms of knowledge.....	42
2.2.2.2. Mobile phone data use cases in social sciences	43
2.3. THE TECHNICAL AND ETHICAL CHALLENGES OF MOBILE PHONE DATA	45
2.3.1. Accessibility, transparency, and technical challenges.....	46
2.3.2. Concerns around privacy and ethics	47
2.3.2.1. Bias	47
2.3.2.2. Consent	48
2.3.2.3. Privacy, confidentiality, and disclosure	49
2.4. DATA AGGREGATION: REGIONALISATION TO RECONCILE DISCLOSURE CONTROL AND GRANULARITY.....	51
2.4.1. Aggregation and geomasking as disclosure control.....	51
2.4.1.1. Geomasking and modelling	52
2.4.1.2. Aggregation.....	53
2.4.2. Impact of aggregation on analytical validity and completeness	56
2.4.3. Regionalisation	59

2.4.3.1.	Defining regionalisation	60
2.4.3.1.1.	Regions: formal or functional?	60
2.4.3.1.2.	Regionalisation definition	60
2.4.3.2.	Hierarchical or rules-based methodologies	61
2.4.3.2.1.	Case-study of hierarchical approach: Quadtree algorithms	62
2.4.3.2.2.	Case-study of rules-based approach: Output Areas	63
2.4.3.3.	Regionalisation of in-app mobile phone data: what to consider	64
2.4.3.3.1.	Summary: issues with in-app data	64
2.4.3.3.2.	Opportunities in regionalisation	65
3.	DATA AND PRELIMINARY ANALYSIS	68
3.1.	DATA OVERVIEW.....	68
3.1.1.	<i>Metadata and data description</i>	<i>68</i>
3.1.2.	<i>Filtering out low accuracy impressions</i>	<i>71</i>
3.1.3.	<i>Temporal distribution</i>	<i>73</i>
3.1.4.	<i>App statistics</i>	<i>77</i>
3.2.	DATA DIGESTION PROCESSES.....	79
3.2.1.	<i>Secure access to sensitive data: UCL Data Safe Haven</i>	<i>79</i>
3.2.2.	<i>Cleaning</i>	<i>81</i>
3.2.3.	<i>Typical aggregation process</i>	<i>84</i>
3.2.3.1.	<i>Spatial join to chosen scale.</i>	<i>84</i>
3.2.3.2.	<i>Creating activity counts</i>	<i>85</i>
3.3.	MAKING OF A STANDARD PRACTICE CDRC AGGREGATE FOR FURTHER DATA DESCRIPTIONS	87
3.3.1.	<i>Method and metadata</i>	<i>88</i>
3.3.2.	<i>Data exploration using the subproduct: spatial coverage</i>	<i>89</i>
3.3.2.1.	<i>Spatial coverage and insufficient data events (IDEs)</i>	<i>89</i>
3.3.2.2.	<i>Impact of using an aggregated product for data coverage exploration</i>	<i>92</i>
3.4.	DATA REPRESENTATION, BIAS AND AGGREGATION IMPACT	93
3.4.1.	<i>Population estimates.....</i>	<i>94</i>
3.4.2.	<i>Population representation</i>	<i>96</i>
3.4.2.1.	<i>Methodology</i>	<i>96</i>
3.4.2.2.	<i>Results</i>	<i>97</i>
3.4.3.	<i>Comparison of population assessments differences between the raw dataset and the 1km² data subproduct.....</i>	<i>99</i>
3.4.3.1.	<i>Linking the aggregated subproduct to OAs</i>	<i>99</i>
3.4.3.2.	<i>Comparison results</i>	<i>100</i>
3.5.	CHAPTER SUMMARY	104
4.	REGIONALISATION – EXPLORATIONS.....	105
4.1.	ASSESSMENT OF MAUP IMPACTS ON IN-APP DATA AGGREGATION	106
4.1.1.	<i>Methodology</i>	<i>106</i>
4.1.1.1.	<i>Creation of multiple aggregates of varying scales and zones</i>	<i>106</i>
4.1.1.2.	<i>Linkage with WPZ geographies</i>	<i>108</i>
4.1.2.	<i>Results</i>	<i>111</i>
4.2.	DEFINING THE PRINCIPLES AND PRACTICES FOR ROBUST REGIONALISATION.....	114
4.2.1.	<i>Objectives.....</i>	<i>116</i>
4.2.2.	<i>Constraints.....</i>	<i>116</i>
4.2.3.	<i>Criteria.....</i>	<i>117</i>
4.2.4.	<i>Usability</i>	<i>117</i>
4.3.	EXPLORATIONS OF EXISTING STATISTICAL METHODS	118
4.3.1.	<i>Quadtree based regionalisation.</i>	<i>118</i>
4.3.1.1.	<i>Definition.....</i>	<i>118</i>
4.3.1.2.	<i>Methodology</i>	<i>119</i>
4.3.1.3.	<i>Results</i>	<i>122</i>
4.3.1.4.	<i>Discussion</i>	<i>124</i>
4.3.2.	<i>Voronoi.....</i>	<i>126</i>
4.3.2.1.	<i>Definition.....</i>	<i>127</i>
4.3.2.2.	<i>Methodology</i>	<i>128</i>
4.3.2.2.1.	<i>Clustering the dataset.....</i>	<i>129</i>
4.3.2.2.2.	<i>Obtaining cluster centroids</i>	<i>130</i>
4.3.2.2.3.	<i>Voronoi regions.....</i>	<i>130</i>
4.3.2.3.	<i>Results</i>	<i>131</i>
4.3.2.3.1.	<i>Sensitivity analysis</i>	<i>131</i>

4.3.2.3.2.	Comparison with OSGB 250m ² and OAs.....	133
4.3.2.4.	Discussion.....	134
4.4.	SUMMARY	136
5.	OPERATIONALISING H3: DEVELOPMENT OF A BESPOKE REGIONALISATION METHODOLOGY	137
5.1.	METHODOLOGY	137
5.1.1.	<i>Making of the atomic units</i>	<i>139</i>
5.1.1.1.	Assessing H3 resolutions.....	140
5.1.1.2.	Aggregating data to the atomic unit.....	144
5.1.1.3.	Assigning contextual information to the atomic units	145
5.1.1.3.1.	Assigning administrative boundary information	146
5.1.1.3.1.1.	LSOA and MSOA scale comparison	146
5.1.1.3.1.2.	Linkage and assignment	147
5.1.1.3.2.	Land use assignment	148
5.1.1.3.2.1.	Creating the land-use and terrain profile (topo_code)	148
5.1.1.3.2.2.	Linkage and assignment.....	150
5.1.2.	<i>Preferred neighbour assignment (Making of the PNMx)</i>	<i>150</i>
5.1.2.1.	Filtering IDE units and ranking their neighbours.....	151
5.1.2.2.	Graph theory: finding connected nodes	152
5.1.3.	<i>Merging.....</i>	<i>153</i>
5.2.	OUTPUTS	154
5.2.1.	<i>Region shapefile.....</i>	<i>154</i>
5.2.2.	<i>Functions.....</i>	<i>155</i>
5.3.	VALIDATION AND SENSITIVITY ANALYSIS	158
5.3.1.	<i>Validation: comparison with traditional aggregation methods</i>	<i>158</i>
5.3.2.	<i>Sensitivity analysis using the Jaccard index</i>	<i>162</i>
5.3.2.1.	Definition.....	162
5.3.2.2.	Applying Jaccard to assess volatility	164
5.3.2.2.1.	Changes in data	164
5.3.2.2.2.	Changes in thresholds.....	166
5.3.2.3.	Discussion.....	168
5.4.	ASSESSMENT AGAINST REGIONALISATION PRINCIPLES AND AIMS	169
5.4.1.	<i>Objectives.....</i>	<i>169</i>
5.4.2.	<i>Constraints.....</i>	<i>170</i>
5.4.3.	<i>Criteria.....</i>	<i>170</i>
5.4.4.	<i>Usability</i>	<i>171</i>
5.4.5.	<i>Summary</i>	<i>172</i>
6.	MTUP AND TIMES LOW DATA: ASSESSMENT OF THE H3-BASED REGIONS.....	175
6.1.	MTUP: DEFINITIONS AND IMPLICATIONS FOR THE H3BRs	176
6.1.1.	<i>Defining MTUP</i>	<i>176</i>
6.1.2.	<i>MTUP impacts on the H3BR's performance</i>	<i>178</i>
6.1.2.1.	H3BRs: data-driven, for a day	178
6.1.2.2.	Changes of temporal scale	179
6.1.2.3.	Changes of temporal zones.....	181
6.1.3.	<i>Implications</i>	<i>183</i>
6.2.	HOURLY-BASED H3BR.....	184
6.2.1.	<i>Identifying times of low data</i>	<i>184</i>
6.2.2.	<i>Performance of an Hour-based H3BR</i>	<i>186</i>
6.2.2.1.	Can H3BR regions be made at the hour scale?	186
6.2.2.2.	Comparing H3BR_D and H3BR_H.....	188
6.2.2.3.	Using H3BR_H to map daytime and nighttime activity.....	191
6.3.	TIMES OF LOW DATA AND THEMATIC REGIONALISATION	193
6.3.1.	<i>Controlled temporal input: making regions for specific hours or use cases</i>	<i>193</i>
6.3.2.	<i>Jaccard comparisons of resulting regions</i>	<i>195</i>
6.3.3.	<i>Sunday's 3am and 5pm omissions by regions.....</i>	<i>197</i>
6.4.	MTUP AND DYNAMIC RECONFIGURATIONS OF OTHERWISE FIXED SPACE.....	200
6.4.1.	<i>Methodology.....</i>	<i>201</i>
6.4.1.1.	Data presentation: AddressBase Premium POIs.....	201
6.4.1.2.	Linkage to H3BR_NT and H3BR_DT	203
6.4.2.	<i>Results.....</i>	<i>204</i>

6.4.3.	<i>Implications</i>	208
6.5.	SUMMARY AND DISCUSSION	209
7.	DISCUSSION	211
7.1.	INTRODUCTION	211
7.2.	REFLECTION ON METHODS	212
7.2.1.	<i>Data assessment methods</i>	213
7.2.2.	<i>Reflections on regionalisation methodologies</i>	214
7.2.3.	<i>Reflection on MTUP assessment</i>	218
7.3.	LIMITATIONS	219
7.3.1.	<i>Dataset limitations</i>	219
7.3.2.	<i>Secure environments and software access</i>	220
7.4.	APPLICATIONS AND IMPLICATIONS	221
7.4.1.	<i>Harnessing promising datasets</i>	221
7.4.2.	<i>Understanding in-app data</i>	222
7.4.3.	<i>The MAUP, formal frameworks for aggregation, and dissemination</i>	223
7.4.4.	<i>Applications of the MTUP assessment</i>	224
7.4.5.	<i>Data protection, privacy and ethical implications</i>	224
7.5.	FUTURE PROSPECTS AND CONCLUSION	226
	REFERENCES	228
	APPENDIX	247

List of Figures

CHAPTER 2

Figure 1. Flowchart of data manipulations for removing disclosive information from in-app datasets. The original location points are stored within a secure environment, and only the outputs are removed. For aggregation, counts below a threshold (typically, 10 individuals) are suppressed from the outputs.	52
Figure 2. Synthetic data example of the spatial aggregation process.	54
Figure 3. OSGB coordinate system – illustration of subdivisions (Ordnance Survey, 2020).	55
Figure 4. Comparison of the distance between neighbour centroid for square grids and hexagonal tiles. The distance between neighbours is always the same for hexagons, whereas diagonal neighbours are $\sim 1.4 (\sqrt{2})$ units away for square (given a distance of 1 for their adjacent neighbours)	56
Figure 5. Diagrammatic representation of the impact on aggregating a set of points to area counts, highlighting the scale (left and middle panels) and zoning effects (right panel). Numbers and grayscale indicate the number of points within each spatial unit for that partitioning scheme (darker shade of grey indicating a higher number of points). (Source: Mennis, 2019)	58
Figure 6. Snapshot of Molloy and Moeckel’s 2017 silo regionalisation method.	62

CHAPTER 3

Figure 1. Glossary of terms and example dataset illustrating each component’s role in the data collection	70
Figure 2. Location accuracy measurements of the in-app data impressions (sum of impressions recorded at each accuracy). Approximately 90% of all impressions in the raw dataset have an accuracy of 150m or less (red line)	72
Figure 3. Location accuracy measurements from another in-app dataset (proportion of observation per accuracy)(Wang et al., 2019).	73
Figure 4. Sum of daily impressions across the in-app dataset.	74
Figure 5. Distribution of unique devices recorded daily across the original in-app dataset. A large number of devices are onboarded before 2018, with an equally steep drop mid-2018, and a steady increase of devices through until late 2020.	74
Figure 6. Average number of impressions per device across the in-app dataset. This number increases from less than 10 in 2016-2018, to close to 200 impressions per device in 2020.	75
Figure 7. Mean impressions per month, coloured by season (blue winter, green spring, summer yellow and fall orange). September records the most impressions, with April being the least active month on average.	76

Figure 8. Mean number of impressions per hour throughout the in-app dataset. 5pm (17) is the hour that records the most activity, with a dip between 1-5am. 8am is the first abrupt peak of activity of the day on average.-----	77
Figure 9. Sum of apps in the dataset per year -----	77
Figure 10. Cumulative sum of app contribution to the dataset. The 10 largest apps contribute towards 85% of the dataset, and the 20 largest apps 95%. -----	78
Figure 11. CDRC procedures to access, analyse, output and present controlled datasets, such as the in-app dataset. (Lloyd, 2018) -----	80
Figure 12. Flowchart overview of the data digestion process, with the data digestions steps (comprising cleaning and aggregating) in green and data geodemographic linkage in orange when necessary (secondary step for conducting analysis and data explorations.-----	82
Figure 13. Different methods to count the data when aggregating the dataset. Each point type (circle, triangle, diamond) describes a different unique device, with each device creating multiple impressions.-----	86
Figure 14. Spatial distribution of the data nationally – The highest counts of impressions are recorded in Greater London. Other big metropolitan areas are visible with denser city centres. This helps in part justify why this report focuses on metropolitan areas, specifically London. -----	90
Figure 15. Coverage map, GB scale. Grey-blue grid cells represent activities being consistently below a count of 10 for each day in the dataset (IDEs). Red grid cells are above a count of 10 devices for at least one day of the data period. Missing grid cells show missing data throughout the entire data period (if a grid cell recorded one activity over even a single day over the four years, it would appear blue rather than white). Central Wales and Northern Scotland display large proportions of missing data. The data shows low significance on the national scale, with 85% of grid cells consistently recording counts of 10 or lower-----	91
Figure 16. Count of London OAs distributed by the percentage of their population represented by the night-time in-app data activity counts. For instance, the data seems to capture about 0.6% of the population in 6835 OAs. -----	95
Figure 17. Comparison of population proportion by group (outline) and proportion of activities recorded by group (fill). Below each category, the percentage difference between in-app data activities and resident population. An under-filled bar (and negative difference) shows an under-representation of that group in the in-app data sample, and positive an over-representation-----	97
Figure 18. Maps of London with coloured LOAC groups. Top: OAs with their corresponding LOAC group. Bottom: groups assigned to 1km ² grid cells through areal overlap and population weighted centroids. -----	101
Figure 19. 1km ² OSGB grid map of London, showing the activity level per unit over the analysis period -----	102

Figure 20. Bar plot – proportions of each subgroup’s contribution towards the total activity count for the data period. Blue represents ratios obtained by direct join, and red obtained by the 1 km² grid aggregate analysis. ----- 103

Figure 21. Line Graph – population profiles for the 1km² grid analysis (red) displayed as a function of the direct join to OA profiles (green). Each OA activity count was assigned the value of 100, and each 1km² grid cell activity expressed a percentage of OA activity. ----- 103

CHAPTER 4

Figure 1. Example of aggregated activity counts at the subset period and scale. The aggregation unit here is the 250m² OSGB grid. ----- 107

Figure 2. Difference between the control and aggregate tests. In both cases, aiming to obtain the number of activities per studied areas (areas A and B), but in the case of an aggregated dataset (right) this requires overlapping the aggregate zones to the studied areas and assigning the overlap, whereas direct counts can be assigned to the studied geographic from the control. ----- 108

Figure 3. Miniatures for each scale showing the distribution of WPZ subgroup assigned to grid cells or output areas through weighted spatial joins. ----- 111

Figure 4. Activity means (left) and subgroup proportion (right).----- 112

Figure 5. Ratio comparisons between control and 250m². ----- 113

Figure 6. Taken from the Lagonigro et al., AQuadtree R package description (See Appendix 2), shows three level quadtree splitting cells. The initial cell is on the left, at the centre is the first subdivision and the second quadtree subdivision is on the right (Lagonigro et al., 2020a). ----- 119

Figure 7. Three quadtree objects created using different thresholds, over the same dataset and period. Th=10 (left), th=25 (centre) and th=50 (right). Black cells are residual cells.----- 121

Figure 8. Left - in-app impressions collected across London on the 26th of March 2018. ----- 122

Figure 9. Left - unique devices recorded in London, 26th of March 2018. Right - unique devices, 30th of March 2020. Both maps were plotted using the quadtree hierarchical method, with a threshold of 10 unique devices minimum per cell. ----- 123

Figure 10. Left - unique devices recorded in Westminster over rush hours (8-10am) over the course of a week in September, aggregated and plotted using the quadtree hierarchical method, with a threshold of 10 devices minimum per cell. Right - activity aggregated to 250m² grid cells over the same data period (Sieg and Cheshire, 2021). ----- 124

Figure 12. Example Voronoi diagram. ----- 127

Figure 13. “Procedure for generating Voronoi polygons using HDBSCAN for POI” extracted from Hagen et al., 2022, p.7. ----- 128

Figure 14. Voronoi boundaries made with cluster centroid. Left - Clustering parameters are $e = 10$ MinPts = 10. Right - parameters: $e = 50$, MinPts = 10. A smaller value of e when clustering the dataset results in more compact Voronoi regions. ----- 130

Figure 15. Parameter impacts for the day of data. e20mPts10 is where the line of noise points drops below the number of clusters generated. The aim is to have a maximum number of clusters for a minimum amount of noise points. -----	132
Figure 16. Parameter impacts for the week of data. e5mPts100 presents surprising results with large amounts of discarded noise points and few final clusters. e5mPts10 is the most attractive option, with close to 5000 clusters (which would result in a similar number of regions), and less noise points than the following two options. e10mPts10 and e10mPts20 are still tested for their lower number of noise points. -----	132

CHAPTER 5

Figure 1. Flowchart of the regionalisation algorithm. PNMx signifies Preferred Neighbour Matrix. The divided parts (dotted lines) correspond to the methodology sections of the same number for further detail. This process is repeated for groups of hexagons within pre-selected merge boundaries (MSOAs) to obtain outputs nested in existing geometries. -----	138
Figure 2. Increase of unique devices (total counts) by scale, with examples of spatial divisions of London for H35 and H37. -----	141
Figure 3. Proportion of total counts (pre-omission) for each H3 resolution compared to resolution 14 total counts, plotted with the proportion of counts remaining post omission (due to IDE) at each resolution. -----	143
Figure 4. In-app data distribution at the London scale, September 2019. The Activity count per H310 per day corresponds to the number of unique devices per area. The Thames is white as the H310 are nested inside London MSOAs which are typically clipped alongside the river. Other white hexagons correspond to H310 which never contained in-app data throughout the sample period. -----	145
Figure 5. Densities of average unique counts per LSOA (blue) and MSOA (yellow), across London. -----	147
Figure 6. Diagrammatic illustration of the MSOA code assignment, with a synthetic sample of the data after this step. -----	148
Figure 7. Raster mosaic of terrain information, Greater London. -----	149
Figure 8. Illustration of the topo_code assignment process with codes designed for land use overlaps. -----	150
Figure 9. Example process of assignment of preferred neighbour. -----	152
Figure 10. Example of membership detection through graphing method and group generation. -----	153
Figure 11. H3BRs at resolution 10, nested in and coloured by MSOA. The map shows areas of varying density of activities, with cells remaining unmerged next to other region made of bigger merged groups. The tidal stretch of the River Thames is excluded from the output regions as census units, such as MSOAs, are conventionally excluded from it. -----	155
Figure 12. Process contextualising the use of the functions provided for the regionalisation. -----	156

Figure 13. Pseudocode snippet for the create_PNMx function. -----	157
Figure 14. Pseudocode snippet for functions "select_best_group" and "update_groups" with "get_n_details" being an important element of the first: the one which extracts the characteristics of it's neighbours once more -----	157
Figure 15. Maps plotting the cells omitted due to low count for OSGB250, OAs, H310 and H3BRs. -----	161
Figure 16. OA census geographies omitted due to low data counts (left) compared to area omitted from the H3BRs (right), when aggregating the same day of data points to each. -----	162
Figure 17. Diagrammatic explanation of Jaccard similarity calculations.-----	163
Figure 18. Intersect of the regions generated for Monday and Sunday. The areas in white are those that differ between the two regions generated for each day. Everything else remains identical between both. Monday and Sunday have Jaccard coefficient of ~0.25.-----	165
Figure 19. Comparison of Jaccard index between regions made with data from different days of the week ('Avg' corresponding to the 'average day' across September – mean of all days). Higher values on the y axis correspond to a greater overlap with the regions on the x axis.-----	166
Figure 20. Comparison of Jaccard index between regions made with varying thresholds. Higher values are more similar to other regions than lower values (for example 45 is has less overlap with other regions than 51) -----	167

CHAPTER 6

Figure 1. Comparison of proportions of counts and regions omitted for different days of data (temporal zones), for H3BR and OSGB.-----	178
Figure 2. Illustration of the making of the hour blocks. In order to capture significant blocks of time, a rolling average is kept for the hour blocks where required. For example, for the average 3-hour block, the activity count is calculated for all consecutive 3-hour blocks starting at midnight (0-2, 3-5 and so on) and all activity across these blocks are averaged to get the 'typical 3-hour block activity'. Where the block number is not a fraction of 24, a random hour is exempted (such as for the 5-hour block). -----	180
Figure 3. Percentage of regions omitted by hour-block for H3BR and OSGB250-----	181
Figure 4. Percentage of counts and regions omitted per hour for OSGB250 and H3BR-----	182
Figure 5. Proportions of hourly activity of a typical September day in London (left) and hourly activity for a typical week (right) -----	184
Figure 6. Map of H3BR regions coloured by time attributes. Centred on the London Borough of Camden. -----	186
Figure 7. Counts and regions omitted per MSOAs per hour. -----	187
Figure 8. Comparison of H3BR_D and H3BR_H at the London scale, with a focus on Camden and surrounding areas for ease of comparison of the partitioning differences. -----	189

Figure 9. Comparison of percentage omitted counts and regions per hour - H3BR_H added-----	190
Figure 10. Bivariate choropleth maps showing night and day activity levels across Greater London, aggregated at H3BR_H. Night activities range from 10 to 100 and day activities 10-150. Greys are under 10. Boroughs are outlined in white, and the zoom focuses on central areas.-----	192
Figure 11. Diagram simplification of region reassignment when data is added. In the top case, regions with data counts largely above 10 will be see little difference if an activity is or two is added. This is similar to comparing two days of data. However, starting H310s with low data, particularly counts hovering around 10, will be creating very different output regions if one count or two are added. This simulates what can happen when one hour sees slightly more counts than a previous hour when regionalising both separately, as the starting counts are lower, and the differences larger than between days.-----	196
Figure 12. Bar plots showing the proportions of omitted counts and regions per region type when aggregating 3am (left) and 5pm (right) Sunday data. -----	198
Figure 13. Maps of dominant ABP per region for each region types, focus on Camden. In order, from left to right and top to bottom: (1) POI original points coloured by ABP class (2) H3BR_H coloured by their majority (dominant) POI's class (3) same thing for H3BR_DT and (4) H3BRD_NT. Though the underlying POIs are the same as in the first map for all, the majority class varies spatially depending on the delineation chosen. -----	204
Figure 14. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H over Greater London.-----	205
Figure 15. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H across Camden.-----	206
Figure 16. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H across Westminster. -----	207
Figure 17. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H across Tower Hamlets. -----	208

List of Tables

CHAPTER 2

Table 1. Strengths and weaknesses of mobile phone location data for use in research (as described by the ONS).-----	45
---	----

CHAPTER 3

Table 1. Metadata table - Initial summary of descriptive statistics of the in-app dataset.-----	68
Table 2. List and description of variables in the raw in-app dataset (alphabetical order). Variables in bold are the key ones filtered for the making of data aggregates in Section 3.2.3. Other variables presented here are used for assessing the dataset's quality an accuracy and provide overall descriptive statistics presented throughout this chapter.-----	69
Table 3. Example tibble of the data at this stage.-----	83
Table 4. Assigned polygon IDs (Unit ID) and geometry to device ID hash, replacing individual impressions' latitude and longitude.-----	85
Table 5. Activity counts (sum of unique device IDs) per geographic unit per day per hour. The geometry can be kept for mapping purposes. -----	87
Table 6. Metadata table for the aggregated subproduct. -----	88
Table 7. Percentage of data omitted each year due to IDEs in the aggregated subproduct dataset. Comparison of GB extent and metropolitan areas. -----	92
Table 8. LOAC supergroups and associated groups with their colour code, as found in the LOAC documentation (Longley and Singleton, 2018).-----	96
Table 9. Activity counts per grid cell, with the grid cell's corresponding LOAC (synthetic example). Summing n_activities per LOAC returns the number of activities for each group.-----	99

CHAPTER 4

Table 1. Scales of aggregation studied-----	107
Table 2. WPZ classification subgroups used for this study, as categorised and described in the WPZ technical report (Singleton et al., 2017). -----	109
Table 3. Synthetic example of the activity count per WPZ subgroup, obtained for the aggregate made with the 500m ² grid (id500m). -----	110
Table 4. Principles to guide the definition of in-app data bespoke regions. Adapted from (Casado Díaz and Coombes, 2011). -----	115
Table 5. Sensitivity analysis results. *(e=epsilon, MinPts = minimum points). Shaded in yellow, the parameters kept for testing day of data Voronoi, blue those selected for the week of data, green is shared by both. -----	131
Table 6. Comparing the granularity and omission counts for the Voronoi regions, OSGB250 and OA. -----	133

CHAPTER 5

Table 1. H3 resolution statistics: average area and edge length of hexagons per resolution.-----	140
Table 2. Summary statistics of LSOA and MSOA activity counts-----	146
Table 3. topo_code assignment to their corresponding land use. The red entries illustrate the purpose of having decomposable numbers where layers overlap for ease of interpretation. -----	149
Table 4. Synthetic sample of processed dataset at atomic unit level to be used for regionalisation. -	150
Table 5. Comparing data omission when aggregating to H310-based regions against OA and LSOA aggregation. -----	159
Table 6. Remaining data points post omission of low counts for aggregates of similar scale but different zoning. -----	160
Table 7. Comparison matrix assessing if the principles are met by the regionalisation methods -----	172

CHAPTER 6

Table 1. Descriptive statistics of the specific hourly H3BRs. Mean per H310 prior to regionalisation, and number of resulting H3BR units post regionalisation-----	194
Table 2. Jaccard similarity indexes between hourly region sets. Colours correspond to relative similarity as indicated by the legend.-----	196
Table 3. Primary and Secondary POI classification categories as described by the OS ABP dataset. Classes which have low presence across London, or low relevance to the analysis, were not considered in the profiles. All 58 classes, including the ones not considered here, are listed in Appendix 3. -----	201

List of Abbreviations

ABP	AddressBase Premium
CDR	Call Detail Record
CDRC	Consumer Data Research Centre
CMA	Competition and Market Authority
CPU	Central Processing Unit
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DSH	Data Safe Haven
ED	Enumeration District
ESRC	Economic and Social Research Council
GB	Great Britain
GDRP	General Data Protection Regulation
GIS	Geographic Information Sciences / Systems
GLA	Greater London Authority
GPS	Global Positioning System
H3BR	H3 Based Regions
IDE	Insufficient Data Event
INSEE	Institut National de la Statistique et des Études Économiques
iOS	iPhone Operating System
LLMA	Local Labour Market Areas
LOAC	London Output Area Classifications
LSOA	Lower Super Output Area
MAUP	Modifiable Areal Unit Problem
MSOA	Middle Layer Super Output Area
MTUP	Modifiable Temporal Unit Problem
NPS	Network Positioning System
OA	Output Area
ONS	Office for National Statistics
OS	Ordnance Survey
OSGB	Ordnance Survey National Grid
PNMx	Preferred Neighbour Matrix
RAM	Random-access Memory
STP	Space Time Prism
TFL	Transport for London
UCL	University College London
UK	United Kingdom
WPZ	Workplace Zones

1. Introduction

Datasets generated by GPS-enabled mobile phones are increasingly used in research as such devices have not only become ubiquitous, but also proven their value across a range of applications in mobility analysis, such as dynamic population mapping, epidemic modelling and computational social sciences (Lazer *et al.*, 2009; Kitchin, 2013; Watts, 2013; Deville *et al.*, 2014; Bengtsson *et al.*, 2015; Jeffrey *et al.*, 2020). However, the high spatial and temporal precision of the mobile phone GPS location data make it personally disclosive, so to prevent the risk of disclosure a geomasking or aggregation process is required.

Data anonymisation - by removing personally identifiable information (such as names) - is often insufficient to prevent the re-identification of individuals when granular location data is gathered (de Montjoye *et al.*, 2018). In this context, aggregation assigns the precise coordinate pairs to an area-based geography – such as a grid – and counts the instances of these pairs as a proxy of activity within each area (Jeffrey *et al.*, 2020; Kishore *et al.*, 2020). A further step commonly occurs when the devices generating the location pairs are known and can therefore be aggregated to a ‘device count’ per areal unit, with devices treated as analogous to individuals.

Aggregated products are created for third party access, but unlike more traditional population datasets, no specific regionalisation has been made to aggregate and represent these new forms of data in a consistent way. This is important because the decisions taken in this process, such as the size of the areal units used, will have impacts on the analysis that follows, as it is a classic manifestation of the interactions of scaling and zoning detailed by Openshaw (1977, 1981). This is better understood in the context of the Modifiable Areal Unit Problem (MAUP) and can be further compounded by data redaction procedures that are commonly applied to further limit risks of disclosure, for example by filtering out the areas with device counts of less than 10 (a common threshold in this context). Small deviations in unit size can therefore amount to large variations in the areas that are redacted vs those that remain (Helbich *et al.*, 2021).

One approach to minimise the impact of the MAUP on aggregate datasets is to consider the scale and zone at which the data is aggregated. Regions that are made to closely fit the data and promote granularity could help create aggregates that correspond better to the original dataset. The England and Wales Census Output Areas (OAs) are one example of regionalisation strategies developed for a specific dataset and use case since they were conceived to closely fit the underlying spatial distribution of the data (Martin, 2000).

The overarching aim of this thesis is to develop and propose regions for an in-app mobile phone dataset, a specific subcategory of mobile phone GPS location data. These regions should fit the data's spatial distribution, much like the OAs do for census data, and help aggregate them in more consistent ways to reconcile data protection and preservation. Data-driven methodologies exist for this type of exercise, and some are presented and assessed throughout this work. Nonetheless, another key objective is to propose a more generalisable approach, challenging the current restrictions of bespoke regionalisation methods, which tend to be mostly proprietary and project specific. Considering the recent increased interest of researching consumer datasets, another aim is to maximize the ease of access to this unique dataset. This would help establish a framework for future researchers to analyse and interpret similar data, and in turn helps spark discussions around data dissemination standard practices.

The two key purposes of this work are thus to (i) develop a transferable method for the regionalisation of sensitive and dynamic mobility datasets and (ii) more specifically regionalise and provide a safe aggregated version of the granular dataset utilised for conducting this research. These objectives contribute towards increasing wider access and understanding of these growing, complex datasets and promote their safe use across the research community.

1.1. Aims

As explained above, the purpose of this research is to explore questions pertaining to the regionalisation of in-app data, provide a methodology for the creation of data-driven bespoke regions and give access to an aggregated case study dataset. This work anchors itself inside the fields of mobility studies and geographic analysis, which inform and impact the decision-making throughout the method's development. One such example is the consideration of time scales in a second time, to acknowledge the impact of time granularity on the dataset's usability in mobility studies. Benefitting from a rare access to a consumer in-app mobile phone dataset, the research also seeks to provide thorough exploration of the dataset, and, ultimately, an aggregate for safe use. On this basis, the thesis has five broad aims:

- (1) To review current practices in mobility analysis, highlighting opportunities for progression with new forms of data. This helps justify the need for safe data dissemination (Chapter 2).
- (2) To evaluate an individual-level in-app location dataset and investigate the effects of the MAUP on its aggregation (Chapter 3 and 4).

- (3) To regionalise the dataset, preserving privacy, granularity and analytical completeness and validity (Chapters 4 and 5).
- (4) To propose a transparent and versatile methodology for regionalising and aggregating large point datasets which could be applied to other comparable datasets and urban areas (Chapter 5).
- (5) To assess the impact of the Modifiable Temporal Unit Problem (MTUP) on the method's generalisation and propose recommendations for optimal use of the regions through changing temporal dimensions. Use the method as an opportunity to demonstrate and quantify MTUP (Chapter 6).

The key outputs from this thesis are reported in Chapters 5 and 6. Indeed, these chapters detail the regionalisation methodology and describe its outputs. However, their value and relevance are appreciated through the research foundations and crucial contexts comprised in Chapters 2, 3 and 4. A more complete thesis outline is provided below, outlining each chapter's key contents.

1.2. Thesis Structure

1.2.1. Chapter 2: Literature review

This chapter lays the foundational concepts of the thesis, focusing on time geography, and explores the evolving landscape of mobility analysis. The purpose of this literature review is to make explicit the narrative thread woven throughout the thesis, which follows these key points: (1) mobility analysis is crucial in geographic research and supports many fields ranging from social science to urban planning, (2) these analyses rely on complex or limited datasets capable of capturing granular information in both space and time (3) new forms of data have appeared which can greatly improve mobility research (4) these new forms of data constitute a paradigm shift and add in complexity, both technical and ethical. (5) Resolving some of these challenges requires extensive data alteration and (6) some of these alterations (here aggregation and regionalisation) provide promising solutions to reconcile privacy and accuracy when applying these new forms of data, in turn safely harnessing their established research potential. To develop these points, the literature review first provides an overview of research conducted in mobility analysis, focusing on traditional datasets, to highlight the paradigm shifts resulting from the uptake of new forms of data. This is important as the in-app mobile phone dataset presented throughout the rest of the work is largely considered as a potential source of novel

mobility insights. This is followed by a discussion around the challenges posed by the integration of these datasets in social and geographic research, with a particular focus on limited access and privacy protection. The discussion in this chapter later shifts to the strategies employed to mitigate these privacy risks, primarily through data aggregation. The final key section of the literature review describes the aggregation and regionalisation concepts applied throughout the thesis. For example, definitions of functional regions are provided, and key methodological elements, such as the H3 indexing system, are introduced in anticipation for their use in later chapters. The literature review thus seeks to provide ample theoretical background, setting the stage for a detailed examination of the dataset in the context of mobility research, and regionalisation in the context of data protection.

1.2.2. Chapter 3: Data Presentation and Preliminary Analysis

Chapter 3 presents the in-app mobile phone dataset provided for this research. Throughout the thesis, all tests are run on the dataset detailed there. The attributes of the dataset are provided, and descriptive analysis is conducted to familiarise ourselves with key in-app data characteristics. This data description is particularly important as the data-driven regionalisation methods proposed in future chapters are contingent on the dataset used for the shape of their outputs. This chapter then provides the main methodologies used to digest the data for analysis, taking the raw data through typical cleaning and aggregation processes. Further analyses are conducted using a processed version of the dataset, to assess the data's representativeness of the population. Finally, building on this last representativeness assessment, the chapter concludes on a comparison of results obtained from conducting this test with the raw data, compared to the standard practice aggregate. The preliminary analysis presented in this chapter thus serves two purposes: (1) describing the in-app data utilised throughout the work and (2) illustrating the impact of using common aggregates to assess data quality. At the end of Chapter 3, some concepts discussed in Chapter 2 have been applied in the context of the in-app data, providing justification for the research conducted in Chapter 4.

1.2.3. Chapter 4: Regionalisation - Explorations

Chapter 4 provides a more complete assessment of the MAUP's impact on aggregated products derived from the in-app data. It takes root in the uncertainties highlighted in Chapter 3 and compares analysis results using different aggregate scales and zones to quantify the issue. From there onwards, the necessity for regions which fit the data more closely is made evident. The

chapter then reflects on discussions provided in Chapter 2 around regionalisation methodologies and defines the key principles and practices to consider in constructing robust functional regions in this context. These criteria are used later in the chapter to assess two common data-driven and tractable regionalisation methods. The regionalisation principles appear throughout the rest of this work, as the final regions outputted by the thesis are later assessed against them. This chapter is exploratory in nature: it describes the various tests conducted in the search for a reusable regionalisation methodology, applicable to in-app data. These explorations were necessary to inform the making of the bespoke methodology provided in Chapter 5, and produced interesting examples to assess the credibility of the core regionalisation principles defined here.

1.2.4. Chapter 5: Operationalising H3 – Development of a bespoke regionalisation methodology

Chapter 5 presents the first regionalisation methodology built on H3 units for regionalising in-app mobile phone datasets. This is predominantly a methodological chapter, with the first half describing in detail the regionalisation algorithm devised to account for the criteria and discussions set out by previous chapters. This is achieved by creating atomic spatial units at which to initially aggregate the dataset, and combining them based on ranked criteria to account for size and terrain. The resulting regions are presented, yielding positive results. They preserve a significant amount of data when compared to traditional aggregates. Following the presentation of the results, a sensitive analysis is conducted to assess the method's volatility, and the chapter closes on a reiteration of the principles and criteria presented in Chapter 4, as the proposed bespoke regions are compared with the previous regionalisation attempts. The necessity to better account for temporal dynamics is discussed, particularly in the context of time granularity and mobility analysis: this pivots towards the temporal research conducted in Chapter 6.

1.2.5. Chapter 6: MTUP and times of low data – assessment of the H3-based regions

Following from the conclusions of Chapter 5, Chapter 6 assesses how changes of temporal scales and zones impact the regions' good fit to the in-app data. This chapter is largely exploratory and seeks to draw attention to key MTUP effects by using the new regions as a helpful analytical tool. This chapter begins by defining MTUP in the context of aggregation. It then evaluates the new regions across different temporal dimensions, to determine if the

method's benefits observed in Chapter 5 persist across various temporal dimensions. Building on this, regions are built for hourly data to assess whether changing the temporal input still produces coherent and effective regions for preserving data at times of low activity. These new regions are tested using previously developed methods to evaluate region performance. Later, points of interest are assigned to each new hourly region set, and the resulting spatial compositions are compared, demonstrating that different regions provide varying perspectives of the same fixed space. The conclusions highlight the importance of considering MTUP alongside MAUP when selecting aggregation and regionalisation scales. This chapter aims to offer key recommendations for the optimal use of the Chapter 5 method and showcase the potential of the algorithm as a tool for exploring MAUP and MTUP effects in tandem, guiding future efforts toward a more comprehensive space-time tessellation.

1.2.6. Chapter 7: Discussion

The final chapter is a summary discussion of the thesis' main takeaways. It seeks to reinforce the conclusions and insights drawn from this work. Key methodological contributions are detailed, and the limitations of the research are discussed, with a focus on data uncertainties and access restrictions. Potential applications for the regionalisation methodology, as well as the implications of the academic contributions as a whole, are presented. This chapter concludes on future avenues of research related to this thesis.

1.3. Notes on Software and Code

Most analyses in this thesis were undertaken in R Software for Statistical Computing (v4.4.0, R Core Team, 2024), an open-source program freely downloadable from www.r-project.org. Associated codes are available upon request, with the core regionalisation algorithm accessible from <https://github.com/lousieg/H3BR>. Other software utilised included QGIS.

1.4. Ethics

This research was approved by the UCL Research Ethics committee (Project ID: 4763/002)

2. Literature review

This chapter lays the foundational concepts of this thesis, focusing on mobility studies and time geography, and introducing the evolving data landscape of mobility analysis. It seeks to clarify the necessity to regionalise, aggregate and disseminate mobile phone datasets to harness their application potential safely, and does so by chronologically detailing the role these datasets could play in improving and developing informative mobility analysis across fields. Thus, after defining the key concepts, this chapter presents the traditional methodologies and types of data utilised in conducting mobility research. Subsequently, contemporary large-scale location datasets are introduced, examining their impact on the discipline. These new data sources, such as the mobile phone in-app location data studied throughout this thesis, enable new timely insights but also introduce significant technical and ethical challenges, particularly concerning privacy and disclosure control. After detailing these concerns, the chapter introduces the strategies employed to mitigate these privacy risks, primarily through data aggregation. However, it is shown that spatial aggregation can compromise data integrity, prompting a deeper investigation into data-driven regionalisation as a potential technical and ethical solution to this hindrance. Regionalisation for aggregation is hence presented as an opportunity to create more specific spatial units, which could reconcile the need for privacy with the preservation of the data's granularity and underlying characteristics valuable to mobility analysis. This chapter concludes by summarising the research objectives and setting the stage for a detailed examination of regionalisation strategies for in-app mobile phone data throughout this thesis.

2.1. Conceptual framework – time geography and mobility research

2.1.1. Key concepts and definitions

This thesis' overarching aim is to consider more closely how sensitive datasets are aggregated, to ensure their safe use by many. As the dataset analysed throughout (mobile phone in-app dataset) is considered as a novel form of mobility data, we here define what will be understood as 'mobility' throughout this work. We also lay out key concepts in time geography as these

contextualise much of mobility research conducted within the social sciences as a whole (Thrift, 1977). The importance of such mobility analysis is then demonstrated through the presentation of key studies, to justify the necessity to continue providing timely data. These studies also provide guidance regarding the ways mobility datasets should be handled to preserve and promote the key insights derived from their use.

2.1.1.1. Mobility

The term mobility is polysemous: it may describe the direction, potential and means behind the movement of individuals or groups, define an observed or predicted movement over differing timescales, or designate a whole field of study in itself. It is perhaps one of the most shared concepts across social science disciplines (geography, sociology, urban studies etc.) (Lassave and Haumont, 2001). In geography, mobility studies describe the movement of people (rather than goods or information). They are concerned with the underlying mechanisms and motivation behind movement along with the social, economic, and environmental impacts it has on places (Massey, 2005; Urry, 2007; Cook, 2018). Mobility studies can range from local to global - including tourism, migration, commuting etc - but tend to be mostly framed by innovation and methods of transport and communication (Lévy, 2000). There is no strict consensus as to when mobility started to become a key concern in social sciences. However, the emergence of transportation geography and urban planning research in the 1950s arguably laid the foundations for analysing daily urban movement of people (Garrison *et al.*, 1959; Berry, 1965). Within urban studies themselves, mobility is often approached through the lens of daily patterns of movements (commuting) at a local scale (Segaud *et al.*, 2003). This helps filter from the multiple definitions of the term mobility to outline a more readily applicable concept to frame this project: the travel methods and practices of a population within its regular (daily) context. This research adopts this definition to frame the assessment of datasets applied in the context of mobility studies: preserving the data features which can inform on these practices at their granular scales, and within daily time frames.

2.1.1.2. Time Geography

Research in mobility, travel patterns and behaviour has existed for decades. By introducing Time-geography in the 1960s and 70s, Hägerstrand proposes the idea that time should be considered with equal importance to space in geographic analysis and mobility studies

(Hägerstrand, 1970; Thrift, 1977). Time-geography is a physicalist and constraint-based approach to society which mostly relies on the concept of time-budgets: namely that individuals have limited time and spatial resources to perform tasks and goals, and that these constraints are the primary reasons behind some of one's decision-making and time-allocating (Hägerstrand, 1974; Miller, 2017). Among some of the main conditions for defining this concept, Hägerstrand mentions "the fact that every task has a duration" and that "movement between points in space consumes time" (Hägerstrand, 1975a; Thrift, 1977). All one's goals require some amount of time to complete, and must happen somewhere in space; there might also be some time allocated in reaching said space, and temporal constraints on its accessibility. This is summarised in literature as three main types of constraints: *capability constraints* (the ability to move or operate vehicles, the need to eat and sleep etc.), *coupling constraints* (certain activities requiring different groups and people to intersect, such as work) and *authority constraints* (such as a place's opening hours or rule of laws) (Hägerstrand, 1970; Thrift, 1977; Neutens *et al.*, 2011).

Within this context, the Space-Time Prism (STP) is defined as "a representation of the constraints limiting the time within which the individual can act" (Oxford Reference, 2011). With the development of STP, Hägerstrand focuses on measuring person-based accessibility as opposed to traditional place-based accessibility (Hägerstrand, 1975). This shift brings about new questions in understanding the movement of people and occupancy of spaces, with more focus on how to define places based on the times at which they are accessed, and the ways they are utilised and occupied. Thus, within mobility analysis, scholars like Hägerstrand drive a plural view of accessibility and movement: namely one that couples both space and time, but also person-based with place-based geographies. To approach research across these dimensions requires data which also contains varied facets, a challenge described in Section 2.1.3.2 of this literature review chapter. The complexity of conducting such types of analysis is rewarded by the proven merits of mobility analysis.

2.1.2. The value of mobility analysis

The value of mobility analysis lies in its potential to help inform on a wide variety of behaviours and traits. Understanding the flows of movements and mobility habits of large groups informs planners and policy makers, and STP concepts appear frequently in mobility research, particularly in the context of transport planning (Chai, 2013; Sahebgharani, Mohammadi and

Haghshenas, 2019; Qin and Liao, 2021). However, beyond legislative of policy purposes, the value of mobility analysis also lies in its potential to reveal telling characteristics of the individuals behind the movement patterns described.

Numerous studies have demonstrated the existence of "temporal rhythms" in human behaviour when considering the spatiotemporal factors and constraints on daily life described above by Hägerstrand. Lefebvre (2004) introduced the concept of "rhythmanalysis" which examines spatiotemporal rhythms across different scales, from individual to global scales. His theory suggests that everyday life is shaped by cyclic and predictable habits, schedules, and routines, as societal functioning requires the synchronization of practices to achieve goals (Edensor, 2016). Moreover, rhythmanalysis posits that different routines within spatial contexts are interconnected with individual identities. Urban rhythms, such as public transport schedules or workplace hours shape unique activity patterns derived from the daily movements of different types of people (commuters, working parents, students, seniors etc.), and require social synchronisation in organising travel (Grieco and Urry, 2012; Edensor, 2016). Places can thus be seen as spatial and temporal intersections for daily tasks. The distinctive characteristics of a place can be identified through its "polyrhythmic ensemble" (Crang, 2001). In recent years, these polyrhythmic ensembles have been increasingly applied to the understanding of individual socioeconomic dynamics. Studies have demonstrated repetition and predictability in individual activities, as well as variations in temporal rhythms among different social categories due to their varying space-time constraints (Axhausen, 1995; Kwan, 1999). The speed and method of movement of an individual can also be considered in itself as a marker of social characteristics.

French geographer Jean Ollivro further emphasises the value of time in understanding one's mobility choices: reaching a destination quickly (by accessing various means of transport), or on the contrary being allowed the choice to "take our time" and go slowly, can both be indicator of social class and opportunity of choice (Ollivro, 2005). Furthermore, understanding the ways spaces are occupied, and by whom, at what times, can help inform on a variety of commercial purposes (Berry *et al.*, 1962; Silva *et al.*, 2017).

A visit to a certain place at a certain time can be revealing of a person's characteristics as described above, but can also help predict where an individual may go next from there onwards based on previous visits and patterns of behaviour (González *et al.*, 2008). This informs decision making related to footfall and retail, shop openings and transport providers. Time

geography concepts such as STP can not only help identify these socio-economic patterns, but also analyse short- and long-term consequences of wider disturbances, such as pandemics (See Klapka *et al.*'s 2020 study on Time-geography approaches to assessing COVID-19 impacts), or accessibility to an environment (Miller, 2017). Thus, understanding mobility in both space and time, and analysing everyday activities helps reveal the ordered rhythms of both people and places, highlighting variations among social groups, and informing on large scopes of research and decision making (Lefebvre, 2004; Ollivro, 2005; Lager et al., 2016).

2.1.3. Traditional methods in capturing mobility.

Mobility research relies on the collection of individual data to understand the movement patterns, directions, and motivations described above. Traditionally, geographers and social scientists have relied on surveys and censuses to unearth mobility trends. Here, the common types of data traditionally used are described. This section covers national censuses and the information they can provide for mobility studies, and then delves deeper into the other forms of data previously used to complete the missing information between census years. The focus is here on the data types which were most commonly used before recent technological developments enabled passive data collection and larger datasets.

2.1.3.1. Census

Traditionally, the most important source of population estimates has been population censuses (Office for National Statistics, 2016a). Census data is collected periodically by governments, and includes detailed information relating to population counts, household compositions, and other demographic characteristics. While it is primarily composed of static characteristics at a specific time (residence or age for instance), it is still often used to understand general movements of population. This can be done for overarching migration patterns (for example, recording the nationality of respondents) but also the diurnal and larger scale¹ patterns this thesis takes interest in. Commuting and travel patterns can be inferred from census data related

¹ Note on geographic scale and resolution: In geography, the terms 'smaller scale' and 'larger scale' can be misleading. A 'smaller scale' map represents a larger area with less detail (e.g., 1:500,000), whereas a 'larger scale' map covers a smaller area but with greater detail (e.g., 1:50,000). This inverse relationship is due to scale being a ratio of map distance to actual ground distance. Similarly, 'resolution' in geographic context refers to the level of detail a map conveys; lower resolution means less granularity and a broader geographic coverage (akin to a smaller scale), while higher resolution provides more detailed features, corresponding to a smaller area coverage (similar to a larger scale map).

to employment or modes of transport used for daily movement (Office for National Statistics, 2016a). It can also help validate hypotheses of rhythm analysis described above by joining socioeconomic census data to responses related to commuting.

As it collects from a large proportion of the population, a national census is one of the most reliable sources of data for sociodemographic analysis (Office for National Statistics, 2016a). Both the England and Wales and the United States census aim to record the count of resident population in their country, and include questionnaires focusing primarily on housing and demographic information. Historically, the England and Wales census also includes further details pertaining to occupation, health, ethnicity and education. However, it is published once every 10 years: this makes it difficult, if not impossible, to capture transient or temporary movement such as hourly, daily, weekly, or even monthly and yearly patterns (Lenormand *et al.*, 2014). The main US census, collected by the Census Bureau, is also issued once every 10 years. However, since 2000, the Census Bureau has also started employing a short-form census for all households, with a longer, more detailed survey (the American Community Survey) conducted annually on a smaller percentage of the population. Other countries such as France have adopted a rolling census system: surveying a different sample of the population each year (Roux, 2020). This spreads the cost of running the surveys, allows for continuous updates and can accommodate changes in the questionnaire more flexibly. However, full results for any new questions are delayed until after five years of data collection, although some national and regional data can be generated more quickly, and sampling different portions of the population each year can lead to inconsistencies and potential errors that wouldn't occur from a full census.

Mobility patterns can evolve rapidly, and census data cannot capture real-time dynamics or emerging mobility trends. Furthermore, the national census is often collected at an administrative level, which can result in a loss of information on more granular scale at which daily mobility patterns may evolve (Minot and Baulch, 2005; Lenormand *et al.*, 2014). There is thus a need to explore avenues for data collection outside of census years and formats.

2.1.3.2. Capturing population information between censuses

To capture information on human movement outside of census data, researchers have been making use of other sources of information such as time-space diaries, archives, transport and traffic data, or small-scale surveys conducted directly by researchers seeking to answer to

specific data needs (Axhausen, 1995; Bertini and El-Geneidy, 2003; Hubrich and Wittwer, 2017).

Time-space diaries, integral to time-geography since the early 20th century, allow individuals to record their movements and activities over time (Axhausen, 1995). They provide a window into everyday life and can take many forms: paper or digital journals, written or photography formats, and in more recent years coordinate locations collection (Axhausen, 1995). They have been pivotal in shifting urban and transport studies from supply-focused to demand-focused strategies, aiding in route choice analysis, and understanding social and temporal dynamics in cities (Axhausen, 1995; Kwan, 1999; Schwanen, 2009). Another use case of time diaries includes health studies, where they help track how mobility patterns relate to physical activity and exposure to environments, providing insights necessary for public health interventions and lifestyle health assessments (Jiron and Carrasco, 2020). They also aid in understanding community dynamics and the impact of spatial configurations on social behaviours (Jiron and Carrasco, 2020).

Beyond time-space diaries, other types of data have been used creatively to analyse mobility in more granular ways. One such example includes the tracking of dollar bill journeys across the United States as a marker for mobility (Brockmann et al., 2006). Specific mobility behaviours may incur passive data generation (such as purchasing a train ticket), largely in the form of archives from transport companies (Bertini and El-Geneidy, 2003). However, these specific behaviours may not allow for the breadth of coverage allowed through diaries, which can be designed to collect information outside of travel and transactional behaviours (Axhausen, 1995; Kwan, 1999).

However, these forms of data collection present salient issues: they require the generalisation of small samples of the population (sometimes only a few participants) to draw wider conclusions, they can be costly (especially in the case of research-specific surveys), and they may be subjected to issues of internal validity, as they rely on individuals accurately recording events over sometimes long periods (Miller, 2010; Kitchin, 2013). Though they offer insights outside of census years and can be curated for specific research questions, these sources of data present little opportunities for scaling up, or guaranteeing reproducibility (Miller, 2010).

2.2. The impact of new forms of data on mobility analysis

Recent history has seen the emergence of new types of data, such as social network data, mobile phone GPS locations, or other consumer datasets (collected through internet usage, travel, or loyalty card information etc.) (Lazer *et al.*, 2009; Miller, 2010). The rapid uptake of technologies such as smart phones supports a transition towards “computational” social sciences, reducing the need to rely on the smaller and sparser data described earlier (Kitchin, 2013). Most of these novel datasets contain both spatial and temporal information and are of particular interest to geographic analysis and mobility studies (Goodchild, 2013; Miller and Goodchild, 2015). This appearance of large-scale human generated datasets has been discussed as a great opportunity, but also a sizeable technical and ethical challenge for social researchers (Lazer *et al.*, 2009; Kitchin, 2013).

This section describes the provenance of these new forms of data before diving into how they have shifted the ways geographic research is conducted. It provides a literature review of social science and geographic research conducted using mobile phone data in particular, the data type this project makes use of later. This section aims to demonstrate the potential of these datasets in complementing traditional statistics and data-collection methods, particularly in creating opportunities for research at higher temporal and spatial resolutions.

2.2.1. New forms of data: provenance and definitions

2.2.1.1. New, ‘Big’ data:

Big data refers to datasets which cannot be analysed using traditional methods due to their complexity and volume (Jain *et al.*, 2016). They are often passively generated by day-to-day activity, collected through Wi-Fi networks, location-based services, social networking, transactions (online purchases and credit card metadata), mobile applications (apps), amongst others. Their advantages come mostly from their *velocity* (frequently and continuously collected), their *variety* (diverse data sources), and their *resolution* (often more detailed than traditional datasets in both temporal and spatial scales) (González-Bailón, 2013; Kitchin, 2013; Jain *et al.*, 2016). Rather than their sample sizes in themselves, their value lies in their passive collection which reduces the issues pertaining to bias introduced by actively participating in a study (Miller, 2010). ‘Big data’, as a term, thus encompasses multiple categories of new forms of data. The rest of this work focuses on a subtype of consumer data, mobile phone in-app data. This is defined in more detail below.

2.2.1.2. Consumer data definition

According to the UK government's Competition and Markets Authority (CMA), consumer data is defined as “any information firms might collect from and about consumers that is used, or intended to be used, to support commercial activities” (CMA, 2015). This includes passively generated data and third-party data such as mobile phone data in its various forms (call detail records, in-app information etc.), or data generated as part of travel, leisure activities, communication etc. Typically, one principal challenge of using consumer datasets is their accessibility. Most often collected by commercial organisations, these datasets are not open source, often expensive, and rarely offered to researchers (Lansley and Cheshire, 2018). However, as a result of the COVID-19 pandemic, access to these big mobile phone datasets was facilitated, with the aim of informing policy and timely health responses, and determining the effectiveness of lockdown measures. Programs such as Google Mobility reports promoted the use of aggregated mobile phone locations for research purposes (Google, *COVID-19 Community Mobility Reports*, 2020). This recontextualised the accessibility of consumer datasets in research, encouraging their wider use.

2.2.1.3. Mobile phone data

Mobile phone data is becoming increasingly ubiquitous, as more than 95% of the UK population now uses mobile phones (Ofcom, 2020; Raento *et al.*, 2009). As mobile device data collection increases, it has the potential to reveal detailed information on the everyday spaces and lives of individuals (Cheshire, 2020).

Mobile phone datasets usually fall under one of two categories: *call detail records* (CDR) data and *in-app Global Positioning System* (GPS) data (Kishore *et al.*, 2020). CDR data is gathered by mobile operators and offers an approximate location collected by the cell towers when a phone connects to the mobile network. In-app location data is obtained from GPS sensors in smartphones. It can be collected for specific services (such as Google Maps routing), or collected by apps and sent to third party location data aggregation companies. These two types of mobile phone data are examples of *location data*: namely data which carry specific geographical coordinates and situate events and devices in space. Throughout the thesis, we use the term *in-app data* to refer to mobile phone in-app GPS data. The specific dataset is presented in Chapter 3.

In-app data is significantly less representative of the wider population than CDR data, as it relies on the use of specific apps to collect locations. However, while CDR datasets are much more comprehensive in terms of their coverage, their locational precision is dependent on the density of cell towers in the vicinity and never attains the levels of accuracy and precision provided by a GPS dataset, considered more appropriate for granular mobility analysis (Kishore *et al.*, 2020). Furthermore, CDR data is often only distributed by mobile operators, whereas the GPS data market is composed of more diverse actors (such as data brokers or social media platforms), making them more accessible (See Keegan and Ng, 2021 for a list of data providers). In 2022, the global location data intelligence market was estimated at over USD 16 billion and is forecast to grow at a rate of 15.6% over the next 7 years, mainly driven by the growing penetration of smart devices in the population and the increased availability of in-app data (GVR, 2022).

2.2.2. How these new datasets reshape geographic research.

2.2.2.1. Paradigm shift - new forms of knowledge

These datasets present opportunities for researchers, but also impact research practices. The appearance of large location datasets (such as the in-app data described above) constitutes a paradigm shift in computational social science, from scarce to big data, and from sample selection of data for answering specific questions to the reconversion of pre-existing datasets into research-ready content (Goodchild, 2007). This can be considered a move away from idiographic towards nomothetic approaches to conducting research: from mass data to conclusion rather than from hypothesis to data collection. This data-driven approach has been coined the ‘Fourth Paradigm of Science’ (Kitchin, 2014a). Subsequently, data-driven geographies have emerged, with the potential to answer questions previously not conceivable in the data-scarce landscape, especially at granular scales, or in analysing daily and ‘mundane’ occurrences thanks to passive and frequent collection (Cook, 2018; Wang and Chen, 2018).

New forms of data and software have also created new spaces of virtual engagement, called “code spaces”, as well as new avenues to collect information on the individuals in these spaces (Kitchin and Dodge, 2011). Code spaces range from workplaces reliant on web apps, to travel systems requiring digital forms of payment, to mobile apps that facilitate everyday movement and interactions (Kitchin and Dodge, 2011; Cheshire, 2020). For example, using a running app to record physical activity turns the city into a code space, adding another dimension of

interactions between individuals mediated by the increasing interconnection between people and software (Cheshire, 2020). This also adds another dimension in the way research can be conducted with these datasets: they create new spaces and fields of study which would be inexistant without the presence of the software, or the behaviours formed around it. In this sense, data becomes a form of feedback-loop, being both the research tool and the research topic at the same time.

2.2.2.2. Mobile phone data use cases in social sciences

Mobile phone datasets (CDR and in-app GPS data) have provided insights across various fields, as concisely summarised by the Office for National Statistics (ONS) in a working paper series on statistical uses for mobile phone data (Office for National Statistics, 2020). Amongst multiple statistical applications, they propose a list of international papers attesting to the usability of mobile phone data on the topics of population estimates, mobility, land use and epidemiology.

In a study conducted on Belgian population, De Meersman et al. demonstrated that aggregated mobile phone data can provide accurate population estimates complementing traditional statistics (De Meersman *et al.*, 2016). Deville et al. created estimates of population densities at national scales using CDR from Portugal and France. Salat, et al. (2020) produced similar outputs in Senegal to compensate for missing, or incomplete census data. They show that comparing population densities with average phone activities return low correlations, but propose a method using daily, weekly, and annual activity curves, providing much more accurate estimates (Salat *et al.*, 2020). These papers demonstrate how mobile phone data can serve as a proxy for evaluating populations outside of census years, at more granular temporal scales, and even to support the census with more detailed information (Oyabu *et al.*, 2013; Bwambale *et al.*, 2020).

Mobility analysis often requires timely, granular data. Countless papers in the field have thus used mobile phone data specifically to trace movement (Birenboim and Shoval, 2016). These datasets have shown that human mobility patterns can be predictable and dependant on previous movements, supporting the rhythmanalysis theories discussed in Section 2.1.2 (Lefebvre, 2004; González *et al.*, 2008; Alessandretti *et al.*, 2020). To name a few examples, Silva et al. (2017) use mobility data from mobile apps to predict patterns in new-venue visitations. Becker et al. (2013) use mobile phone datasets to map daily commuting patterns,

and a paper by Willberg et al. (2021) traces the escape from Finnish cities over the COVID-19 pandemic.

Building from the need of population statistics to inform the built environment, Tooru and Kawakami (2013) use mobile phone data in urban planning, advising on where best to implement new bus networks based on needs, or analysing how the land use of a city shifts over the course of a day as residents move out of residential areas and into the dynamic centre for work. Reades et al. (2007) use these datasets to construct a different vision of the city as a dynamic system, particularly in letting researchers represent real-time dynamism changes in city centres. They apply clustering methodologies to analyse the neighbourhoods and venues of Rome most active at various times of the day and week, building a good conception of tourist flows. Similarly, Sevtsuk and Ratti (2010) use mobile phone data to infer on the periodicity of cities, echoing themes of space time prisms evoked earlier (see Section 2.1.1.2), and highlighting the potential of these datasets for novel temporal mobility analysis. The Alan Turing Institute is currently conducting research on the development of digital twin cities, using various data sources including mobile phone location dataset. Such a digital twin would allow researchers and planners to model the transport, health, and supply chain of cities (Alan Turing Institute, 2021)

Over the COVID-19 pandemic, mobile phone data helped produce insights on the spread of disease and the efficiency of lockdown measures (Basellini *et al.*, 2020; Iacus *et al.*, 2021). Some research coupled mobile phone data with models of virus transmission to predict the rate of virus penetration within the population and propose measures in accordance (Gibbs *et al.*, 2021; Iacus *et al.*, 2021). Oliver et al. (2020) used mobile phone data to inform public health action over the pandemic. Chang et al. (2021) paired similar datasets with mobility metrics to inform reopenings in the United States. Several other papers used mobile phone data applied to mobility analysis to estimate the level of compliance with lockdown rules (Jeffrey *et al.*, 2020; Kang *et al.*, 2020). Similar studies were conducted in China, testifying for the reproducibility of the methods (Zhou *et al.*, 2020). In such unprecedented times, the timeliness and granular information provided by mobile phone data was a strong tool to visualise unpredictable patterns in real-time, and at various scales (global, national, city or neighbourhood). In the UK, the last census having been in 2011, new types of data were a necessity to make up for a gap in statistical data over the 2019-2020 period of COVID-19 measures (Iacus *et al.*, 2021). Notably, mobile phone data had been used to support

epidemiological research well before COVID- 19 (Wesolowski *et al.*, 2015). Bengtsson et al. used similar data to predict the spread of cholera in 2015 in Haiti (Bengtsson *et al.*, 2015). Beyond epidemiology, mobile phone datasets have been used in healthcare to estimate one's exposure to behaviour-based health factors, such as obesity in a 2022 study by Zhou et al. relating it to physical activity measurements (Zhou *et al.*, 2020).

To summarise, there is a plethora of studies which make use of mobile phone data, across disciplines. These datasets not only provide complementary information between census years and support traditional statistics: they also allow for novel analysis at more granular temporal and spatial scales, particularly in the context of urban studies, mobility and geodemographics. Their usage has drastically increased over the past 3 years with their timely epidemiological applications, particularly in the context of the COVID-19 pandemic.

2.3. The technical and ethical challenges of mobile phone data

These datasets present opportunities as detailed above, but their sensitive nature means that they are accompanied by additional challenges around coping with increased uncertainty and the ethical concerns that come with using data that has been generated as a by-product of other activities (González-Bailón, 2013; Kitchen, 2014a). Table 1 describes mobile phone data's strengths and weaknesses as laid out by the ONS, (Office for National Statistics, 2020).

Table 1. Strengths and weaknesses of mobile phone location data for use in research (as described by the ONS).

Strengths	Weaknesses
Timeliness and frequency of collection	Complexity of access (commercial sensitivities, cost, infrastructure etc.)
Passive data collection	Concerns around data protection, security, ethics.
High coverage	Inference (needing to presume the nature of a movement, for example)
Higher accuracy of collection than surveys	Bias and Sample Bias
Small geographies (OA and lower)	Uncertain quality of estimates
Widely applicable	
Consistent over time	

The weaknesses are thus separated into two categories to further discuss below:

- (1) Issues surrounding accessibility, transparency, and technical challenges.
- (2) Concerns around data privacy and ethics.

This section seeks to address these challenges, and provide a background on the ethical discussions surrounding privacy and disclosure control. Firstly, the issues surrounding the difficult accessibility of the data, the reduced transparency behind its collection and dissemination and the technical challenges faced by researchers seeking to utilise them are introduced. In a second part, the privacy and ethical concerns brought about by their use are treated, with a more specific dive into disclosure control and regulations.

2.3.1. Accessibility, transparency, and technical challenges

“The challenge of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity.”
(Kitchin, 2014, p.2)

Consumer datasets are proprietary by nature (CMA, 2015; Lansley and Cheshire, 2018). Collected, cleaned, and processed by third parties, a lot of mobile phone data aggregates thus vary in quality and transparency. As articulated by Willberg et al.: “data products with undisclosed methodologies create new challenges, such as what exactly data represents as well as the compatibility of the terminology” (Willberg *et al.*, 2021, p.4). Brunsdon and Comber describe the consequences of this lack of transparency. All aspects of an “answer” generated through research should be able to be tested – by remaining in a “black box” third party collection and aggregation methods do not follow this principle and make the research less reproducible and verifiable (Brunsdon and Comber, 2020). Handling big data critically implies understanding all aspects of the data’s journey, from generation to digestion and cleaning all the way to final analysis results.

This lack of understanding of, or sometimes access to, metadata results in the potentially inconsistent quality of the data provided. Data quality issues can originate from data collection methods, whether technical or human-induced errors (Lansley and Cheshire, 2018). Consumer data stemming from new technologies has not always been subjected to rigorous quality controls suitable to meet research standards, often prioritising volume of data over quality (Dalton and Thatcher, 2015).

Another complication of using mobile phone datasets or other large datasets for research is the technical knowledge required to process them (Oyabu *et al.*, 2013). Social scientists and

geographers are not traditionally trained in handling computationally intensive databases. Understanding, treating, and communicating these datasets requires a technical numeracy, creating more resistance for their wider uptake in social science research. They also require a knowledge of the technology collecting the data – what sensors are being detected and how, are they GPS or cell tower generated etc. (Raento *et al*, 2009). It is often not sustainable to have a large majority of researchers work with sensitive data as both purchasing the datasets from the providers and training the research staff incur a cost, making these datasets expensive in both financial and human resources.

2.3.2. Concerns around privacy and ethics

2.3.2.1. Bias

Big datasets, like any human generated data, are subject to bias. More data does not necessarily equate to a decrease in data bias (Crawford, 2013; Kitchin, 2014a). As stated by Crawford *et al.*: “data may seek to be exhaustive and capture everything, but it will always be subject to the technology and collection techniques used”. Consumer datasets are especially prone to these issues. It is not uncommon for a minority of the population to be contributing to a majority of consumer data (Lansley and Cheshire, 2018). Mobile phone datasets may be biased towards the populations which use the devices the most and underrepresent certain age groups and demographics.

Bias may also appear from data interpretations and choices made when aggregating, cleaning, and processing. Data does not arise from nowhere and does not speak for itself; a theoretically perfect dataset may still be interpreted with bias stemming from the researcher’s personal experience (positionality) (Kitchin, 2014b). It is important to systematically recontextualise results, and not treat big data as all-encompassing (Crampton *et al.*, 2013; Kitchin, 2014a, 2014b). Brunsdon and Comber view robust data science as a practice which promotes transparency and reproducibility (Brunsdon and Comber, 2020). Transparency helps limit the impact of the researcher’s positionality on the interpretation of data (Sutherland *et al.*, 2012). Reproducibility is ensured in turn by transparency – by providing methods, one ensures all aspects of an analysis can be tested and verified, and the research can be reproduced (McNutt, 2014; Brunsdon and Comber, 2020). This can help limit the impact of researcher positionality and potential biases and uncertainties of the data on final analysis. They also support future decision-making, informing research conducted with these datasets.

Charters have been drafted to outline new standards of practice to protect individuals and public interest alike when conducting research using location data such as in-app data. The Locus charter, provided by EthicalGeo, is one such initiative, proposing a better understanding of the risks brought about by location-based analysis in geography (EthicalGeo, 2021). One of its founding principles is understanding impacts, highlighting users' responsibility in understanding social context and knowing who might be affected by the use of the data. Numerous studies have highlighted the ways in which data collection biases may leave whole populations unaccounted for in analysis relying on big data (Taylor, 2015; D'Ignazio and Klein, 2020; Williams, 2020).

2.3.2.2. Consent

Since its coming into effect in May 2018 in the UK, the General Data Protection Regulation (GDPR) has been regulating data protection and privacy. The GDPR specifies that consent must be “freely given, specific, informed, and unambiguous”. It adds “consent cannot be assumed from inaction” (Art 7, GDPR, 2018). In the case of research-specific datasets, consent is usually collected once at the start of the study by the researchers, and participants are made aware of their right to withdraw. For passively generated data, what constitutes as “consent” is hard to define, particularly if the data is subsequently used by other parties than the ones collecting it (Degirmenci, 2020). Mobile apps must comply with GDPR and request user's consent (often in the form of a notification) to collect their locations. However, the user is not necessarily informed of future usage of this data – they may not be explicitly notified that the locations are being used for a statistical study, for instance (Georgiadou *et al.*, 2019). About “transversal consent”, van Dijck says: “the notion of trust becomes more problematic because people's faith is extended to other public institutions (e.g., academic research and law enforcement) that handle their (meta)data.” (van Dijck, 2014, p. 1).

This uncertainty around informed consent raises legitimate concerns around data usage, information, and control over one's data. An article from MIT displays how outrage was sparked in Singapore after COVID-19-specific tracking data was used for criminal investigations by the local government (Han, 2021). These concerns explain the complications in accessing certain big consumer datasets which are protected by commercial agreements, but also consent restrictions. Raento *et al.* (2009) suggest that mobile phone data research

paradoxically benefits from the unobtrusiveness of mobile phones – by forgetting they are a part of a study, users behave in a more natural way. However, this unawareness conflicts with the concept of informed consent. A middle ground must thus be met – researchers either deem it acceptable and ethical to use data where permission to record has been given once, such as for in-app datasets, or choose to implement more recurrent, and potentially obtrusive consent requests. This second option may be preferable when dealing with especially sensitive topics or vulnerable users (Raento *et al.*, 2009; Han, 2021).

Contextualising bias and consent in this work helps us grasp what is at stake when using in-app data. This research focuses mostly on privacy and disclosure control, which we define more in depth below. However, it is important to note that ethical use of location data does not solely equate to a non-disclosive use. There is a responsibility to ensure that the data has been collected appropriately, and that biases inherent to that data, its collection, and its processing must be considered at every stage.

2.3.2.3. Privacy, confidentiality, and disclosure

Privacy is “the thorniest challenge” of location data usage (Lazer *et al.*, 2009, p.4). The use of any granular location dataset is subjected to disclosure risk mitigation strategies and privacy legislations. Eurostat defines the ‘disclosure problem’ as ‘the possibility of identifying individuals through released statistical information’ (Helmpecht and Schackis, 1996). The UK Information Commissioner’s Office (ICO) also considers personal data as data that can identify an individual either as is or in combination with other information. The ICO highlights that pseudonymised data, though reducing risk, is still personal and thus still subject to the GDPR (ICO, 2022). Tracking individuals, and releasing resulting information in any form, carries ethical and legal implications and must be done with knowledge of and control over the technologies and information involved (Kitchin, 2013; Song *et al.*, 2014). Privacy encompasses the ability to share information selectively, not publicly. Maintaining subject’s confidentiality is paramount to ethical research, and a widely debated topic in the wake of the uptake of big location datasets (Jain *et al.*, 2016)

Previously, anonymisation was deemed an adequate method to protect user’s confidentiality and privacy. However, evaluations of varied anonymisation and encryption methods revealed that it is often possible to reverse-engineer carefully anonymised data. Only 4 locations of an anonymised user will suffice to retrace the individual’s identity in a mobile phone dataset (Zang

and Bolot, 2011; de Montjoye *et al.*, 2018). For in-app data more specifically, a study by Sekara *et al.* (2021) sampled 3.5 million app users across 33 countries and demonstrated that 91.2% of individuals could be re-identified using other public information, if the data was only anonymised rather than coarsened (see also Achara *et al.*, 2015). Mobile in-app data, though pseudonymised, thus falls under ICO and Eurostat definitions of personal data and are most often only accessed by third parties (including researchers) through restricted access. (ICO, 2022). Beyond this, the public have generally expressed concerns regarding a lack of understanding and control over the data collected about them (Almuhimedi *et al.*, 2015; Zhang and McKenzie, 2022).

Governmental regulations provide actionable steps data users must take to protect individual privacy. However, these sometimes do not account for issues of re-identification described above. In practice, a variety of methodologies beyond anonymisation are used to guarantee privacy protection and minimise the harms of sensitive information disclosure. Four main models of secured access are listed below (de Montjoye *et al.*, 2018):

1. Limited release: only transformed data, containing no sensitive information, is provided.
2. Pre-computed indicators: researchers obtain indicators aggregated across individuals rather than individual data.
3. Remote access: analysis happens within a controlled environment, and only aggregated results are outputted after a disclosure check.
4. Question-and-answer: researchers access aggregated data through specific queries and do not have access to the original, sensitive points.

For individual-level data, the primary approach to preventing the disclosure of sensitive information is to strictly restrict the access to the dataset, specifying the types of analysis that can be conducted, and requiring that outputs and visualisations be checked for disclosure risks prior to release. Following this limited release model, in the Consumer Data Research Centre (CDRC), in-app data is stored on secured servers and is only available to trained researchers, with data outputted only after disclosure checks. These disclosure checks rely on data manipulations to guarantee compliance with disclosure controls so that only data transformed to prevent disclosure can be removed from the controlled environment to be shared with third parties (Griffiths *et al.*, 2019). More detail is provided on the types of checks performed for this research in the data presentation chapter (Chapter 3).

Failing to properly protect sensitive location information carries personal, social, and legal consequences: individuals in these datasets must be protected, and their right to privacy remains a priority. This explains why individual-level datasets are strictly regulated, despite this creating friction against aims of accessibility, reproducibility and transparency listed previously in Section 2.3.1. The next section presents the data-alteration techniques which can allow for safe access to these datasets outside secure environments. Aggregation is introduced as a common way to reconcile some of the challenges listed above: a data-alteration method which provides non-disclosive insights into mobility data.

2.4. Data aggregation: regionalisation to reconcile disclosure control and granularity.

As explained above, issues pertaining to disclosure control make mobile phone location data particularly sensitive and difficult to access by the wider research community. In this section, two traditional ways of creating safe-for-use data products out of sensitive data are presented: geomasking and aggregation. However, these processes impact data quality along with analytical validity and completeness. This section provides contextual literature on spatial scale problems and highlights how aggregation processes bring about inconsistencies of scale and zone impacting both data and analysis. In a final part, existing efforts are highlighted in developing geostatistical and regionalisation methodologies for the creation of bespoke units to preserve data quality during the aggregation process, particularly for mobility data.

2.4.1. Aggregation and geomasking as disclosure control

The UK Government and Office for National Statistics (ONS) both recommend restricted or controlled access to sensitive or potentially disclosive datasets (Office for National Statistics, 2017; ICO, 2022). Beyond controlled access, data can be shared if the disclosive elements are removed. Often, for location data, this process takes the form of spatially aggregating data products to a coarser scale, removing individual points and instead providing a count of events for an area. This type of aggregation is a common and proven method for preventing spatial data disclosure (Skinner *et al.*, 1994; Domingo-Ferrer and Mateo-Sanz, 2002; Office for National Statistics, 2017; de Montjoye *et al.*, 2018; Forgó *et al.*, 2021). Other ways of masking locations (geomasking) are also used for disclosure control, often consisting of randomising the data or ‘jittering’ the locations to mask the original points (Young *et al.*, 2009; Hut *et al.*,

2020; Wang *et al.*, 2022). Figure 1 thus summarises the three main processes for transforming data to prevent disclosure: aggregation, geomasking, and modelling. The three techniques presented may be used independently, but may also build upon one another, for instance, a model could be trained with aggregated rather than individual data

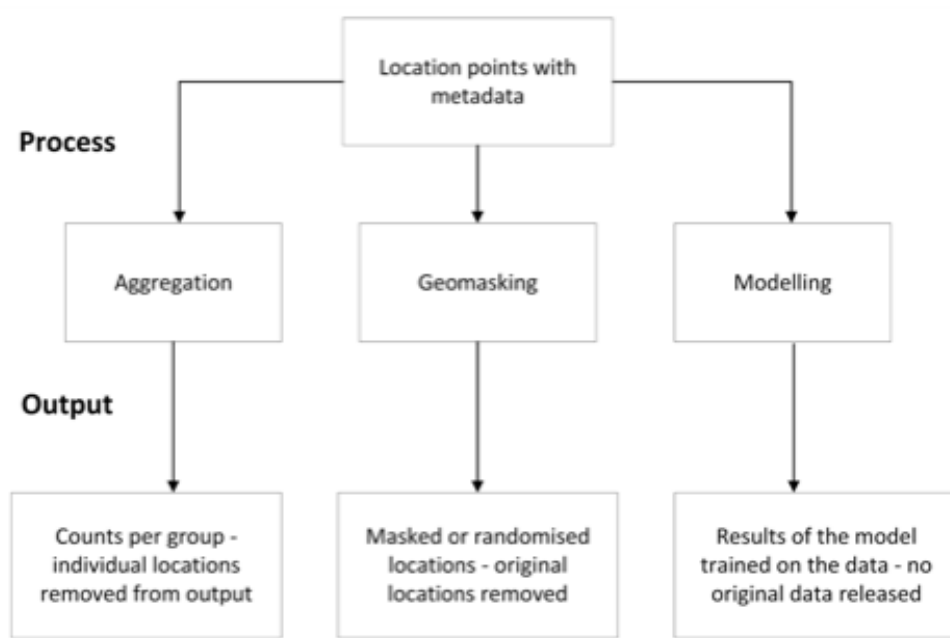


Figure 1. Flowchart of data manipulations for removing disclosive information from in-app datasets. The original location points are stored within a secure environment, and only the outputs are removed. For aggregation, counts below a threshold (typically, 10 individuals) are suppressed from the outputs.

2.4.1.1. Geomasking and modelling

Geomasking describes a variety of processes aiming to mask the original coordinates comprised in location data whilst preserving the spatial relationship between the original points (Zandbergen, 2014). This can be done by changing scale, rotating coordinates, or adding random noise to the original coordinates, each method presenting various advantages depending on aims (Zandbergen, 2014; Wang *et al.*, 2022). Geomasking is preferable when studying the relationship between spatial locations rather than the count of occurrences and is often used for health data analysis (Allshouse *et al.*, 2010).

In a modelling process, an algorithm is trained using in-app location data to produce predictions about a system. Only the model's results are outputted, thereby obscuring the individual-level

data used to inform the model. In the context of sensitive in-app data, a model can be trained inside a secure environment and these data do not need to be aggregated or otherwise altered in order to inform the model. This is a preferred method for real-time analysis but requires regularly updated datasets (Lange and Perez, 2020).

However, for data dissemination, aggregation (the first process in Figure 1) has traditionally been the preferred candidate: it is a more straightforward and reproducible data transformation method, and allows for the creation of reusable data products for further use (de Montjoye *et al.*, 2018; Wang *et al.*, 2022). For many applications, geomasking and modelling techniques can prevent independent verification of results and reproducibility (McNutt, 2014; Phillips and Knoppers, 2019). For instance, the United States Census bureau has recently adopted differential privacy methodologies for protecting individual information in census results, a decision which has been criticised for its potentially negative impacts on data quality and usability (Garfinkel, 2022; Ruggles and Van Riper, 2022). The rest of this work focuses on spatial aggregation: generating counts of points per given area

2.4.1.2. Aggregation

Spatial aggregation, in the context of in-app data, consists of scaling up the information from device level to group level: individual-level location points within spatial units are summarised to describe the number of devices or the level of activity within each area. A minimum number of points (typically 10) are required to fall within each spatial unit. If this threshold is not met larger units are required, or the points are redacted from any output to prevent disclosure (the technical details of this process are provided in Chapter 3). Although aggregation masks individual information, it also increases the level of uncertainty in a dataset in exchange for increased privacy. Figure 2 is an illustrative example of a typical spatial aggregation process: counting the number of points per area and turning the unidimensional data (list of coordinates) into a 2-dimensional (areal) dataset, coarsening its resolution.

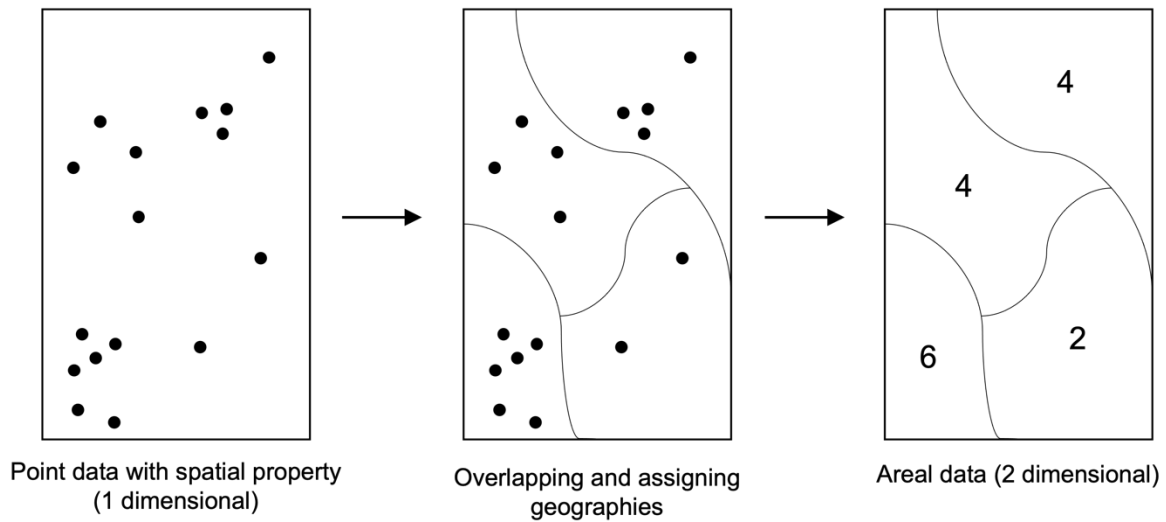


Figure 2. Synthetic data example of the spatial aggregation process.

Such aggregation can be performed at different scales using different types of units. World population density is traditionally calculated and delivered in “number of inhabitants per square kilometre” in this manner (Brunet *et al.*, 1993). Square grids rely on grid reference systems, which represent the earth on a planar surface following certain projections, with each system defining specific locations on earth as points of reference. In the United Kingdom (UK), the Ordnance Survey National Grid reference system (OSGB) is the most common grid reference system independent from latitude and longitude. It is based on *Eastings* and *Northings*, coordinates which correspond to the relative north and east position of a location from the most south-western point of the grid. The Ordnance Survey ‘Guide to Coordinate Systems in Great Britain’ provides more details on the genesis and use of OSGB cells, with comparison to other grid reference systems. The OSGB cells can be divided into smaller squares, with each being given a specific grid digit. Figure 3 demonstrates these subdivisions and the OSGB Eastings and Northings system (Ordnance Survey, 2020).

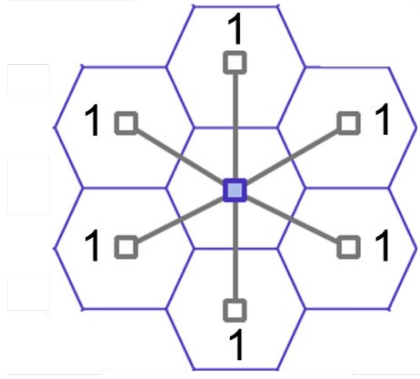
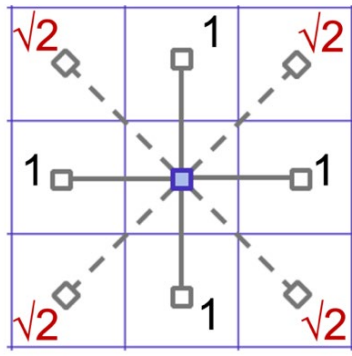


Figure 4. Comparison of the distance between neighbour centroid for square grids and hexagonal tiles. The distance between neighbours is always the same for hexagons, whereas diagonal neighbours are ~ 1.4 ($\sqrt{2}$) units away for square (given a distance of 1 for their adjacent neighbours)

Hexagonal tiles also fit curved surfaces better, making them more appropriate for tiling the globe (Snyder, 1992; Sahr, 2011). They have comparable shapes and sizes across all cities, where different cities might otherwise use different grid reference systems for adapting square tiles (Brodsky and Uber Technologies Inc., 2015; Bondaruk *et al.*, 2019).

These advantages make hexagons particularly appropriate for aggregating datasets with movement (vehicles, individuals, transport flows), which justifies the taxi service Uber's development of a global hexagonal indexing system (named H3) to aggregate their datasets (Brodsky and Uber Technologies Inc., 2015). H3 combines the advantages of a hexagonal global grid system with a hierarchical indexing system, to provide different resolutions of tiles to choose from to perform aggregation (Brodsky, 2018). A user guide and detailed blogpost provide more technical information on the creation of the H3 indexing system, and demonstrates the use of hexagonal tiling for data-aggregation (Brodsky and Uber Technologies Inc., 2015; Brodsky, 2018; Bondaruk *et al.*, 2019).

2.4.2. Impact of aggregation on analytical validity and completeness

Spatial aggregation implies a change of scale from one dimensional (point) data to two-dimensional (areal) data (see Figure 2). Any such shift of scale brings about spatial scale problems related to the integration of data obtained at various scales, fundamental to data-driven geography (Atkinson and Tate, 2000). Atkinson and Tate (2000) highlight that the term *scale* is often interchangeable for a wide variety of meanings. In the following, it refers to both 'amount of detail' (granularity) and spatial extent of a geography (see also Goodchild and Proctor, 1997 for precisions on this dual definition of scale). Furthermore, *Scales of*

measurement refers to the scale at which the data is collected (sometimes referred to as *resolution*), and *scales of spatial variation* refers to changes of scales of any kind. The scale of measurement may not be stable across the whole dataset: some data can be collected at larger or smaller scales, at different resolutions, or in the case of the in-app data, at different accuracy levels. However, the core problem of spatial scale in geo-statistics comes from *changes* in scale from the original dataset (Openshaw, 1979; Mennis, 2019). In itself, a dataset represents an observation of a physical phenomenon: it relies on a ‘support’ or a tool (a chosen output spatial scale) to represent the dataset. Changes in that scale create more uncertainty when it comes to determining how much of analytical conclusions are a result of the actual phenomena or a result of the ‘support’ and subsequent data treatments that come with changing said support (Mennis, 2019).

These issues of scale and zoning can be caused by multiple data-manipulation processes, but one arises especially from the spatial aggregation process: the Modifiable Areal Unit Problem (MAUP). The MAUP was coined by Stan Openshaw in 1979, and states that changes of scale and zone when aggregating point datasets impacts analysis (Openshaw, 1979), particularly through aggregation (Openshaw, 1981, 1984). The MAUP is composed of two distinct effects: the scale effect, and the zoning effect, illustrated by

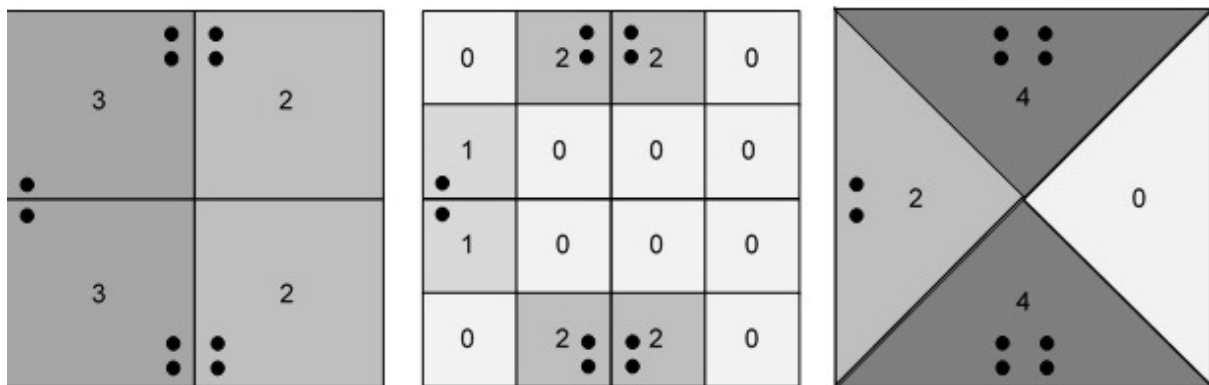


Figure 5. The scale effect describes the variation in results caused by changing the scale of the dataset (making the units of measurement larger or smaller). The zoning effect describes a variation in results which occurs when the scale remains the same (stable number of units), but the boundaries of the units change. In the case of sensitive data, this can happen when the data is first aggregated for disclosure control, but also when it is subsequently reshaped to link with other datasets and geographies. In fact, Atkinson and Tate (2000) highlight that researchers are increasingly required to change the scale of measurement from one to another to allow for comparisons across datasets. There are no clear or direct solutions to these ‘issues of support’,

where changing the tool of the research impacts analysis (Atkinson and Tate, 2000; Mennis, 2019).

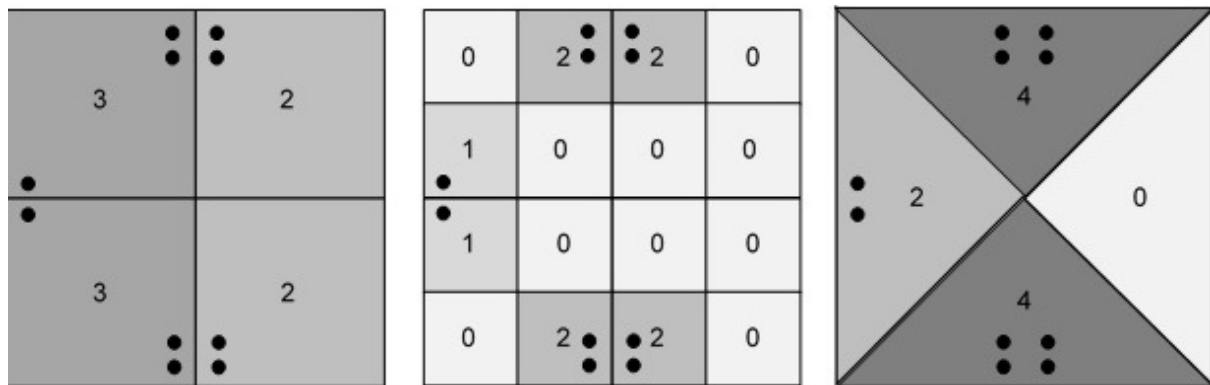


Figure 5. Diagrammatic representation of the impact on aggregating a set of points to area counts, highlighting the scale (left and middle panels) and zoning effects (right panel). Numbers and grayscale indicate the number of points within each spatial unit for that partitioning scheme (darker shade of grey indicating a higher number of points). (Source: Mennis, 2019)

This product of aggregation is not only spatial, but also temporal, where a change of temporal scale (for instance aggregating the data hourly, or monthly) impacts results (Modifiable Temporal Unit Problem, MTUP) (Cöltekin *et al.*, 2011). The MAUP and MTUP's impact on descriptive statistics has been demonstrated over the years for multiple types of location data and is an outstanding challenge for geographers (Openshaw, 1977; Fotheringham *et al.*, 1995; Qi and Wu, 1996; Dusek, 2004; Minot and Baulch, 2005). When trying to conduct mobility analysis, for which the specificities of granular traces and behaviours happening over space and time are valuable, such alterations of the data are especially consequential (Openshaw, 1979; Viegas *et al.*, 2009; Haley, 2017; Biehl *et al.*, 2018).

More precisely, these (often arbitrary) changes of scale and/or zone impact the data's *analytical completeness* and *analytical validity*. Purdam and Elliot (2007) define these two concepts in the context of disclosure control on data utility:

- (1) A reduction of analytical completeness signifies that 'analyses that might have been conducted with un-recoded data cannot be [conducted with the aggregated data]' (Purdam and Elliot, 2007, p. 1102). For instance, a change of geographic scale resulting in the loss of granularity is a reduction of analytical completeness, as one can no longer conduct analysis at the finer scale.

- (2) Analytical validity, on the other hand, is lost if the disclosure control method ‘changes a dataset to the point at which a user reaches a different conclusion from the same analysis’ (Purdam and Elliot, 2007, p. 1102).

They find that disclosure control measures such as aggregating data have a significant impact on both analytical completeness and validity, echoing Openshaw’s emphasis on point data aggregation’s vulnerabilities to MAUP (Openshaw, 1979). The aggregation process not only involves a change of scale, which can impact the outcome of research conducted with the aggregated dataset, but also often relies on scales and zones not properly fitted to the data they seek to represent (Dusek, 2004; Minot and Baulch, 2005).

The MAUP is generally accepted as an unavoidable challenge in geography, rather than something which can be wholly resolved. Nevertheless, advancements in GIS technology and statistical methods may help reduce its impact on spatial disclosure methods. The increasing availability of in-app data generated by mobile phones contributes to a broader trend in geographical research: digital cartography and GIS technologies have transformed the way geographers study and map populations, by enabling more complex spatial statistical analysis and data exploration (Goodchild, 2018). However, few tools exist with which to consider issues of rescaling and zoning when it comes to the safe aggregation of new, large, human-generated location datasets (Lagonigro *et al.*, 2020). One method for reducing this impact consists in finding the units and scales which fit the data most closely, hoping to reduce the impact of the aggregation process on analytical completeness and validity. This can be done through the delineation of specific aggregation units: regionalisation.

2.4.3. Regionalisation

“We need a geography today which helps us to see ourselves, our fellow passengers, and our total environment in a more coherent way than we are presently capable of doing. To me the answer seems to lie in the study of the interwoven distribution of states and events in coherent blocks of space-time – in other words a regional synthesis with a time-depth”

(Hägerstrand, 1975b, p. 27)

2.4.3.1. Defining regionalisation

2.4.3.1.1. Regions: formal or functional?

The term ‘region’ can be broadly categorised into two groups: formal or functional regions. Formal regions are areas which share the same attribute, often representative of a certain culture, administration of political boundary. Countries and States are often an example of formal regions. Functional regions, however, are defined as “areas organised by the horizontal functional relations (flows, interactions) that are maximised within a region and minimised across its borders so that the principles of internal cohesiveness and external separation regarding spatial interactions are met” (Farmer and Fotheringham, 2011, p. 2732). In this sense, functional regions serve a specific (often analytical) purpose, and reflect selected spatial behaviours in a geographic space (Smart, 1974; Hales, 2001). An example of functional regions often put forward in regionalisation literature are the labour-market areas (LLMAs), designed to analyse daily travel-to-work flows (Smart, 1974; Casado Díaz and Coombes, 2011). The methodology and principles behind the delineation of the LLMAs will be further explored in Chapter 4, where they are applied to the in-app data. The rest of this work thus focuses on functional regions, as it aims to explore the benefits of bespoke functional regions for efficiently aggregating location data.

2.4.3.1.2. Regionalisation definition

Regionalisation describes the process of drawing region boundaries. In the case of functional regions, the delineation of regions is often informed by the phenomenon the regions seek to represent, such as in the case of the LLMAs. However, administrative boundaries can also be made with data in mind, such as the French *Régions*, built by combining *départements* together based on shared characteristics (Brennetot and Ruffray, 2015). Serge Antoine, the geographer tasked to generate these regions in the 1960s used a wide variety of criteria, including telephone communications between *départments* to determine which ones were to be combined based on their connectivity and similarities (Antoine and Weill, 1968; Piastra, 2015).

Regionalisation was also proposed as a way to prevent disclosure control in the context of the differencing problem (Duke-Williams and Rees, 1998). The differencing problem identified that releasing a dataset at different geographies may reveal sensitive information by subtraction, even if the original two geographies were deemed safe. Duke-Williams and Rees (1998) demonstrated that publishing census statistics for different zones close in size both impact the

data and the safety of the output unless the zones are carefully designed to minimise both issues. They also corroborated earlier discussions concerning changes of scale and zones for linkage purposes and comparative analysis, stating that this should also be considered in the zone design. Regionalising datasets thus could not only help highlight the data's behaviour or preserve analytical completeness and validity of the outputs created, but also minimise potential disclosure risks from the 'differencing problem'.

2.4.3.2. Hierarchical or rules-based methodologies

The making of data-driven functional regions can be broadly classified into two main approaches: hierarchical methods or rules-based methods, both of which consist in applying one or multiple rules to the dataset or geography to help unearth underlying patterns. Hierarchical methods work with the same rule 'from start to finish', whereas rules-based methods apply different rules at different stages (Casado Díaz and Coombes, 2011). Usually, hierarchical methods group areas step-by-step, aiming to increase the area's statistics, such as population size. A hierarchical algorithm will start with a set number of regions, and either combine or subdivide them to follow the initial rule. Casado Díaz and Coombes (2011) provide a detailed review of international attempts at both rules-based and hierarchical examples applied to the creation of LLMAAs across different countries. Similarly to the administrative *Régions*, an example of a hierarchical method are the French *zones d'emploi* (equivalent to LLMAAs), which were made by combining together *communes* (the smallest French administrative territorial division) until the total resident population reached a count of 25,000 inhabitants (Eurostat and Coombes, 1992; INSEE, 2020). The first case study below details a hierarchical method, the quadtree algorithm.

With rules-based methods, the number of regions is not set at the beginning of the analysis: the algorithms tend to re-balance the overall regions based on their interactions with other regions, and some rules need to be weighted to control for their impact on the final result. In this sense, rules-based methods rely on more decision making and potentially arbitrarily-set thresholds than hierarchical methods typically do. However, Casado Díaz and Coombes (2011) established in their review that despite this, rules-based methods performed better than hierarchical methods in delineating the unique details of LLMAAs in both space and time. The second case study details the rules-based method behind the making of the England and Wales Output Areas.

2.4.3.2.1. Case-study of hierarchical approach: Quadtree algorithms

A recent example of a hierarchical approach to region delineation is the silo regionalisation algorithm developed by Molloy and Moeckel (2017). Their work stems from adapted rasterization methods used in transport planning, and is based on iteratively dividing square grids which respect juridical boundaries and topology. They used an adapted quadtree algorithm, which subdivides units until a threshold limit is reached (technical descriptions of quadtree algorithms will be further explored in Chapter 4, Section 4.3.). This creates smaller units where there is enough data to subdivide without risking disclosure. Figure 6 illustrates an output of this methodology.

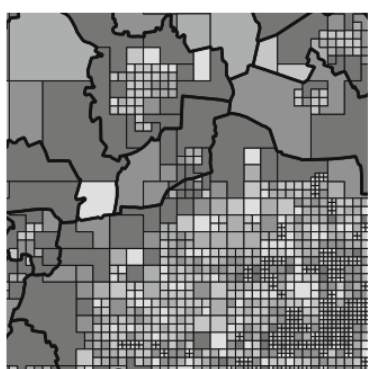


Figure 6. Snapshot of Molloy and Moeckel's 2017 silo regionalisation method.

However, despite the flexibility of the method, multiple papers on zoning strategies have found rasterization-based regionalisation to be lacking. Molloy and Moeckel note four main drawbacks of this zoning method. (1) raster cells may overlap multiple jurisdictions, adding complexity by creating another non-nested set of zones. This complicates linkage with socioeconomic data. (2) Populations are distributed by the area percentage of overlapping zones, assuming that socioeconomic data is distributed equally throughout zones. (3) Choosing the resolution of cells to find the desired one for an analysis relies on trial and error. (4) When zones exceed threshold values, they are split into rasters of equal sizes, not considering the possibility that most of the population resides in one of those cells rather than equally split between them. These shortcomings are further assessed in Chapter 4 (Section 4.3.1), where a quadtree hierarchical algorithm will be applied to a sample of in-app data.

2.4.3.2.2. Case-study of rules-based approach: Output Areas

Rules-based methods, though more complex to develop, have proven to be more fitting for functional regionalisation (Casado Díaz and Coombes, 2011). The England and Wales census Output Areas (OA) created by the ONS are an example of rules-based spatial units optimised for aggregating and disseminating a specific dataset. OAs provide a basis for understanding the rules-based process of creating bespoke regions for aggregation purposes, and highlight the decision making required in the definition of regionalisation rules. Where hierarchical methods may help quickly distinguish clusters, rules-based methods can accommodate for various future use of the regions by optimising for additional factors (Halás *et al.*, 2015).

OAs were designed to be functional geographic units of roughly similar population sizes (minimum 100 residents) and compact in shape (Martin, 2010; Office for National Statistics, 2016b). Traditionally, census geographies were drawn around enumeration districts (ED), which were delineated by the collection areas of census workers: they did not describe underlying residential population, but rather the zone each enumerator was responsible for (Morphet, 1993; Martin, 2010). With the shift to OAs, the aim was to create regions with internal social homogeneity, maximising heterogeneity between regions (intra-unit homogeneity and extra-unit heterogeneity), and utilise the geographical components in the data to make outputs more representative of underlying populations. This was achieved with a methodology that combines small atomic units (postcode building blocks) together based on specific conditions and rules (Martin, 2000; Cockings *et al.*, 2013): shape compactness, population size, but also underlying terrain in an aim to respect environmental borders such as rivers and roads (Martin, 2010). Later, the regions created were further grouped following similar conditions of homogeneity to create larger regions, first Lower Super Output Areas (LSOAs) (between 1,000 and 3,000 residents), and larger Middle Super Output Areas (MSOAs) (average populations of 7,800) (Office for National Statistics, 2016b). This allowed for more linkage opportunities to other scales and versatility in their analytical uses (Cockings *et al.*, 2011; Walford and Hayles, 2012).

As touched upon earlier, rules-based methods can create different outputs depending on the order of priority of each rule. According to David Martin, one such conflict for census geographies was trying to keep the OAs compact and achieve the best results for population size or homogeneity criteria. The OAs are not the same shapes or sizes depending on whether or not the priority is set on the OA population count or the homogeneity of that population

(Martin, 2000). However, these changes have also allowed for OAs to be updated between the 2001 and 2011 census, and again for the 2021 census release.

OAs, LSOAs and MSOAs are some of the most used geographies for UK spatial analysis, thanks to their versatility, and best fit to census data (Martin, 2010). However, aggregating in-app data to OAs might not help preserve the data's granularity or address the issues pertaining to MAUP discussed above, as the OAs were designed to fit an entirely different dataset. Thus, in-app data could benefit from a similar regionalisation method: a rules-based approach which would consider data disclosure practices, underlying terrain, and homogeneity metrics, and would allow for bespoke, granular, and timely analysis using a spatial aggregate.

2.4.3.3. Regionalisation of in-app mobile phone data: what to consider

2.4.3.3.1. Summary: issues with in-app data

The issues brought about by MAUP, and loss of analytical validity and completeness cannot be completely removed so long as aggregation is necessary for disclosure control. However, their effects on analysis can be accounted for and minimised: aggregation scales and zoning processes must be well understood and specifically developed to best fit the datasets studied (Martin, 2000, 2010; Cockings and Martin, 2005; Kishore *et al.*, 2020). This would help the research's reproducibility, on condition that the aggregation process is transparent, and aggregated products available for independent verification. However, in-app location datasets remain difficult to access, and are often aggregated by the organisations generating them. This process regularly happens in a black box, and to scales and zones not decided upon by the researchers requesting access to the data, making it hard to control for the MAUP's impact. Most in-app datasets provided to researchers are aggregated to arbitrary grids. Meta Inc location data, for example, is aggregated to standard tiles rather than bespoke regions or administrative boundaries (Maas *et al.*, 2019). The widespread use of arbitrary grid also makes it unlikely that the output products are linkable to existing geographies or made to closely fit the original datasets, which often result in data loss. This in turn makes them unlikely to be informative for mobility analysis.

However, the access to the original disclosive datasets is subject to barriers, both financial and technical. Secure remote access is often mandatory, with researchers having to use the datasets within secure environments, receive specific training and depend on control checks. This all adds to the (often high) initial cost of purchasing the data. It is thus usually not sustainable for

researchers to request or consider using the original datapoints where an aggregated product might reduce these costs.

One could propose an aggregated product at the scale of the UK's most used geographies, the OAs. This would likely be of interest to researchers working with geodemographic information provided at the OA scale. However, as highlighted by Helbich et al. (2021), using residential regions like OAs for non-residential data, such as in-app data, may prove irrelevant in a variety of studies interested in non-residential interactions outside homes (Helbich, *et al.*, 2021). To address this Coombes et al. developed *daily urban systems* (the ancestor of LLMAs) as early as 1982, highlighting the importance of creating a geography which best represent daily activity as opposed to residential information (Coombes *et al.*, 1982). They argue that, when it comes to non-residential information, OAs are not made with varying population sizes in mind, losing their trademark homogeneity and stable populations. In-app mobile phone datasets would benefit from statistically neutral regions, especially in an aim to reduce MAUP impacts and provide accurate activity spaces regardless of geodemographic characteristics (Kishore *et al.*, 2020).

Finally, mobile phone datasets are often considered as mobility datasets. The key studies making use of these datasets (a number of them presented above in Section 2.2.2.2), are those which rely the most on new forms of data for granular and timely insights, often related in some way to time geography (investigating where people go, when and why, and the consequences, whether social or epidemiological etc.). Outputs made from these datasets may be in the form of static units, but they must still capture some of this dynamic aspect of the underlying dataset they seek to regionalise.

2.4.3.3.2. Opportunities in regionalisation

By benefiting from an access to an original individual-level in-app dataset, this project presents multiple opportunities. Firstly, by retracing the aggregation process from the location points to the output product, it is possible to assess the impact of the scale and zoning effects on these new types of data. Secondly, there are two main outputs possible from this regionalisation exercise: (1) An appropriate aggregated product for dissemination and wider use of in-app data in research and (2) a flexible and reusable methodology for regionalising and aggregating other sensitive point datasets.

Regionalisation has traditionally been a manual process, with new zones created very rarely due to the involvement required. It remains an often proprietary process for individual projects. In their literature review on the topic of zone system designs, Molloy and Moeckel highlight that the “significant variety of approaches is a tribute to the complexity and importance of the process” (Molloy and Moeckel, 2017).

Nevertheless, recent technological improvements have provided new opportunities for the development of regionalisation processes and zone designs. Multiple studies, especially in the field of transport and mobility, have investigated MAUP in spatial analysis and spatial modelling (Viegas *et al.*, 2009; Lovelace *et al.*, 2014). Regionalisation has thus evolved to become more project-specific in order to reduce the case-to-case impact of MAUP on analysis, especially when using new forms of big data (Molloy and Moeckel, 2017). Examples of recent regionalisation algorithms include ordnance survey grids, dasymetric mapping, areal weighting, or polygon-based regionalisation (see Sahr *et al.*, 2003; Tiede and Strobl, 2006; Hallisey *et al.*, 2017; Järv *et al.*, 2017; Bustos *et al.*, 2020).

Data-driven regionalisation solutions are thus put forward as a viable way to mitigate MAUP, especially as the technology evolves to make these techniques more accessible. However, a commonly shared and reproducible system of zone creation would benefit researchers in facilitating the technical aspects of regionalisation. There is an opportunity in developing a regionalisation methodology which remains data-driven and project specific (thus flexible and malleable to the analysis conducted), but relies on one same transparent algorithm. As described by Halás *et al.* (2015), methods which do not require complex mathematical adjustments allow researchers to decide, according to their field knowledge, which parameters and zoning approaches to apply. By developing a regionalisation method for in-app mobile phone data, this project also aims to propose transferable methods which could be applied to other sensitive mobility datasets.

Research Objectives

- (1) Assess an individual-level in-app dataset and explore the impact of MAUP on its aggregation. (Chapters 3 and 4)
 - a. Describe the data's characteristics, with mobility studies in mind.
 - b. Assess the data quality.
 - c. Assess industry standard aggregates against the individual-level data.
 - d. Visualise the effect of MAUP by comparing different zonings and scales.

- (2) Regionalise the dataset to preserve privacy, granularity and analytical completeness and validity. Propose a transparent, tractable and versatile methodology for aggregating large location datasets which could be applied to other comparable datasets and urban areas. (Chapters 4 and 5)
 - a. Define the essential criteria that an appropriate functional region should meet for this data and purpose.
 - b. Test existing reproducible hierarchical methodologies and assess them.
 - c. Develop a rules-based method following the above-defined criteria.
 - d. Assess bespoke regionalisation method against traditional aggregation scales and zones.

- (3) Assess the impact of MTUP on the regionalisation method's volatility, confirming whether it remains relevant through temporal changes and for times of low data. Demonstrate the use of the regionalisation method as a research tool to inform on and visualise key MTUP effects in the context of aggregation (Chap 6).

3. Data and Preliminary Analysis

This chapter presents the in-app mobile phone dataset that underpins the substantive analysis conducted throughout this thesis. The metadata description of the raw in-app dataset, along with key terminology, is outlined before detailing the initial evaluation and exploration (distributions, trends and coverage) of the data. Secondly, the data cleaning process is presented, composed of the data digestion stage followed by data aggregation. As this in-app dataset is sensitive, a key objective of the data processing is to create safe aggregates for dissemination outside of secure environments. These notions are introduced in more detail and the cleaning and aggregation processes are applied to the making of an industry standard aggregated data subproduct published on the CDRC datastore. Finally, using the resulting data subproduct, the last section further investigates the data’s representativeness and bias within Greater London and assesses the subproduct’s quality when conducting analysis. At the end of this chapter, a clear understanding of the in-app dataset and the making of its aggregated subproduct should be attained, with details on their benefits and shortcomings. The final assessment compares the sensitive raw data with the more accessible and non-disclosive aggregated subproduct: it highlights the ways in which common aggregation techniques negatively impact analysis to motivate this thesis’ search for bespoke data aggregation techniques.

3.1. Data overview

3.1.1. Metadata and data description

Table 1. Metadata table - Initial summary of descriptive statistics of the in-app dataset.

Field	Value
Provider:	Huq Industries
File format:	.json.gz (newline delimited json)
Size:	778 GB
Number of variables:	35
Number of items:	1704 (1554 unique files)
Period covered:	05-07-2016 to 05-10-2020
Scale:	National (Great Britain)
Number of records:	6925977175
Average number of records:	4456871

The dataset, supplied by Huq Industries, comprises mobile phone GPS (Global Positioning System) locations collected by third party applications (apps) installed on the users' devices (See Figure 1 for full glossary). When an app requiring a location feature is being used, location, timestamp, phone ID and app ID are stored in a database. The apps seek the users' consent to store this information. The original dataset is composed of 35 variables ranging from phone service provider to reverse geotagged location characteristics and more. This was provided in 1554 daily files of zipped newline delimited json format (.json.gz). The research conducted in this thesis focuses on the impact of aggregation on the dataset's analytical completeness and validity, thus a list and description of the variables relevant to this study are available in Table 2. Other variables were filtered out to respect the data minimisation principle (personal data shall be limited to what is necessary for the purpose of a stated analysis, Article 5(1)(c)[GDPR, 2018]). The 778 GB of data include entries for each record registered in Great Britain (GB) from 2016-07-05 to 2020-10-05, with irregular time series. The data samples provided comprise almost 7 billion records (6,925, 977,000). The timestamps are stored in two columns with the date in "YYYY-MM-DD" format and the time in "HH:MM:SS UTC". Location is divided into two columns of latitude and longitude respectively.

Table 2. List and description of variables in the raw in-app dataset (alphabetical order). Variables in bold are the key ones filtered for the making of data aggregates in Section 3.2.3. Other variables presented here are used for assessing the dataset's quality and accuracy, and provide overall descriptive statistics presented throughout this chapter.

Name	Data type	Note(s)
app_id_hash	string	A hashed id that is consistent for events created from any given app
device_iid_hash	string	An anonymous, consistent, hashed identifier for the user device
human_readable_os	string	The converted human readable version of the device operating system
impression_acc	float	Device GPS accuracy at the time of measurement
impression_lat	float	The latitude of location as provided by the device at the time of event creation
impression_lng	float	The longitude of location as provided by the device at the time of event creation
timestamp	timestamp	The local datetime of the event, calculated from the reported event time, the location, and the server time.

Thanks to the device id and app id variables, we can see how many impressions are collected by apps on average and how active specific devices may be compared to others. Figure 1 provides a glossary of the key terms used throughout the thesis to describe the different components of the dataset, along with a mock data frame relating each term to their role in the data samples provided.

Term	Definition
Record, Impression, Event	A single location point recorded in the dataset.
App	Mobile application which, when used, generates and collects the impressions.
User	Individual in the dataset (person using an app on a mobile device)
Unique devices	Mobile devices uniquely identified by the hash ids. Used as a proxy for users.
Activity	Throughout the work, ‘activity’ refers to the count of unique devices over a specific time and place (see Section 3.2.3.2).
Raw data	The in-app location data prior to any processing: the records, unfiltered, as they were provided by the data provider, within a secure environment (see Section 3.2.1)
Processed data, Subproduct	Aggregated output dataset derived from the raw data, post cleaning and processing.

device_iid_hash*	impression_lat	impression_lng	timestamp	app_id_hash
D_1	12.345	-123.4	YY-MM-DD	APP_1
D_2	67.89	-678	YY-MM-DD	APP_2
D_2	12.345	-123.4	YY-MM-DD	APP_1

Each line is a record

Apps collect records for multiple devices

***Devices are used as a proxy for counting users**

Figure 1. Glossary of terms and example dataset illustrating each component’s role in the data collection

Each line of the raw dataset is a record (impression), with the columns containing its descriptive information (location of the record, time, app and device which generated it etc.). A single device may generate multiple impressions, especially if different apps are used on the same phone. Throughout the thesis, we consider the devices as proxies for users, assuming one device per user. There is, on average, around 74k unique devices registered each day, though this number varies greatly throughout the 4-year period (this is described further in Section 3.1.3).

3.1.2. Filtering out low accuracy impressions

Investigating the impressions' accuracies allows us to understand and describe the collection process and technologies involved in the creation of the dataset. It also aids in filtering out low accuracy impressions, which would provide imprecise information (Diggelen, 2009; Wang *et al.*, 2019). For this dataset, the *impression_acc* variable describes the recorded GPS location's accuracy in metres. For example, an *impression_acc* of 200 metres indicates that the specific location recorded is within a 200m radius of the latitude and longitude recorded. The smaller the *impression_acc*, the more accurate the location recorded is. Negative accuracies (<0) were filtered out of the dataset as they indicate a device's hardware malfunction, given that accuracy values are supposed to only be positive (Diggelen, 2009). The sum of impressions per accuracy recorded is calculated, to obtain a distribution of data accuracies as plotted in Figure 2. The red line highlights 150 metres.

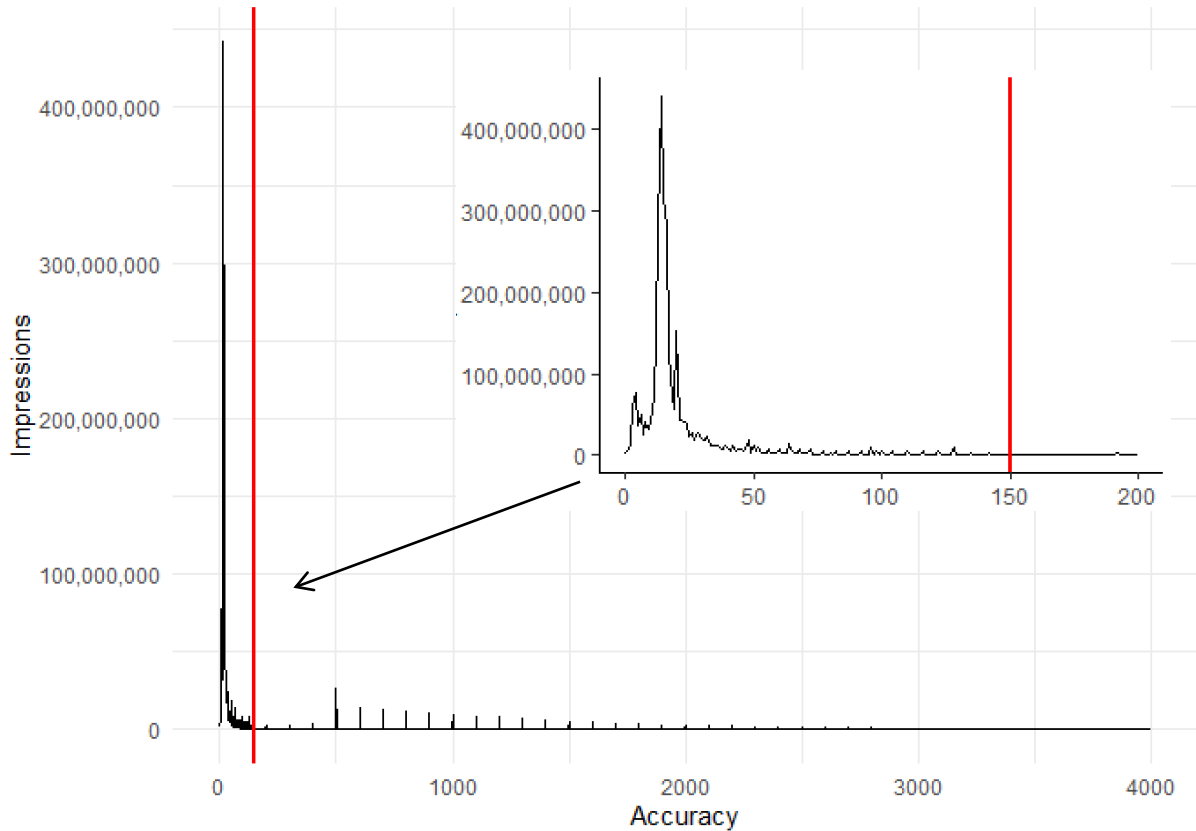


Figure 2. Location accuracy measurements of the in-app data impressions (sum of impressions recorded at each accuracy). Approximately 90% of all impressions in the raw dataset have an accuracy of 150m or less (red line)

The regularly spaced peaks on the x axis seen in Figure 2, from roughly 500m onwards, show that this dataset is generated through both GPS and network positioning system (NPS) technologies. NPS makes use of cell tower network and other surrounding transmitters (e.g. Wi-Fi) to triangulate the device's location (See Diggelen, 2009 and Vallina-Rodriguez *et al.*, 2013, for more thorough descriptions and assessments of both GPS and NPS systems). Typically, NPS-collected locations have lower precision and accuracy than GPS, which explains the bins visible on Figure 2 (Diggelen, 2009; Wang *et al.*, 2019). 87% of the dataset's locations are accurate by 100 metres or less, and 89.6% of points are equal to or below 150m accuracies, left of the red line marker of Figure 2. This number only increases by 0.4% between 150 and 200m (90% of the data points equal to or below 200m). These figures correspond to typical accuracy measures of in-app data: Wang *et al.* (2019) performed a similar assessment of their app-based data, finding that 'about 15% of the observations have location accuracy larger than 100 meters', and plotting similar peaks highlighting the presence of both GPS and NPS sensors (See Figure 3, extracted from their analysis, presenting similar results to Figure 2 above).

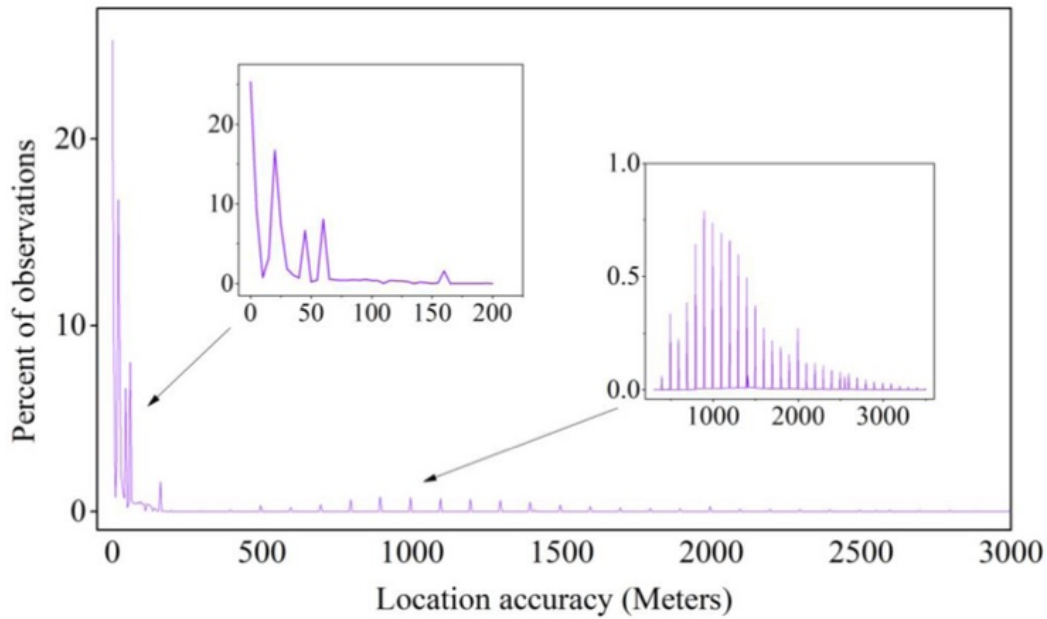


Figure 3. Location accuracy measurements from another in-app dataset (proportion of observation per accuracy)(Wang et al., 2019).

Following this assessment, the dataset is filtered to retain only the impressions with accuracy measurements of 150m or below. There is no specific cut off recommended, but other app-based studies filter location points with a maximum accuracy ranging between 100 and 200m for studying individual mobility patterns (which require high accuracy)(Wang *et al.*, 2019). Accuracy filtering should typically be informed by both the data and the analysis conducted with it. For this in-app dataset, 150m is the point at which the cumulative sum stabilises, with almost 90% of the data having accuracies below this value. The analysis in this chapter is thus conducted on the filtered dataset containing the large majority of impressions with accuracy levels of 150m or less.

3.1.3. Temporal distribution

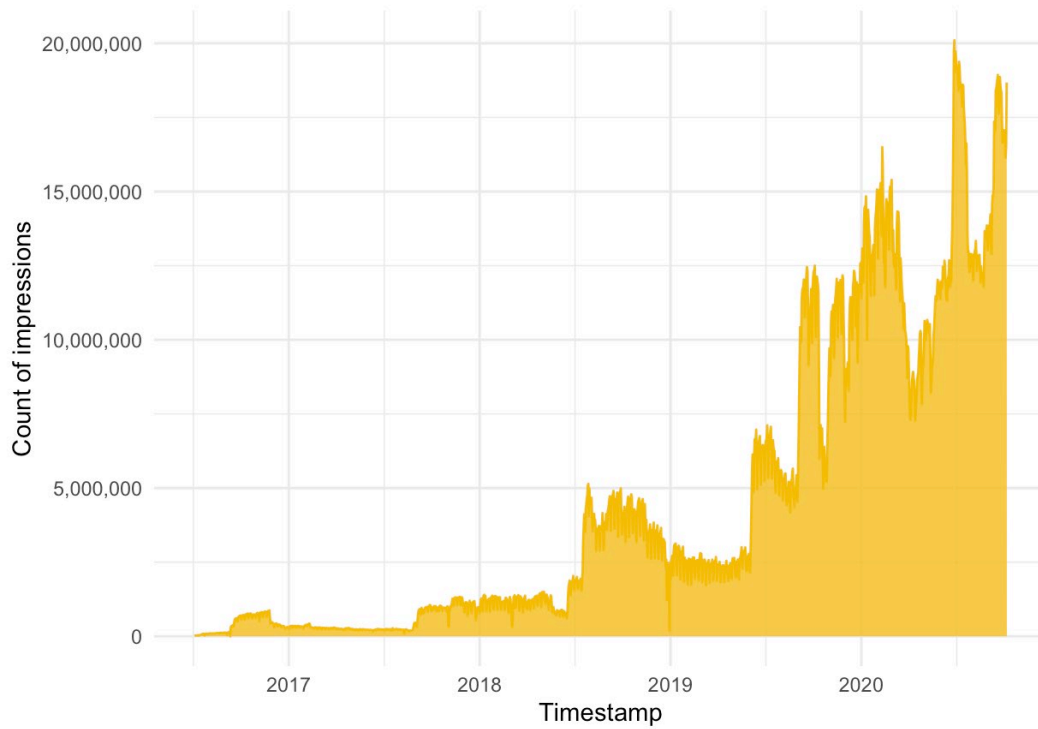


Figure 4. Sum of daily impressions across the in-app dataset.

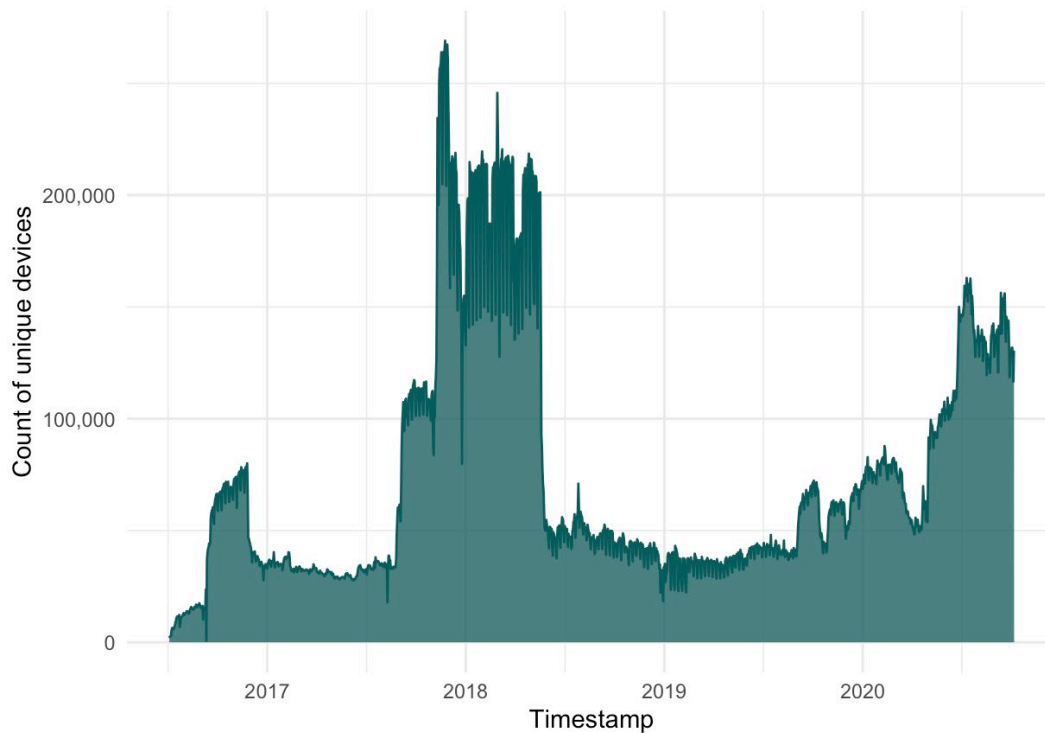


Figure 5. Distribution of unique devices recorded daily across the original in-app dataset. A large number of devices are onboarded before 2018, with an equally steep drop mid-2018, and a steady increase of devices through until late 2020.

Between 2016 and 2020, the number of impressions recorded each day in the dataset steadily increased, going from a couple hundred impressions at the start to 20 million by the end of the data period (Figure 4). However, the number of unique devices is not proportional to the number of impressions throughout the 4 years (Figure 5). This indicates that the number of records generated by each device is not stable throughout the data period (Figure 6). Comparing Figure 4 and Figure 5, we notice that, although the number of devices peaks in 2018, the number of impressions increases throughout the data period and does not have a specific peak in 2018. In fact, the number of impressions per device increases significantly after 2018 (Figure 6) – this indicates that fewer users became, over time, responsible for more impressions in the dataset (fewer devices but more impression per device).

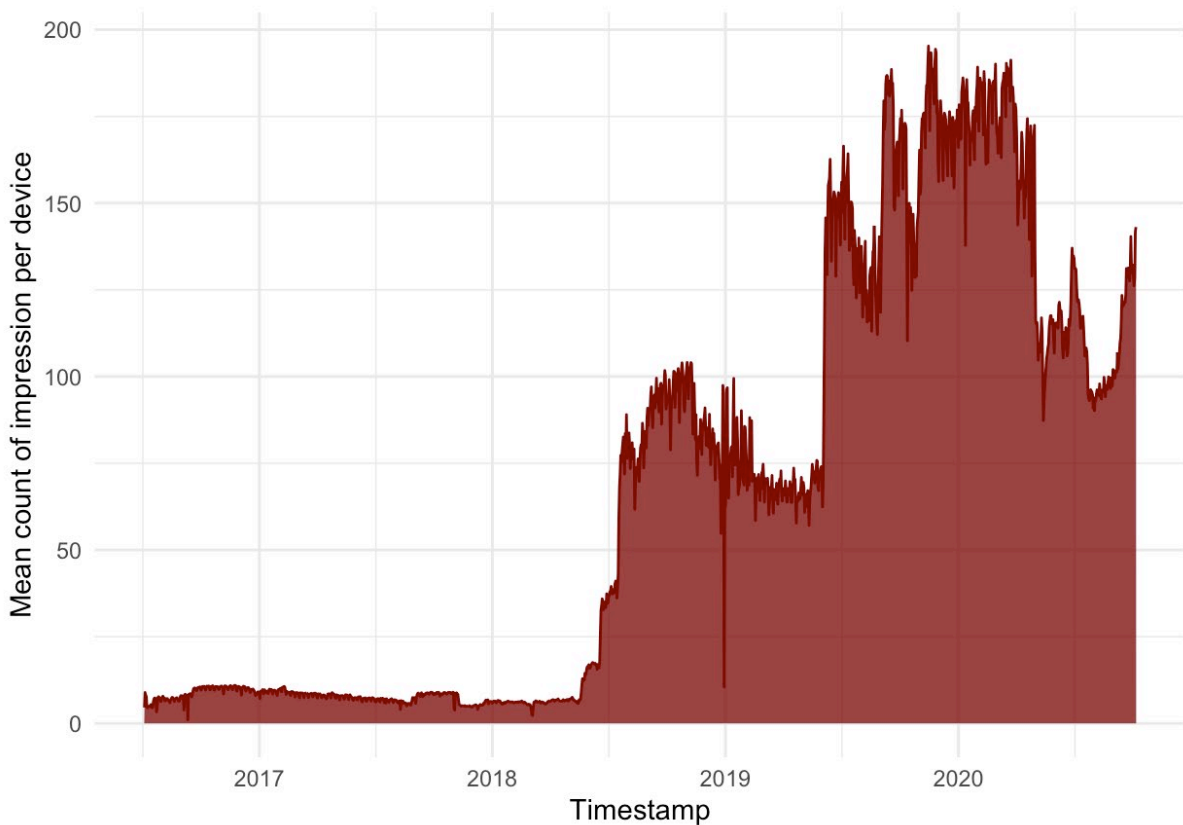


Figure 6. Average number of impressions per device across the in-app dataset. This number increases from less than 10 in 2016-2018, to close to 200 impressions per device in 2020.

These distributions imply that the increasing count of impressions is neither due to an uptake of additional devices generating a similar number of records, nor due to a steady increase of impressions per device: rather the number of impressions per devices varies greatly over the data period. This could be explained by operating systems regulations or app onboardings impacting the impression counts. Mobile phone operating systems (iOS and Android) are

sampled unevenly throughout the data period: the drop in record visible in mid 2018 could be attributed to a change in Apple’s privacy regulations on third party data resulting in the removal of almost all iOS devices from the dataset from this date onwards (Apple, 2022). Data is also collected from a varying sample of apps, the identities of which are commercially sensitive. Apps may be added to or deleted from the dataset over the sample period (which is discussed further in Section 3.1.4). This, along with disparate coverage and mobile phone uptake, results in general inconsistencies in impression counts per devices over the period covered.

Impression counts in the dataset also vary monthly and hourly. The following results break down these temporal distributions. Figure 7 plots the mean impression per month across the 4-year period, and Figure 8 displays the mean per hour. The resulting values are consistent with usual population periodicity within urban centres (peaks at rush hours for instance), towards which the dataset is biased (See 3.3.2.1) (Sevtsuk and Ratti, 2010; Silva *et al.*, 2017; Transport For London, 2020).

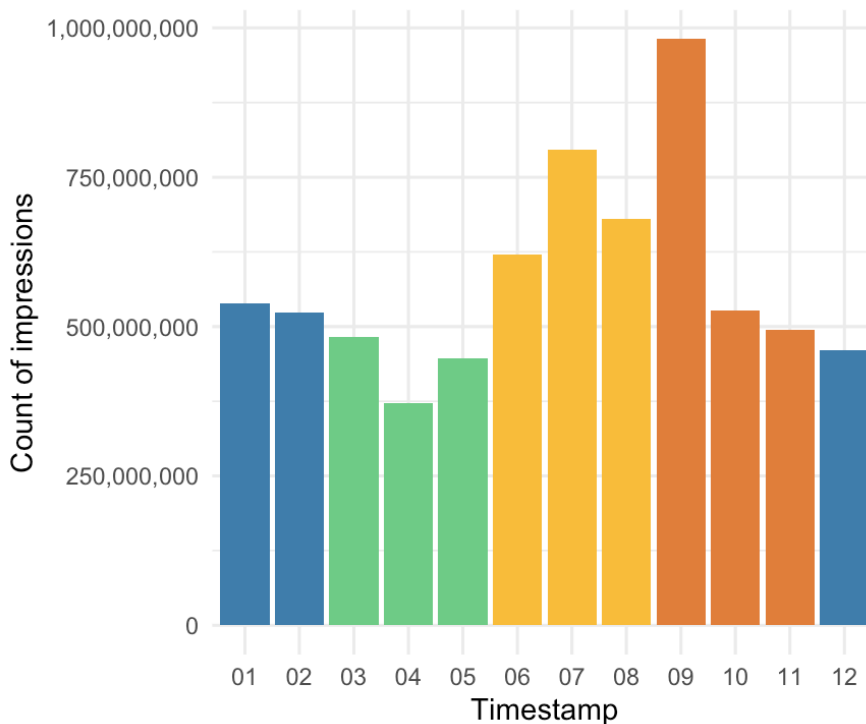


Figure 7. Mean impressions per month, coloured by season (blue winter, green spring, summer yellow and fall orange). September records the most impressions, with April being the least active month on average.

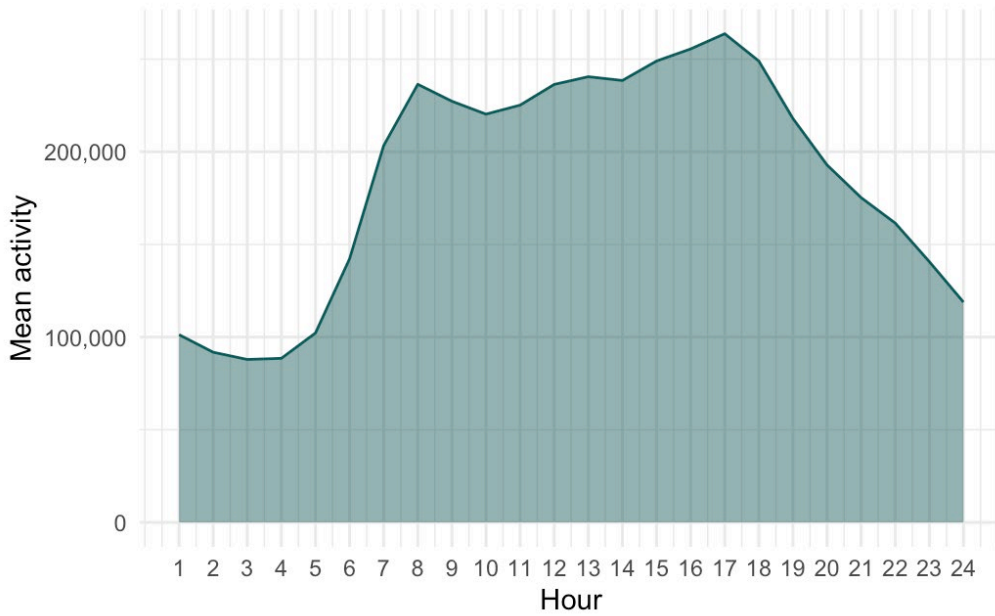


Figure 8. Mean number of impressions per hour throughout the in-app dataset. 5pm (17) is the hour that records the most activity, with a dip between 1-5am. 8am is the first abrupt peak of activity of the day on average.

3.1.4. App statistics

Some apps record user locations to propose location-based services, perform geo-tagging or provide targeted marketing (Vallina-Rodriguez *et al.*, 2013; Silva *et al.*, 2017). This in-app dataset is dependent on the partnered mobile apps which collect the impressions and share them with the data provider. There are 2376 apps in total across the 4-year period of the dataset. However, this number is inconsistent, with apps being introduced and removed from the dataset frequently. Figure 9 plots the sum of apps present in the dataset per year. There are on average 1240 apps per year, though Figure 9 shows this number varies greatly year by year.

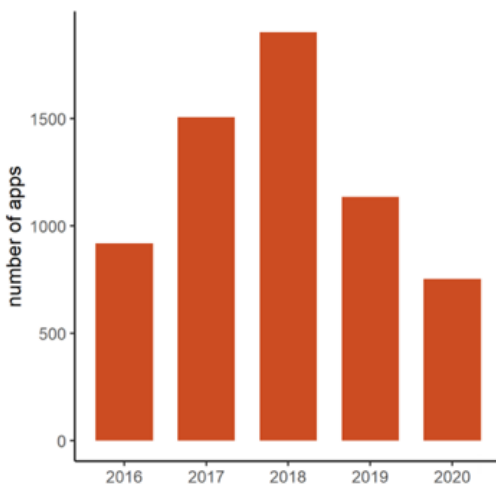


Figure 9. Sum of apps in the dataset per year

Only a minority of apps produce a majority of the dataset. Figure 10 displays the cumulative sum of the largest apps' contributions to the dataset, illustrating that the top 20 apps account for over 95% of the dataset. More specifically, the top 5 apps alone account for 65% of all impressions, and the top 10 generate 84%. Thus, despite a portfolio of over 2000 apps, a large majority of the in-app data is generated by the 10 to 20 most productive apps. By plotting a couple of device trajectories from these large apps, it can be inferred that these are the types of apps which collect data passively and record impressions at regular time intervals (such as traffic apps) (Wang and Chen, 2018). 300 apps record less than 10 impressions across the full 4 years, amongst which 91 only record a single impression. Keeping the top 20 apps thus allows to filter out the unique and sporadic impressions whose behaviours are difficult to explain without knowledge of the specific apps collecting them. This also helps ensure the dataset is composed of only regularly updated locations for a more stable number of individuals.

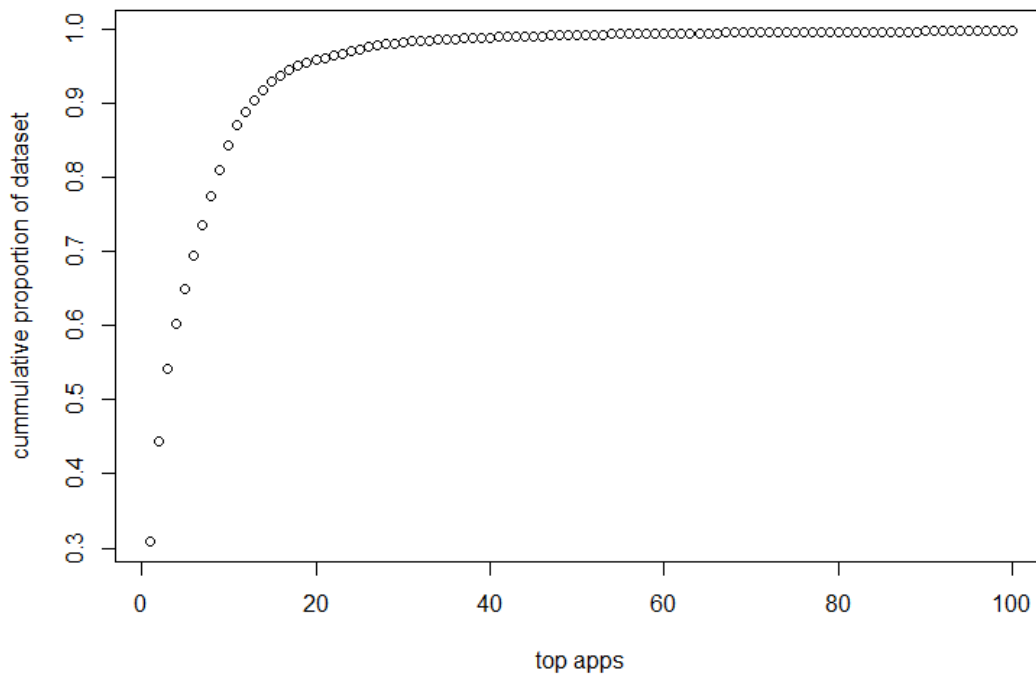


Figure 10. Cumulative sum of app contribution to the dataset. The 10 largest apps contribute towards 85% of the dataset, and the 20 largest apps 95%.

3.2. Data digestion processes

To conduct and present further analysis using the in-app dataset, the raw data must be processed to prevent commercial and individual disclosure. Typically, this involves data reformatting and filtering steps to facilitate further work with large volumes of data, and some form of spatial aggregation for disclosure control and data presentation. In this section, the methodology employed in cleaning and aggregating the in-app dataset is presented. It is first contextualised within the secure laboratory from which the data is safely accessed prior to aggregation. Then, the aggregation process followed for creating counts of events per geographic area is described step by step. The cleaning and aggregation methodologies proposed here are later applied to the creation of a typical data aggregate which aligns with the best-practice standards of the CDRC, to further describe the dataset and assess its representativeness at varying scales. These data cleaning and aggregation steps are repeated throughout the thesis to create aggregates of the dataset at various stages, and support the core methods of this research.

3.2.1. Secure access to sensitive data: UCL Data Safe Haven

Due to the disclosure risks carried (both individual and commercial), this sensitive in-app data was accessed through the secure services of the Consumer Data Research Centre (CDRC), specifically via the UCL Data Safe Haven (DSH) secure facility. The DSH is a remote secure service, certified to the ISO27001 information security standard. Initial access to sensitive data in the DSH involves thorough screenings and training to ensure that only qualified researchers can access the dataset, as outlined in the CDRC User Guide (Version 7.0) (Appendix 1). After having been granted access to the secure environment, researchers must get approval for their proposed data uses, and all analyses must be conducted within the confines of the secure laboratory. To extract data from the DSH portal, it must first meet several statistical disclosure controls, such as aggregating data to larger geographical areas, suppressing sensitive areas, ensuring percentages do not reveal sensitive units, and adhering to a minimum count threshold of no less than 10 for data involving counts, as detailed in the user guide. The statistical disclosure controls performed are verified by two appointed CDRC data scientists, who carry output checks to ensure the processed datasets are non-disclosive and adhere to the principle of data minimisation. If the data output is to be published in any form (including presentations), it must also be approved by the data provider. Figure 11 provides an overview of these

processes, from data access to approved data output, as described by Lloyd in their UCL CDRC thesis which also followed similar disclosure control procedures (Lloyd, 2018).

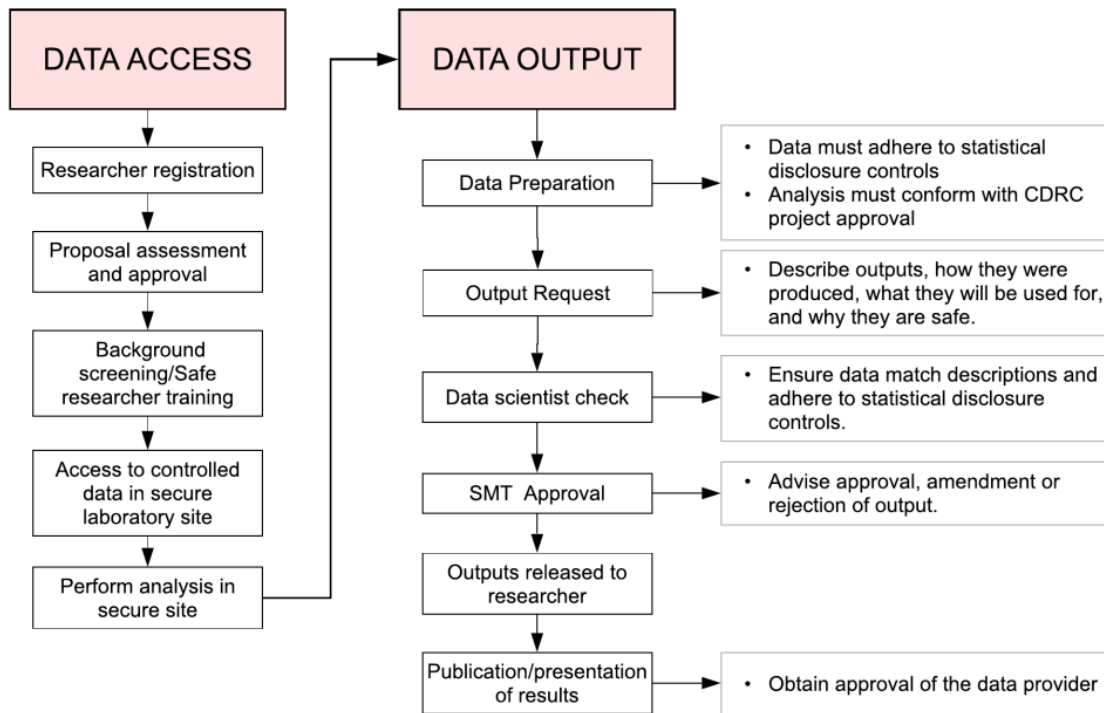


Figure 11. CDRC procedures to access, analyse, output and present controlled datasets, such as the in-app dataset. (Lloyd, 2018)

Certain processing choices made throughout the thesis are also explained by the necessity to process the data through the DSH. Data is stored, processed and managed within the security of the remote system, which restricts the software or programming packages available for data processing. All coding packages used for this project were approved before installation on the DSH, and are listed in Appendix 2. The computational resources are also shared across DSH users, limiting the size and length of running computations. With a dataset comprising billions of records and 1554 individual files, a significant hurdle of using the in-app data was overcoming size restrictions and creating filtered data samples to avoid running analysis on the entirety of the dataset within the DSH. An increase in the allocated RAM within the DSH was requested for this project (doubling from 32 to 64GB of RAM, and adding 4 CPUs) to allow for the cleaning and aggregation processes described below.

Due to the data access and output procedures, any data presented as part of this thesis has been restricted to comply with both statistical and commercial disclosure controls outlined above. The following sections describe the industry standard methodology for spatially aggregating a

sensitive dataset such as the in-app dataset to adhere to these controls. Creating safe aggregates of such datasets allow for dissemination outside of the secure laboratories, thus allowing both presentation of results and granting access to samples of the data to a larger number of researchers. Aggregated subproducts are a way to offer glimpses of data otherwise only accessible through limited release models (de Montjoye *et al.*, 2018).

3.2.2. Cleaning

The primary aim of the cleaning process was to remove duplicated data and convert the dataset into a more exploitable format, namely comma separated value (csv) files containing the relevant information slimmed down to reduce processing times. Only keeping the variables relevant for the intended analysis and aggregation (those listed in Table 2) made the dataset 7.5 times lighter than the unzipped files provided, from an average file size of 1.5GB to 0.2 GB. The following description details the processes illustrated in Figure 12, more specifically the first column (cleaning).

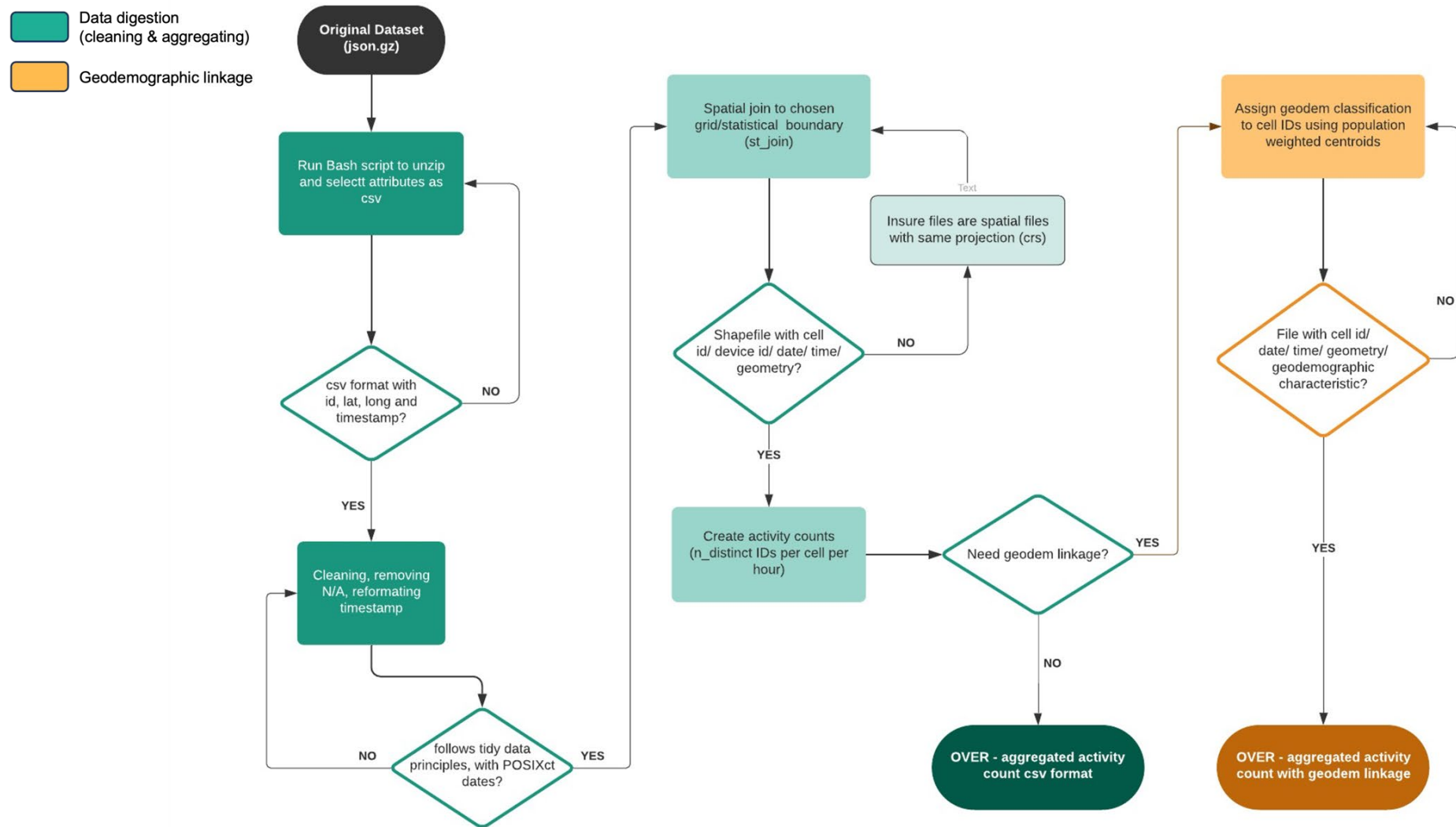


Figure 12. Flowchart overview of the data digestion process, with the data digestions steps (comprising cleaning and aggregating) in green and data geodemographic linkage in orange when necessary (secondary step for conducting analysis and data explorations).

The first step was to convert the daily datafiles from zipped json.gz to csv files. Due to the size of the original dataset, this was achieved in the bash command line within the DSH’s Linux machine, and by parallelizing the file conversion task to be processed and divided between various nodes. More precisely, rather than a “format conversion”, this step consisted in extracting the relevant columns directly from the json files without fully loading them in, and concatenating them into daily csv files. The columns kept in the cleaned csv files were initially *device ID*, *latitude*, *longitude*, and *timestamp*, as highlighted in Table 2. Each of the 1554 individual files averaged at 1.5 GB zipped which made the reformatting of the data computationally demanding, despite using parallelisation methods. The format of the original data files also presented technical issues. Ndjson (newline delimited json) files require different functions and packages compared to traditional json files, which needed to be approved by the DSH (Aubin, 2020).

All data-exploration steps previously discussed (temporal distributions and app statistics) necessitated the completion of these cleaning steps to access and assess the data. This intermediary dataset (Table 3) follows the principle of clean data as described by Hadley Wickham, where each row is a new observation, each variable is in a separate column, and each cell contain a single value (Wickham, 2014). Keeping latitude and longitude separated in two columns later facilitates geographic linkage in R. Any entry for which either latitude or longitude were missing were removed, as these would be unexploitable geographically. After this step, the dataset comprised 1554 individual csv files (one per day in the dataset), averaging 0.2 GB per file.

Table 3. Example tibble of the data at this stage.

device iid hash	impression lat	impression long	timestamp
ABCD1234==	12.34567	-123.4567	2021-01-01 24:00:00 UTC

Though other attributes (such as the location accuracy, app ID, etc.) were used for exploratory analysis, device ID, location and timestamp were the most important features for the creation of aggregated products of the dataset. Keeping only what is necessary not only complies to GDPR recommendations, but also follows best practice in ethical location data usage (GDPR, 2018a; EthicalGeo, 2021). Device IDs are kept at this stage as a necessity in the making of aggregates: they serve to determine whether a given cluster is the result of a crowd or generated from a single device registering multiple activities. Anonymisation of the dataset for third party use (or retrieval from the DSH) requires the removal of these device IDs from any data

outputted from the DSH. The following section describes how those IDs are used to create activity counts before being removed.

3.2.3. Typical aggregation process

This section describes the second half of the data digestion process as illustrated by the flowchart in Figure 12: how to spatially aggregate the clean in-app dataset to create non-disclosive data products. This process follows discussions from Section 2.4.1 of the literature review (Chapter 2), which emphasises the importance of not only removing identifiable information, but also coarsening the data where granularity could be disclosive. The methodology described here follows the principles of spatial aggregation (from unidimensional point data to areal counts) described in Section 2.4.1. It is generally applicable for generating point-in-polygon aggregates using various geographies, and is the core methodology used throughout this thesis when testing various aggregation scales (in Section 3.4.3 of this chapter and throughout Chapter 4) and when making a bespoke regionalisation methodology (Chapters 4 and 5).

3.2.3.1. Spatial join to chosen scale.

Aggregating the dataset involves removing any identifiable components, such as device IDs and individual locations. Traditionally, the main aim of the aggregation process is to attain a level of privacy protection. Practically speaking, other central purposes of aggregation are to facilitate the dataset's usage outside the secure environment and produce subsets which do not require extensive technical knowledge for analysis. Predefined activity counts are an easier metric to apply in wider research than uncleaned, standalone GPS locations (Kishore *et al.*, 2020).

Prior to aggregating, further processing of the cleaned data was needed. Aggregation-specific cleaning included separating the timestamp into date and time attributes, in a POSIXct format (See *lubridate* package in Appendix 2), facilitating hourly grouping and avoiding inconsistencies across the files' date formats. The second stage to the aggregation process was to link the clean dataset to shapefiles of chosen geographies, namely performing *point in polygon operations*. The term "point in polygon operation" is borrowed from computational geometry: it describes the operation of determining whether a given point is contained within the boundaries of a given polygon (Hormann and Agathos, 2001). There is extensive literature

assessing various point in polygon algorithms and their relative computational costs and efficiencies (Huang and Shih, 1997). For this project, `st_join` from the `sf` R package is used to perform the point in polygon operation: it assigns a polygon's ID and geometry to each recorded activity based on its coordinates (See Appendix 2 for `sf` documentation). Table 4 illustrates the data after this step.

Table 4. Assigned polygon IDs (Unit ID) and geometry to device ID hash, replacing individual impressions' latitude and longitude.

Unit ID	device_id_hash	Date (POSIXct format)	Hour	Geometry (of unit)
AA0000	ABCD1234==	01/01/2020	24:00:00	(-12.34567, ...)

Then, using the `data.table` package (Appendix 2), the counts of impressions corresponding to each polygon are summed to create a count of points per polygon, finally removing individual impressions' latitudes and longitudes.

3.2.3.2. Creating activity counts

There are three main ways to count the numbers of points per polygon (Figure 13):

- (1) *Impressions (total counts)*: sum of the total number of records, regardless of device.
- (2) *Unique devices (activity counts)*: the number of devices. This records only one impression per device per polygon, and corresponds to an "activity count" throughout this research. Here, one activity is considered to be representative of the presence of one individual, and duplicates of records each person may produce in one same place and time are ignored. Individuals are counted each time they cross through to other polygons (once per polygon travelled).
- (3) *Modal locations per device*: the most frequent location (unit ID) of a device. With modal location, each device is attributed the polygon it spends most time in, the rest is discarded, ensuring to only record an individual once over the time-period selected. This significantly slims down the dataset.

Figure 13 helps visualise these different types of counts. The three ways represent different levels of precision and data representation and are suitable for different purposes. This project mostly makes use of unique device counts (activity counts) as it seeks to filter out the noise generated by the most active users, and modal locations prevent the examination of hourly or

daily patterns due to the significant reduction of records kept. Unique device counts also considers devices as they travel through regions, which presents benefits for aggregating datasets for mobility analysis (Tizzoni *et al.*, 2014; Wang *et al.*, 2019).

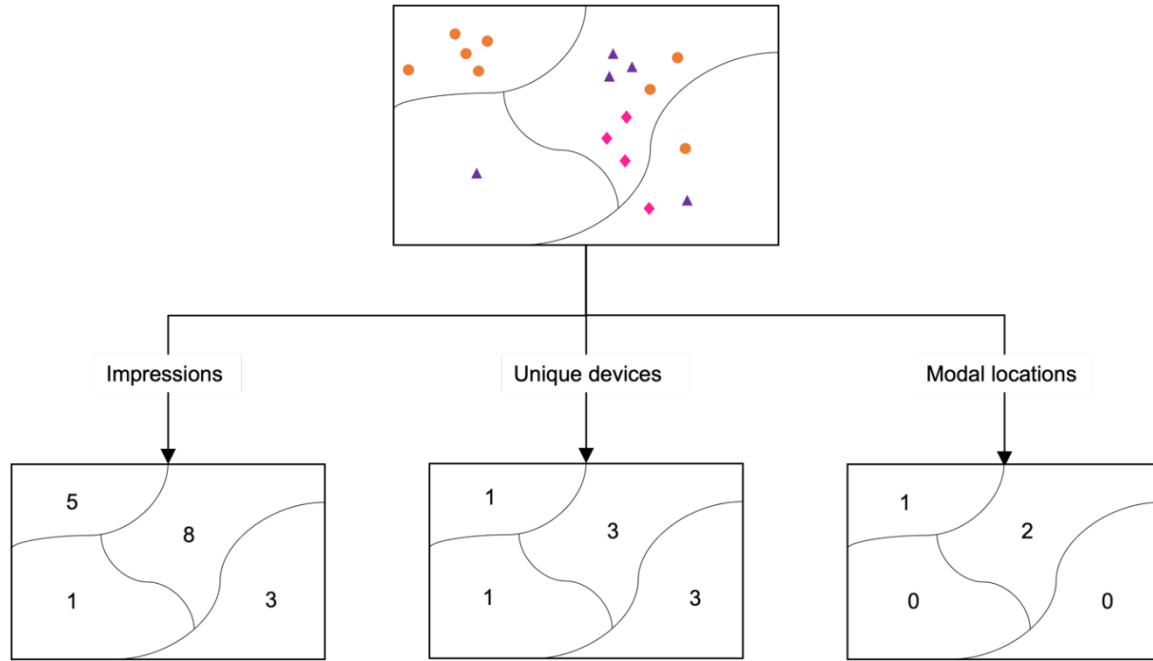


Figure 13. Different methods to count the data when aggregating the dataset. Each point type (circle, triangle, diamond) describes a different unique device, with each device creating multiple impressions.

Thus, the overarching purpose of selecting activity counts (unique device counts) as the measure of population in a polygon is to preserve the majority of data points while filtering out large amounts of noise. However, this also captures a specific definition of what is understood as ‘activity’ or ‘population’ throughout the thesis. At the polygon level, the activity count captures the population in each polygon during a specific time period. This means that, if over this time period the same user travels from polygon to polygon, it is considered in the population of both polygons. This helps estimate a sense of movement throughout the city, but might inflate the activities of the most dynamic populations and does not capture population overall, but population relative to individual polygons. Some behaviours cannot be measured or inferred by the activity count, such as the time a point remained within a polygon over the selected time period (no measure of dwell, or differentiation of users only ‘passing through’), or whether there was movement within the polygon over this time. Thus, the selection of the time period becomes crucial in defining what is captured; this is further explored throughout Chapters 5 and 6.

To count the number of unique device ids per polygon per hour, for example, an aggregation function was written which loops through each daily file, uses point-in-polygon techniques (`st_join`) to assign each impression with a polygon unit ID (for example, a grid cell ID such as OSGBs presented in Chapter 2), counts the number of unique device IDs per geographic unit per hour and returns this count directly into a new data frame. Table 5 provides an example output of this function, which corresponds to the end of the data digestion process in Figure 12.

Table 5. Activity counts (sum of unique device IDs) per geographic unit per day per hour. The geometry can be kept for mapping purposes.

Geo unit ID	Date	Hour	n activities	Geometry (of unit)
AA0000	01/01/2020	24:00:00	10	(-12.34567, ...)
AA0001	01/01/2020	24:00:00	23	(-12.34567, ...)

This aggregated dataset can then be removed from the DSH, or other secured environments if counts below 10 are hidden or removed. This allows for the safer and easier use of initially sensitive and technically demanding datasets. The impacts of aggregation, and the importance of scale and zone selection for this dataset are discussed in depth in a dedicated analysis (Chapter 4).

3.3. Making of a standard practice CDRC aggregate for further data descriptions

To further describe the in-app data and retrieve initial explorations from the DSH, an aggregated subproduct was created using the cleaning and aggregation methodologies described above. This subproduct was created in accordance with standard CDRC practices to facilitate access to a dataset sample for CDRC data users during the initial phases of this research (CDRC, 2022).

CDRC safeguarded datasets can be accessed via remote services with registration to the CDRC data portal and after approval of the target project. These types of datasets do not contain personally sensitive or disclosive data, but are restricted due to commercial sensitivities or license conditions (CDRC, 2022). This specific subproduct was made using the aggregation steps and function described above, and provides activity counts per km² daily, at the GB scale. The OSGB 1km x 1km shapefile was used to aggregate the in-app data into the resulting activity count. The aggregated data sample comprises three variables: grid cell ID, timestamp, and aggregated activity count.

The following sections describe this aggregated subproduct and use it to produce and present further exploratory analysis of the data, which could not be proposed here before aggregation steps due to the sensitive nature of the raw dataset. However, using an aggregated dataset to illustrate the in-app data and its coverage raises concerns pertaining to potential loss of information from the original dataset occurring during aggregation. This is further explored in the final part of this chapter, Section 3.4.3, and throughout future chapters.

3.3.1. Method and metadata

The data subproduct was created by scaling up the cleaning and aggregation processes to create activity count files from the entire in-app dataset file by file. For this, a function in R runs through all data files using multi-core processing. The function divides the task across $n-1$ nodes available on the DSH, retrieves the 4 attributes needed for the creation of activity counts from each csv file, joins them to the chosen shape file (for this subproduct, square kilometre, using OSGB grid cells as a standard unit [Brunet *et al.*, 1993]) and counts individual phones per grid square per day (see Figure 12). This parallelized process ran over the course of 30 hours within the DSH (on 7 nodes equipped with Intel 64 processors), to create aggregated 1km by 1km OSGB grid activity count dataset for each day of data. The daily files were then concatenated using the bash command line to create one csv file containing daily activity counts across GB for the full data period (2016-07-05 to 2020-10-05). This subproduct's metadata is listed in Table 6. Square kilometres are chosen as they are the worldwide traditional standard for representing population densities for countries (See Brunet et al.'s definition of *population density* [1992, p.148]).

Table 6. Metadata table for the aggregated subproduct.

Field	Value
Data Provider	Huq ltd / CDRC
Analytical Units	Mobile Phone GPS locations per square kilometre
Data Format	Character Separated Values (csv)
Temporal Extent	05-07-2016 to 05-10-2020
Geographical Extent	Great Britain
Variables	3 variables across one database
Observations	+64 million records
Grid units (OSGB)	246021

3.3.2. Data exploration using the subproduct: spatial coverage

3.3.2.1. Spatial coverage and insufficient data events (IDEs)

The data's spatial extent covers Great Britain. A large majority of the data is recorded within metropolitan areas, Greater London in particular. For example, in 2019, 54% of all impressions are recorded in Greater London, whereas the ONS population estimates records 12.4% of GB population living in the capital for the same year. Other areas, mostly outside of metropolitan areas, register data and cell coverage, but in very low numbers compared to urban centres, often resulting in insufficient data events (IDE, below 10) preventing conclusive analysis. Furthermore, some areas, particularly in North Scotland, may not record any data throughout the entire four years due to low cell coverage. This points towards a significant bias of the dataset towards metropolitan areas (particularly Greater London). The map in Figure 14 displays the data coverage in total records over the data period, aggregated to the OSGB 1km².

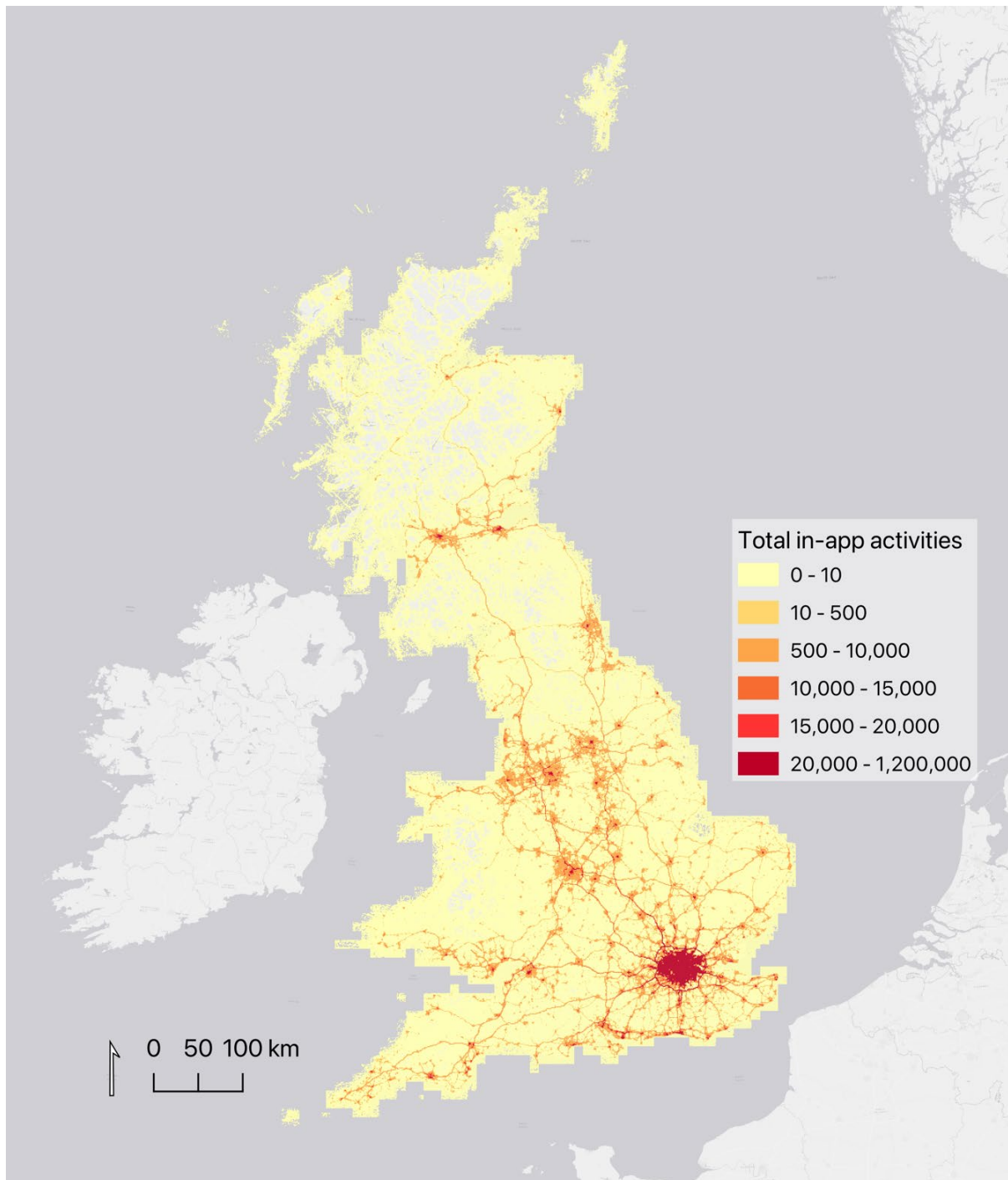


Figure 14. Spatial distribution of the data nationally – The highest counts of impressions are recorded in Greater London. Other big metropolitan areas are visible with denser city centres. This helps in part justify why this report focuses on metropolitan areas, specifically London.

Using the aggregated subproduct (activity counts per km² per day), we can display the proportion of units which fall under the 10-device threshold throughout this version of the dataset, to identify areas which have significant counts for analysis. In themselves, IDEs are a result of the aggregation method (the number of devices counted is directly dependent on the aggregation unit), but they give an indication of relative areas of low or high data across the

UK. Some regions do not have any records at all, particularly Northern Scotland and remote locations in Wales, as explained above. Figure 15 highlights the disparities in the representation of rural and urban areas in this dataset by mapping IDEs throughout the data period.

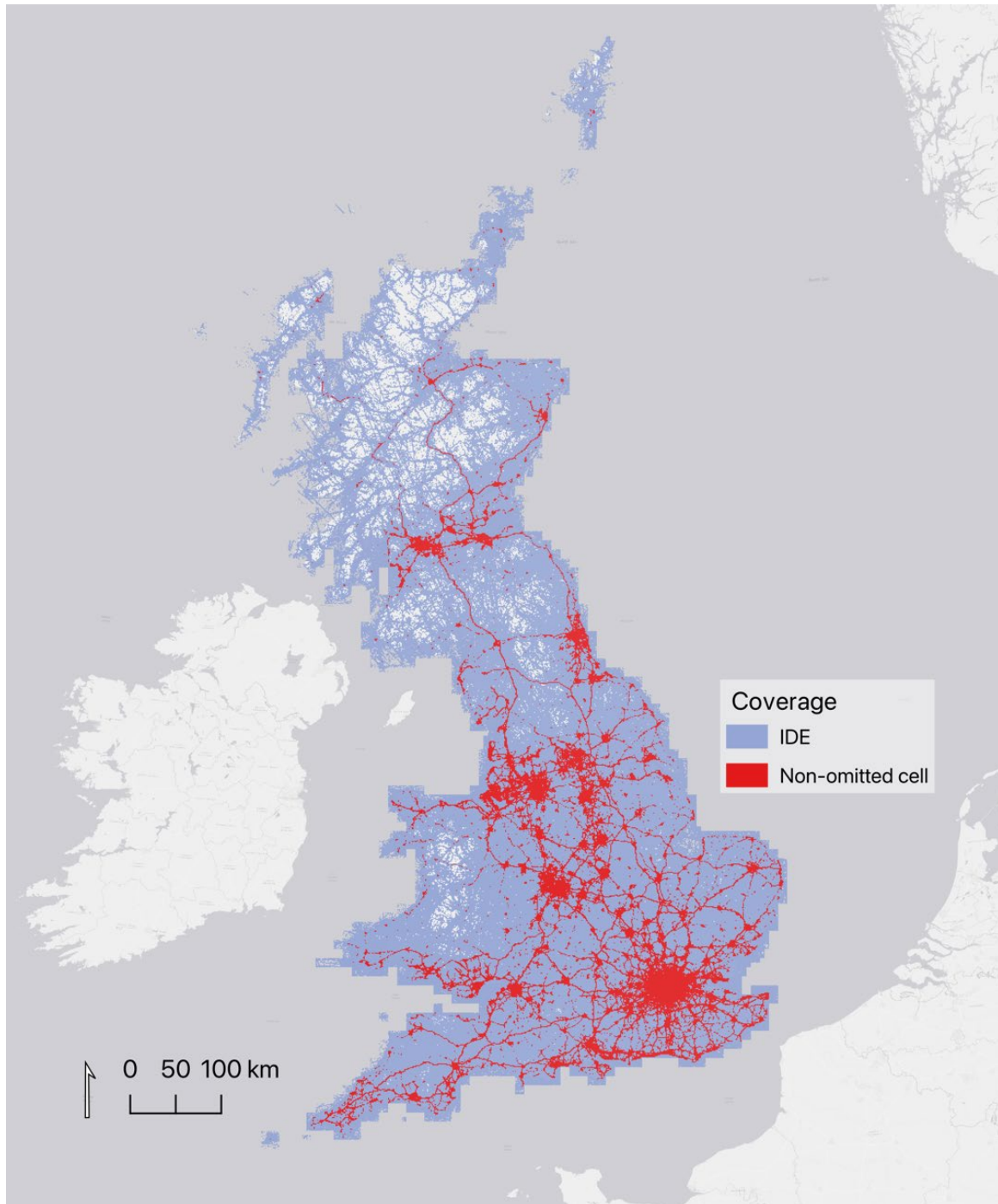


Figure 15. Coverage map, GB scale. Grey-blue grid cells represent activities being consistently below a count of 10 for each day in the dataset (IDEs). Red grid cells are above a count of 10 devices for at least one day of the data period. Missing grid cells show missing data throughout the entire data period (if a grid cell recorded one activity over even a single day over the four years, it would appear blue rather than white). Central Wales and Northern Scotland display large proportions of missing data. The data shows low significance on the national scale, with 85% of grid cells consistently recording counts of 10 or lower

Spatial coverage is also temporally affected, due to inconsistencies in the data distribution over time. As discussed in Section 3.1.3, activities increased from a couple hundred impressions to 20 million throughout the data period, doubling over 2020 alone. Following this general trend in uptake, earlier data have the highest percentage of IDEs (particularly 2016), and data suppression (having to obscure activity counts of less than 10) generally decreases over time, particularly in metropolitan areas. Data suppression levels in metropolitan areas generally fall below 50% by 2020. Table 7 lays out the percentage of IDEs in the full data subproduct each year, comparing it to their proportion in metropolitan areas alone. This shows that, overall, metropolitan area data is more significant, as smaller percentages of the activities need to be removed from the subproduct dataset due to IDEs. These coverage issues and overall low data counts might reduce the value of the dataset in some areas, particularly rural areas in earlier years, where well above 80% of the data needed to be omitted due to IDEs. This calls for a focus on metropolitan and urban areas and often later years (2019-2020) for analysis throughout this thesis.

Table 7. Percentage of data omitted each year due to IDEs in the aggregated subproduct dataset. Comparison of GB extent and metropolitan areas.

Year	% IDEs in full dataset	% IDEs in Met Areas
2016	94.45	82.39
2017	94.57	82.14
2018	92.49	72.51
2019	89.74	66.36
2020	81.80	41.97

3.3.2.2. Impact of using an aggregated product for data coverage exploration

Using the 1km² grid scale subproduct, IDEs concern more than 80% of the overall UK's area. This indicates unequal data distribution in the raw data, but may also be exacerbated by the aggregation process. When making this subproduct, activity counts were aggregated *per day*. Thus, the resulting low activity counts in less active or populated areas may also be affected by the temporal scale chosen. Unique devices could also be counted *per hour* and *summed per day*, but this creates potential repetitions of the same devices over the day. Such choices must be informed: counts per day remove duplicates, but provide lower counts, resulting in data loss in

less densely active areas. Counts per hour summed per day provide more data counts but do not treat duplicates – this means that a minority of users could be generating a majority of events. Repetitions of the same device over a day could be thus interpreted as a cluster of activity, as it would become indistinguishable from multiple devices emitting activity. These issues of duplication are hard to control for and remain a common issue in aggregation strategies of these types of datasets (Slivinskas *et al.*, 2001). For daily or hourly analysis, one would want to have the counts per hour present in the dataset, and risk duplication (Sevtsuk and Ratti, 2010; Kim, 2020). However, for a yearly summary of the data, summing the devices per hour generates more problems than it would solve (adding uncertainty, potentially inflating certain areas where a very active app may appear and disappear throughout the data etc.). These issues pertaining to duplicate counts from the same device remind of the earlier descriptions of the three different ways to count activity when aggregating the data (Section 3.2.3.2), but here over a temporal rather than spatial dimension.

Regardless, the subproduct dataset is non-disclosive, tractable and publicly available: all traits not shared by the raw dataset. However, the necessary aggregation limits the type of analysis which may be conducted using the in-app dataset in this current state. It helped provide an exploratory analysis of the in-app data spatial coverage, but may fall short for analysing the data bias and representativeness, which may require more detailed aggregates than daily counts over 1km² grids.

3.4. Data representation, bias and aggregation impact

The spatial distribution disparities of the data, explored through the data subproduct in the previous section, show the dataset is biased towards urban populations, with most of the data being recorded in Greater London and other metropolitan areas. This section seeks to assess the data representativeness beyond urban vs rural, and investigate which populations may be over or under-represented within the in-app data sample. Firstly, we assess the proportion of overall population represented in the dataset by focusing on Greater London, where a high density of impressions allows for a more representative analysis. Then, the data's population bias is investigated by linking the raw data points to census geographies using the aggregation methodology previously detailed, and comparing the proportions of different population categories between census estimates and the in-app dataset.

To assess the data bias as accurately as possible, the raw dataset is used. This is only possible thanks to this project's access to the original data, and the facilities and services provided through the DSH and the CDRC. As the aggregated subproduct was the first publicly available version of this dataset, we demonstrate how the same data representation and bias analysis would be conducted using this data subproduct. This comparison demonstrates the impact of aggregation on exploratory analysis: the population biases underlying the dataset are not the same when using the raw data or the subproduct. This final section aims to raise concrete concerns about the ways sensitive datasets are commonly accessed. Where only the subproduct is available, researchers have incomplete and altered insights into the data they are using.

3.4.1. Population estimates

To assess the proportion of the population represented by the in-app data, activity counts were compared to geodemographic counts in established geographies. Census data records residential statistics, thus night-time (between 11pm and 5am) impressions counts were retrieved to compare with ONS data from the 2011 census, as night-time data is the most suited for comparison with residential data (Tooru and Kawakami, 2013; Vanhoof *et al.*, 2018). Night-time location is a commonly used heuristics for attributing residential information to location data (particularly mobile phone data) (see Phithakkitnukoon *et al.*, 2012; Calabrese *et al.*, 2013; Kung *et al.*, 2014). This is based on the assumption that “home is the location that has the most activity between x p.m. and y a.m.” and relates a night activity's location to the home location (Vanhoof *et al.*, 2018). Thus, the hours of 11pm and 5am were selected from the raw dataset to tighten this criterion and avoid capturing commuters, and night-time impressions between these times are considered as being emitted from residential locations.

The in-app data for this assessment consisted of a year's worth of night-time impressions, further aggregated to contain only individual devices (no repeated impressions for the same device throughout the year) and spatially aggregated to OA to create annual night-time activity counts per OA. 2018 is selected: the year with the highest count of unique devices in the dataset (see Section 3.1.3) and no significantly disruptive unanticipated displacements such as the Covid-19 pandemic.

The night-time activity counts per OA were created following the cleaning and aggregation methodologies presented throughout Section 3.2: the data was spatially joined to London OA census data using point-in-polygon techniques, to obtain the number of unique night-time

devices per OA. The resulting activity count was compared to the usual resident population (USUALRES) for each OA, collected from the 2011 census. Apart from two OAs in London which had 10-20% of their USUALRES matching the data's activity counts, the large majority of OAs fall in the 0.1-1% range of activity representation. **Error! Reference source not found.** shows the distribution of OAs according to the percentage of USUALRES represented by the in-app data activity count.

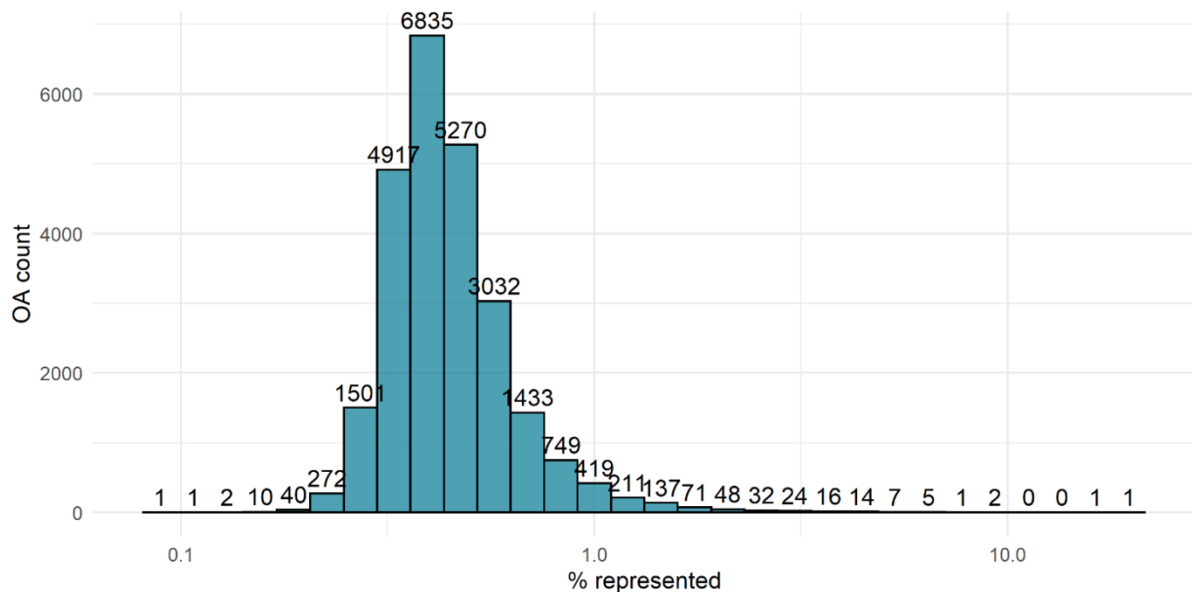


Figure 16. Count of London OAs distributed by the percentage of their population represented by the night-time in-app data activity counts. For instance, the data seems to capture about 0.6% of the population in 6835 OAs.

This analysis concludes that, on average, the dataset accounts for less than 1% of the London population (mostly 0.6%). Overall, despite mobile phones being ubiquitous, individual datasets are not necessarily representative of most of the population, as they still represent a restricted sample (filtered by app types, device types, operating systems, night activity filtering etc.) (Raento, *et al.*, 2009; Lenormand *et al.*, 2014). Furthermore, to compare the activity counts to residential information, a lot of data counts were omitted when filtering for the residential, “night-time” statistics. This is a persistent issue in the literature surrounding the topic of attributing residence location to datasets such as this one (Kung *et al.*, 2014). Not only is the data filtered for night-time (only a small percentage of the whole dataset is eligible to be attributed residence), but the night is most often the period when the least data is generated, reducing the counts further (Vanhooft *et al.*, 2018).

Regardless, in this scenario where the night-time counts reduce the activity recorded greatly, it appears that in-app dataset captures roughly 0.6% of the Greater London population. In 2019,

this would constitute a non-negligible sample of 53,500 people. This number can be expected to be larger as the dataset is biased towards day-time activities as explained above.

3.4.2. Population representation

3.4.2.1. Methodology

If about 1% of the London daytime population is recorded by the in-app dataset, questions pertain as to who is represented in this sample. A proportional and representative sample of 1% of the population could be highly significant for a number of mobility studies, and represent a rich data sample compared to traditional datasets (Kitchin, 2014a; Lenormand *et al.*, 2014; Wang and Chen, 2018).

The same subset of data as with the population estimate analysis was compared to the London Output Area Classification (LOAC) (Longley and Singleton, 2018). The LOAC is a hierarchical structure which uses over 60 census variables to classify London's OAs. It is a useful tool to obtain an overview of the various populations towards which the dataset may be biased, without the need to filter individual geodemographic characteristics. The LOAC was chosen for its open-source nature and versatility which allowed to make up a profile of the night-time in-app data activities by joining them to London OAs as previously described and assigning them the LOAC group associated to the OA. The groups used for this analysis are detailed in Table 8 (Longley and Singleton, 2018). The geodemographic characteristics underlying each group are discussed in the following results section. Further detail on the LOAC classification and the methodology behind its development are available on the London datastore (UCL and GLA, 2015; Longley and Singleton, 2018).

Table 8. LOAC supergroups and associated groups with their colour code, as found in the LOAC documentation (Longley and Singleton, 2018).

Supergroup	Group
A: Intermediate Lifestyles	A1: Struggling suburbs
	A2: Suburban localities
B: High Density and High-Rise Flats	B1: Disadvantaged diaspora
	B2: Bangladeshi enclaves
	B3: Students and minority mix
C: Settled Asians	C1: Asian owner occupiers
	C2: Transport service workers
	C3: East End Asians
	C4: Elderly Asians

D: Urban Elites	D1: Educational advantage
	D2: City central
E: City Vibe	E1: City and student fringe
	E2: Graduation occupation
F: London Life-Cycle	F1: City enclaves
	F2: Affluent suburbs
G: Multi-Ethnic Suburbs	G1: Affordable transitions
	G2: Public sector and service employees
H: Ageing City Fringe	H1: Detached retirement
	H2: Not quite Home Counties

3.4.2.2. Results

The usual resident population (USUALRES) of each OA was aggregated, provided with the LOAC data, to obtain the total count of USUALRES per group. The proportion of census resident population in each group are compared with the proportion of in-app data activity per group for the night-time subset to obtain the following results.

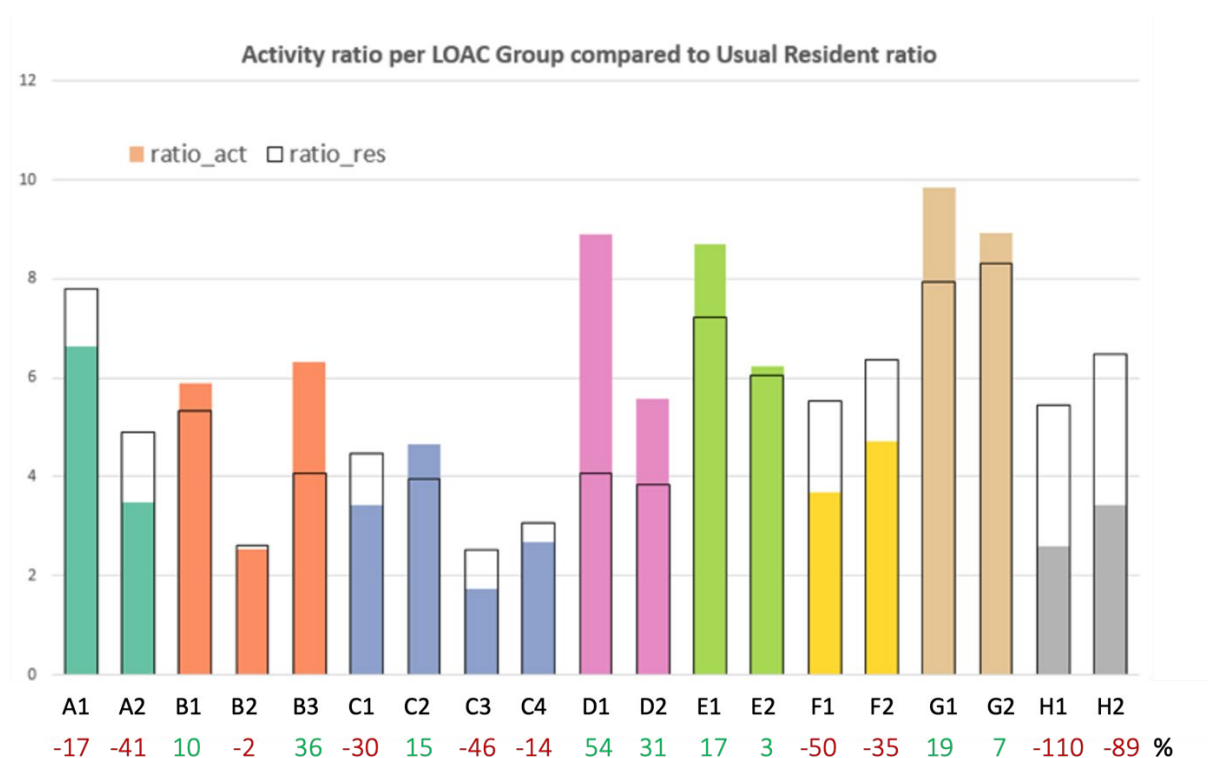


Figure 17. Comparison of population proportion by group (outline) and proportion of activities recorded by group (fill). Below each category, the percentage difference between in-app data activities and resident population. An under-filled bar (and negative difference) shows an under-representation of that group in the in-app data sample, and positive an over-representation

Figure 17 gives an overview of which populations are over-represented in the dataset. At first glance, the B groups activity (high density and high-rise flats) deviates the least from the census USUALRES for this group. D groups are significantly over-represented by the dataset (urban elites) and the ageing city fringe significantly under-represented (H). Diving more precisely into those categories, the most over-represented population (D1) is characterised by overcrowding and student lifestyles (Longley and Singleton, 2018). This supports the idea that consumer generated dataset such as this one may have the majority of their events generated by a minority of the population characterised by specific consumer dynamics (Lansley and Cheshire, 2018). Similarly, B3 is 36% more recorded by the in-app data activity than its census residential population, and is characterised by ethnically diverse, often student populations. The G supergroup, also overcounted in the dataset, is composed of young, diverse populations. H and A, under-represented by 100% and 30% respectively, are generally groups of older populations and families living on the fringe of London. The most underrepresented group is the mostly UK-white and retired population of group H1 (-110% less than USUALRES). In a few lines, older, often white populations living in city edge neighbourhoods are underrepresented by the mobile in-app dataset, which is more biased towards central, young, and often ethnically diverse student populations. This reflects an expected use of the technology and apps collecting the impressions by more dynamic, younger groups (Lansley and Cheshire, 2018; Rosales *et al.*, 2023).

Many assumptions had to be made in the making of the in-app data subset for this analysis. Night-time activities do not reflect residential populations alone, but may also pick up activities from night workers and social outings. The impressions collected between midnight and 5am were processed regardless of what activity could have generated them, and thus may impact the results presented. However, running a correlation test between the residential variable and the in-app data variable returns ~ 0.63 , indicating an encouraging moderately positive relationship. This analysis still provides an overview of the dataset's biases within the better covered urban centres. It confirmed the hypothesis that older, often suburban population are under-represented by these data types, privileging younger, and more centrally situated populations (Rosales *et al.*, 2023). It also reiterated the dataset's bias towards urban populations.

3.4.3. Comparison of population assessments differences between the raw dataset and the 1km² data subproduct

The analysis conducted above to assess the data's population bias was done using a direct linkage of the raw in-app data to OAs and their respective LOAC. This was made possible by the access to the raw data within the DSH. We here assess the result of an identical analysis if it was conducted outside a lab by a researcher using the 1km² aggregated subproduct of this data, provided and described in Section 3.3.1. This represents a more typical experience of using such datasets, not requiring specific training or remote access (de Montjoye *et al.*, 2018). To do so, the analysis above was recreated using the 1km² grid aggregated subproduct. The resulting London population profiles are compared against the ones obtained by the previous analysis using the raw in-app dataset directly aggregated to OAs. This aims to illustrate the impact aggregated products have on research results, and motivates the necessity to reassess the way mobility datasets are typically disseminated.

3.4.3.1. Linking the aggregated subproduct to OAs

The final steps in the Figure 12 flowchart are concerned with geodemographic linkage. Conducting analysis using the subproduct requires linkage to geodemographic statistics and classifications. Keeping the OSGB geometries, or at least providing a list of coordinates for each grid unit in the creation of the subproduct, facilitates potential linkage to geodemographic characteristics. Thus, joining geodemographic information to the grid squares by matching their coordinates suffices, not requiring further point-in-polygon operations. We use `st_join` in R to attribute LOAC classifications to each grid cell. Where multiple classifications overlap a given grid cell, weighted centroids were used to attribute the most significant LOAC based on population counts (Trasberg and Cheshire, 2020). Table 9 is an example of linked data at this stage. To assess the population bias, a subset of the 1km² grid dataset is sampled and linked: one week of data in 2018 in London. This data subproduct no longer has a timestamp as it provides aggregated activity count per grid cell *per day*.

Table 9. Activity counts per grid cell, with the grid cell's corresponding LOAC (synthetic example). Summing `n_activities` per LOAC returns the number of activities for each group.

Grid cell ID	LOAC	Date	n_activities	Geometry (of cell)
AA0000	City Vibe	01/01/2020	10	(-12.34567, ...)
AA0001	Urban Elites	01/01/2020	23	(-12.34567, ...)

In general, such linkage is imperfect, with approximate containment of geographies, or data being lost through the imperfectly overlapping units. This echoes earlier discussions around MAUP and spatial scale problems, especially thorny when integrating data at different scales (Atkinson and Tate, 2000). This added step of integrating data from square grid cells to administrative boundaries adds another layer of uncertainty and multiplies arbitrary decisions. It further reduces analytical completeness and validity (Purdam and Elliot, 2007; Casado Díaz and Coombes, 2011). Where possible, directly conducting the aggregation to the scale at which any other required data is collected minimises such issues (for instance, creating activity counts per output area if the study requires comparison with census data at OA levels, as done previously). However, this is often impossible as only specific data aggregates are made available, often at pre-determined scales and zones requiring further geographic linkage, such as the aggregated subproduct tested here.

3.4.3.2. Comparison results

The resulting activity counts per LOAC group obtained through the aggregated subproduct are compared with the activity counts per group obtained by direct linkage (OA) for the same week. Figure 18 maps the groups assigned to OAs and 1km² grid respectively, and Figure 19 shows the activity count per 1km² grid square.

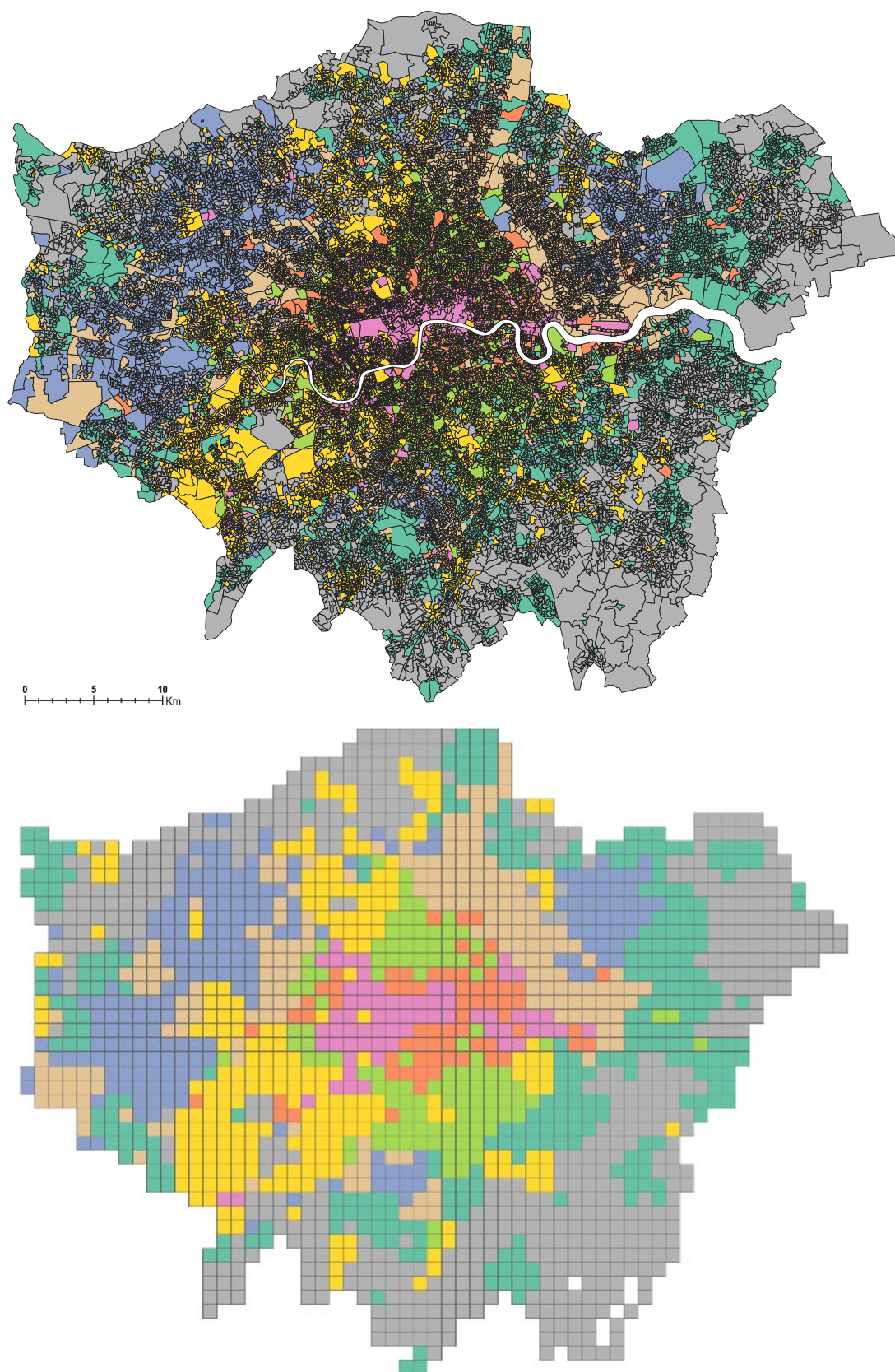


Figure 18. Maps of London with coloured LOAC groups. Top: OAs with their corresponding LOAC group. Bottom: groups assigned to 1km² grid cells through areal overlap and population weighted centroids.

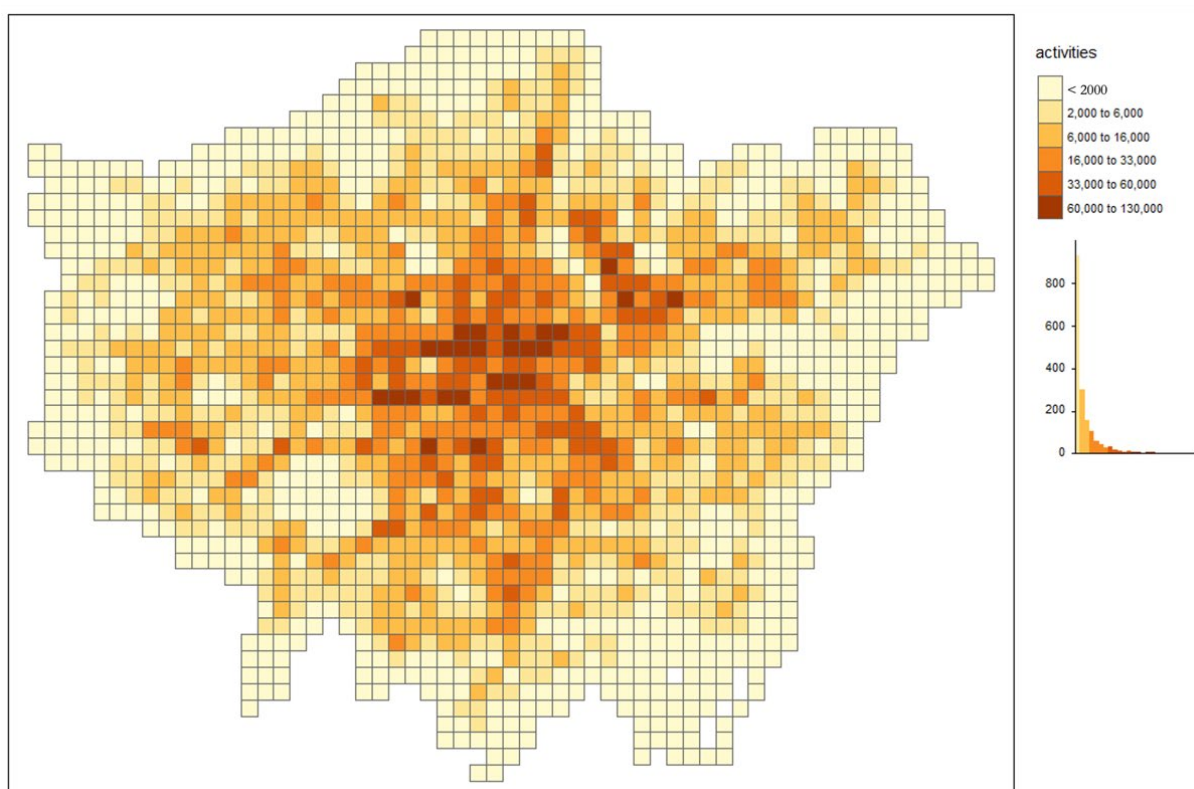


Figure 19. 1km² OSGB grid map of London, showing the total activity level per unit over the analysis period

Though both profiles compared were obtained from the same data (in-app data) and period, the resulting population profiles are different. The first graph (Figure 20) summarises these population profiles. It shows that the subgroups are not equally represented if the analysis was conducted through direct linkage, or through the intermediary of a pre-aggregated dataset. Group B3, for example, accounts for around 4% of overall activity according to the 1km² grid analysis, but close to 7% for the direct join. Figure 21 displays the proportion of each subgroup's population for the OSGB grid as a function of those of the direct OA linkage. Some subgroups' activity counts (B2, C3 and F1 for instance) are between 70 and 100% higher in the population profile generated by the aggregated subproduct results.

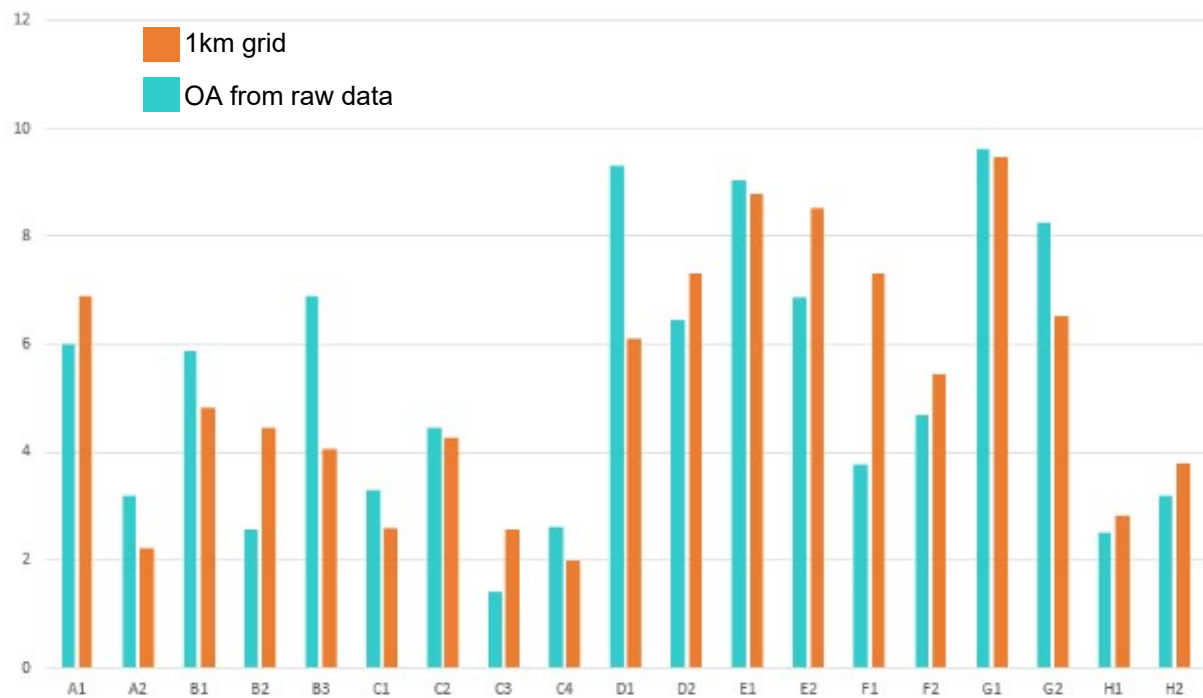


Figure 20. Bar plot – proportions of each subgroup's contribution towards the total activity count for the data period. Blue represents ratios obtained by direct join, and red obtained by the 1 km² grid aggregate analysis.

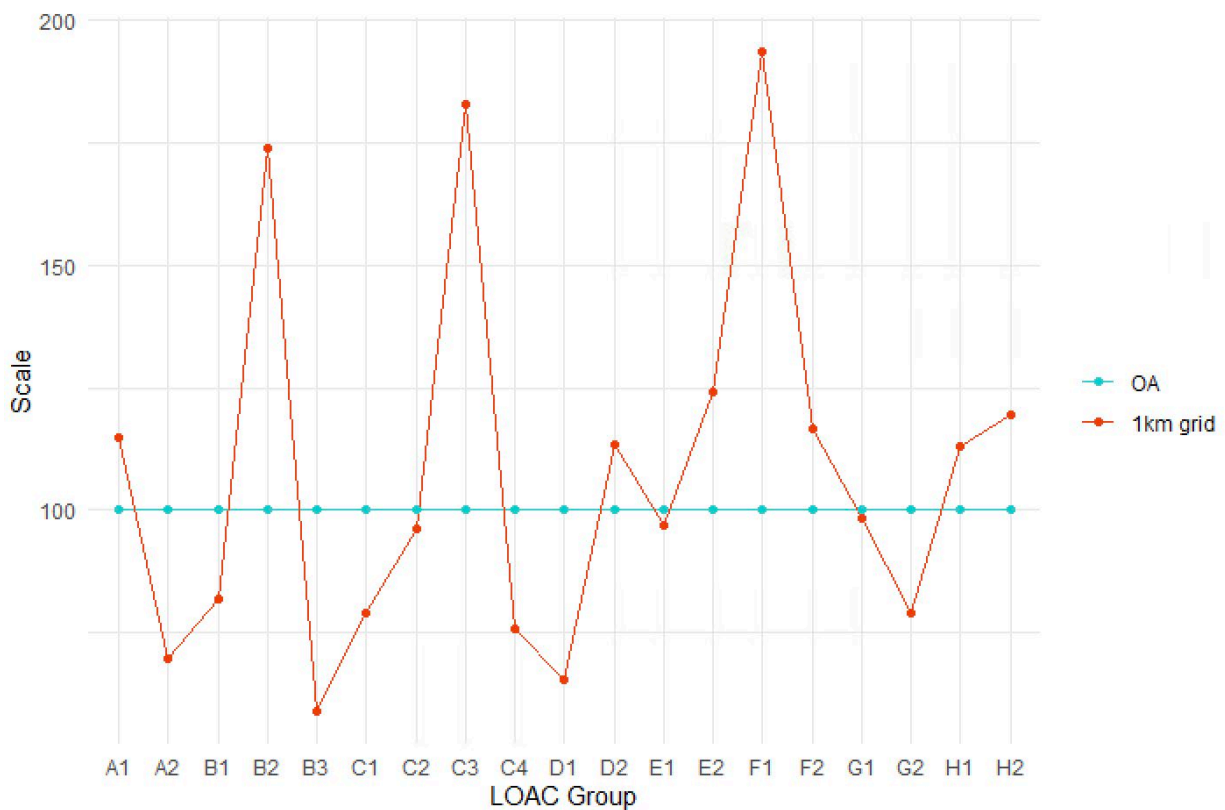


Figure 21. Line Graph – population profiles for the 1km² grid analysis (red) displayed as a function of the direct join to OA profiles (green). Each OA activity count was assigned the value of 100, and each 1km² grid cell activity expressed a percentage of OA activity.

This concludes that aggregation scale and methods have an impact on analysis conducted using the data, but does not precisely highlight the ways in which this can be accounted for, nor inform which aggregation scale to privilege to prevent these differences. Where possible, using the raw data reduces the issues pertaining to geographical linkage and integrating data at different scales. However, this chapter discussed the difficulties in accessing such raw datasets, both in terms of their disclosure control restrictions, and their technical challenges. MAUP is made more problematic by the accessibility issues surrounding these types of sensitive consumer datasets (Brunsdon and Comber, 2020). Without access to the raw data, researchers cannot clearly account for issues highlighted by this comparative analysis, or decide on a more appropriate aggregation scale or zone.

3.5. Chapter summary

This chapter presented the in-app dataset used throughout the rest of the thesis. It explored the data's temporal and spatial coverage, described the app and device statistics and presented the cleaning and aggregation methodologies used to process the dataset. Part of the data exploration required the making of an aggregated subproduct to safely publish samples of the data, particularly to illustrate national coverage. A subproduct was thus developed following standard practice. However, a key outcome of this chapter was the demonstration of a necessity to further consider the way these aggregates are made. The final analysis showed that results obtained with an aggregated dataset are different from those obtained from raw data, sparking concerns surrounding choices of scale and zone in aggregation.

This research recognises the importance of aggregated data products for dissemination. They save great amounts of resources (human, financial and computational) and democratise the access to novel forms of data. However, general aggregates tend to be arbitrary: we have shown how this impacts subsequent analysis for those who can only use the aggregates. This highlights a pragmatic issue in the ways sensitive in-app dataset are accessed. The high hurdles to raw data access are non-negotiable: they safeguard the datasets from misuse and protect individual and commercial interests. However, more could be proposed to reduce the impact of MAUP on the aggregated subproducts made for more general use outside secure environments. The dataset's accuracies being largely below 150m, it is unfortunate to lose such a high granularity by aggregating the raw points to areal units of 1km². Better aggregates could seek to preserve the dataset's presented attributes, and mitigate MAUP effects, especially in the context of novel mobility data.

4. Regionalisation – Explorations

The strength of new forms of data, including in-app data, lies in their potential to study new patterns and behaviours. However, the data's patterns and timeliness are not necessarily preserved through the aggregation process, with existing regions having been built on previously known data distributions and behaviours instead. As covered in Chapter 2, the UK and Wales census OAs were drawn to best represent the underlying distribution and traits of residential populations. In-app data, with its daytime movements and long-tailed distribution, clusters in different places than census data and is used to answer different questions, at different temporal and spatial scales to the census, as demonstrated in Chapter 3 (Reades *et al.*, 2007). Thus, using the census bespoke regions carries with it further risks of loss of analytical validity and completeness, and makes it difficult to control for MAUP effects on analysis. Regionalisation presents opportunities to better aggregate new forms of data for dissemination in research, by creating bespoke regions which allow for granular analysis representative of the original data, whilst accounting for the disclosure risks presented earlier.

This chapter serves as a steppingstone between the demonstration of the impact of MAUP on the dataset and the synthesis of a new regionalisation methodology for in-app data. It first demonstrates why existing regions and aggregation methods are not fit for aggregating in-app data, namely due to their unique distributions and patterns. To do this, multiple aggregated products of the in-app dataset are generated at different zones and comparable scales, to highlight the changes in results between aggregates. In a second time, this chapter provides a list of identified requirements and principles necessary in the making of useful bespoke regions for this dataset. It then explores existing reproducible regionalisation methods, mostly built on tessellation techniques from transport research and physical geography, and reveals the ways in which they do not fulfil the listed target requirements for the making of reusable and comprehensible regions for the in-app dataset. It is revealed that these tessellation methodologies mainly present issues when it comes to region stability over time and linkage to other geographies. This progression leads the chapter to conclude on the necessity of creating a new regionalisation methodology which can account for both time and space concerns and be reusable by future users.

4.1. Assessment of MAUP impacts on in-app data aggregation

This section demonstrates the impact of MAUP on the in-app data when creating safe aggregates for research. As discussed in Chapters 2 and 3, the effects of MAUP are expected to impact data when it is aggregated from point to areal data. This is a well-researched and thoroughly documented phenomenon in geography, however there are few assessments of its impact on in-app data at the scale of the present dataset (Fotheringham *et al.*, 1995; Qi and Wu, 1996; Minot and Baulch, 2005). This analysis demonstrates that different aggregated products will generate different results, despite the original data and research topic being otherwise identical. This raises a key question regarding which of these results is to be considered as the closest representation of the studied behaviour. This subsection seeks to explore both the scale and zoning effects of MAUP. It does so by comparing analysis conducted with direct joins to geodemographic information from the original raw in-app data with aggregated subproducts joined in a second time to the targeted statistics. Aggregates at varying scales are created and compared, using a subset of the dataset. Following the findings from Chapter's 3 descriptive analysis (Section 3.3.2) regarding the dataset's bias towards urban centres, this analysis focused on a London Borough. The choice of a smaller sample (a single borough instead of Greater London) helps in demonstrating the dataset's granular potential and illustrates the effects of the MAUP at smaller scales than traditionally documented.

4.1.1. Methodology

4.1.1.1. Creation of multiple aggregates of varying scales and zones

This analysis was conducted on a subset of the original dataset. The subset chosen was the City of Westminster, in London, over the working days of 02-09-2019 to 08-09-2019 during rush hours (7-10am)(Transport For London, 2020). Westminster is picked as it displays clusters of high-density workplace populations, making it especially relevant for weekday rush hour analysis (Berry *et al.*, 2016). The original point dataset is filtered down to only include these dates and times using the data cleaning methodology described in Chapter 3 (Section 3.2.2).

Spatial aggregates of this dataset were made for OSGB grids and census geographies (OA, LSOA, MSOA) for comparison. Using the aggregation methodology detailed in Chapter 3 (Section 3.2.3), activity counts (number of unique devices per area) per hour are created and summed per day, producing higher activity counts, and reducing the chance of having to omit

large amounts of data when aggregating at the smallest scales. This choice was also more consistent with the aim of conducting analysis over three hours of each weekday (daily counts being less significant to analyse rush hour than summed hourly counts). The different scales used, for OSGB and census geographies respectively, are presented in Table 1, and an example aggregated activity count map (at the OSGB 250 m² scale) is illustrated by Figure 1.

Table 1. Scales of aggregation studied

OSGB GRIDS	Census geography output areas
250 m ² grids	Output area (OA)
500 m ²	Lower layer super output area (LSOA)
1 km ²	Middle layer super output area (MSOA)

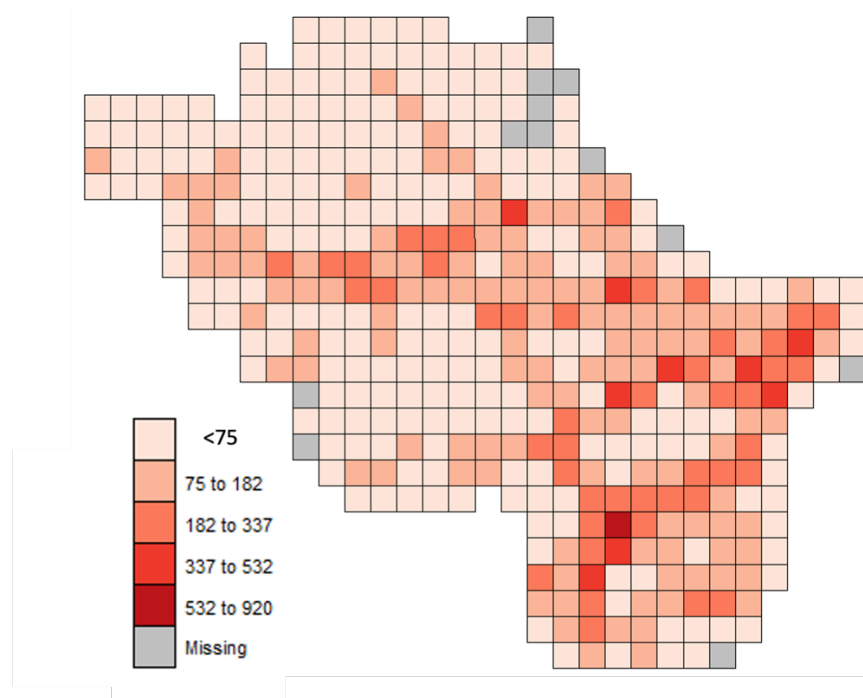


Figure 1. Example of aggregated activity counts at the subset period and scale. The aggregation unit here is the 250m² OSGB grid.

The 1km² grid, the same used for the aggregated subproduct of Chapter 3, was downloaded from Charles Roper's work on OSGB geographies (Roper, 2015). The 500 by 500 and 250 by 250 grids were created in QGIS as divisions of Roper's 1km grid, by dividing the 1km² grid into four 500m² sections, and further dividing these into 250m² units. All three scales follow the British National Grid projection. The OA, LSOA and MSOA political boundaries were downloaded from the London Datastore (UCL and GLA, 2015).

The initial purpose of using OSGB and census geographies in parallel and at various scales was to evaluate them against one another to help establish which of the geographies would be most suited for day-time statistics (as opposed to residential census statistic). The hypothesis for this test was that smaller scales would be a closer fit to a direct linkage using the point data without aggregation, and OSGB cells would perform better than census geographies, as the latter were generated to best fit residential (night-time) statistics.

4.1.1.2. Linkage with WPZ geographies

To test the aggregates' fit to the data, each aggregate is used to conduct the same geodemographic analysis and the outcome of each is compared to results obtained from a control. The control is a direct join of the pre-aggregation raw in-app data points to the geography of the studied characteristic (workplace zone classifications, presented below), whereas the aggregated datasets must be linked to said geography with an extra step. This is similar to the assessment conducted in Chapter 3 Section 3.4.3 which compared the 1km² aggregate to a straight linkage to LOACs. The difference between control linkage and aggregate linkage to the studies statistic is illustrated by Figure 2.

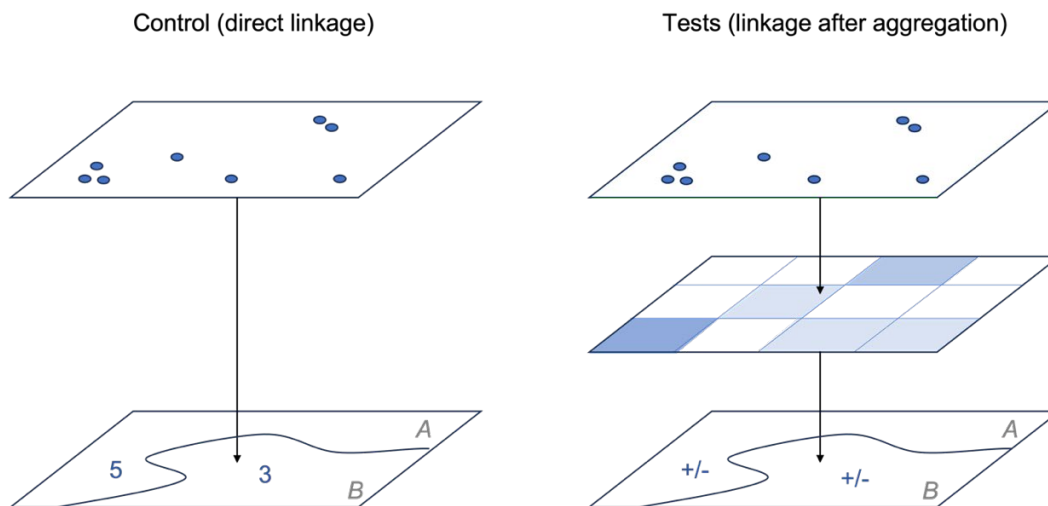


Figure 2. Difference between the control and aggregate tests. In both cases, aiming to obtain the number of activities per studied areas (areas A and B), but in the case of an aggregated dataset (right) this requires overlapping the aggregate zones to the studied areas and assigning the overlap, whereas direct counts can be assigned to the studied geography from the control.

For this analysis, the workplace zone classification (WPZ) is chosen as the studied characteristic. The WPZ is a geodemographic classification used to describe London's working populations and workplace geographies (Singleton et al., 2017). The different groups identified by this

classification are listed in Table 4. WPZ geographies were downloaded from the London Datastore (UCL and GLA, 2017). Each grid cell (and census geographic unit) was assigned the WPZ subgroup which best overlapped the unit. Population weighted centroids linked to workplace counts were used to assign the subgroups, as linking a subgroup to a cell by area overlap alone may not be representative of the underlying demographics and population clusters. Thus, where multiple WPZ subgroup overlapped a given cell, the one with the highest weight (highest population count) was assigned to that cell, ensuring the matching was done by population size rather than by surface area. The centroids were downloaded from the UK government's statistics portal and combined with workplace and worker counts as found in the 2011 census.

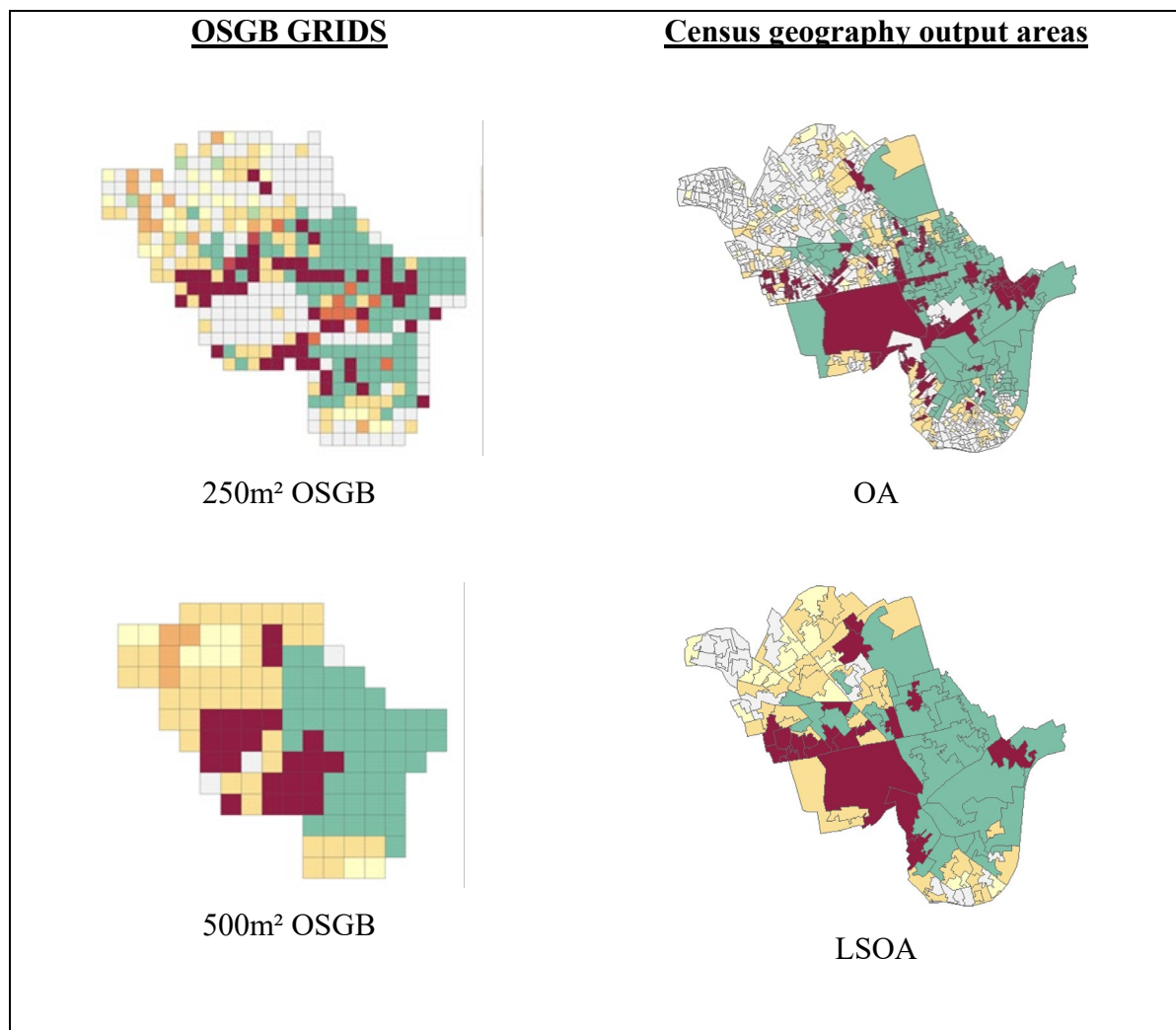
Table 2. WPZ classification subgroups used for this study, as categorised and described in the WPZ technical report (Singleton et al., 2017).

Group – A: Residential Services	A1: Predominantly older, local education and health workers
	A2: Lowly qualified workers in construction and allied local trades
Group – B: City Focus	B1: Dynamic financial centres with extended operating hours
	B2: Professional, retail and leisure Services in dynamic central locations
Group – C: Infrastructure Support	C1: Younger customer service workers in wholesale or retail occupations
	C2: Blue collar, manufacturing, and transport services
Group – D: Integrating and Independent Service Providers	D1: Health care support staff and routine service occupations
	D2: Locally sourced, home helps and domestic or manual workers.
	D3: Travelling or home-based general service providers
Group – E: Metropolitan Destinations	E1: High street destinations and domestic employers
	E2: Accessible retail, leisure, and tourist services

From these linkages, activity counts per WPZ subgroup were computed (number of unique devices registered by WPZ subgroup over the study period in the City of Westminster). This provided information about the differences in activity between working subgroups over rush hour. An example tibble of this data for the 500m² aggregate is showed in Table 3. Figure 3 shows the WPZ subgroup attribution for each aggregate scale and zone tested.

Table 3. Synthetic example of the activity count per WPZ subgroup, obtained for the aggregate made with the 500m² grid (id500m).

id500m	Subgroup	Date	n_activities
3676	High street retail	17/09/2019	10
3678	Health care staff	16/09/2019	23



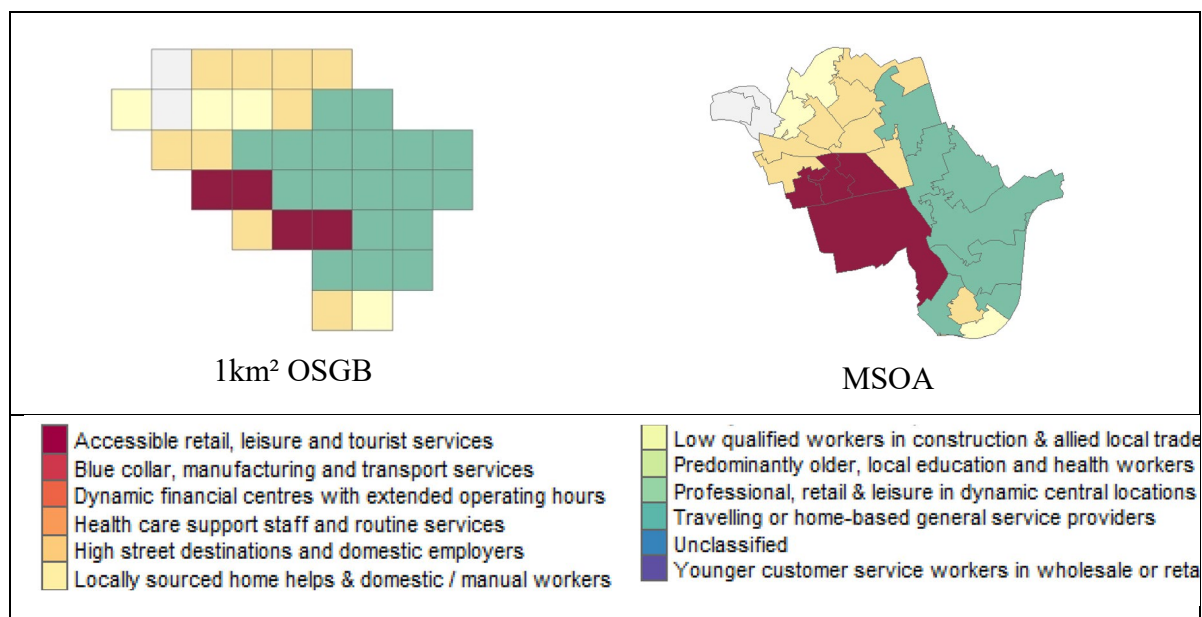


Figure 3. Miniatures for each scale showing the distribution of WPZ subgroup assigned to grid cells or output areas through weighted spatial joins.

4.1.2. Results

Comparing the activities per subgroup over the study period across all aggregated products, the counts vary significantly depending on the aggregate used. This means that, provided with different aggregated subproducts, an analysis on the WPZ subgroup activity of Westminster at rush hours would return different results and population profiles. Figure 4 shows the activity means for the study period per aggregate per subgroup (left). As expected, aggregates present less overall activity than the control, as aggregating activity counts removes repeating data points. Figure 4 (right) shows the percentage of aggregated activities allocated to each subgroup: the 250m² grid cell seems to be closest to the direct WPZ join for their resulting population profiles.

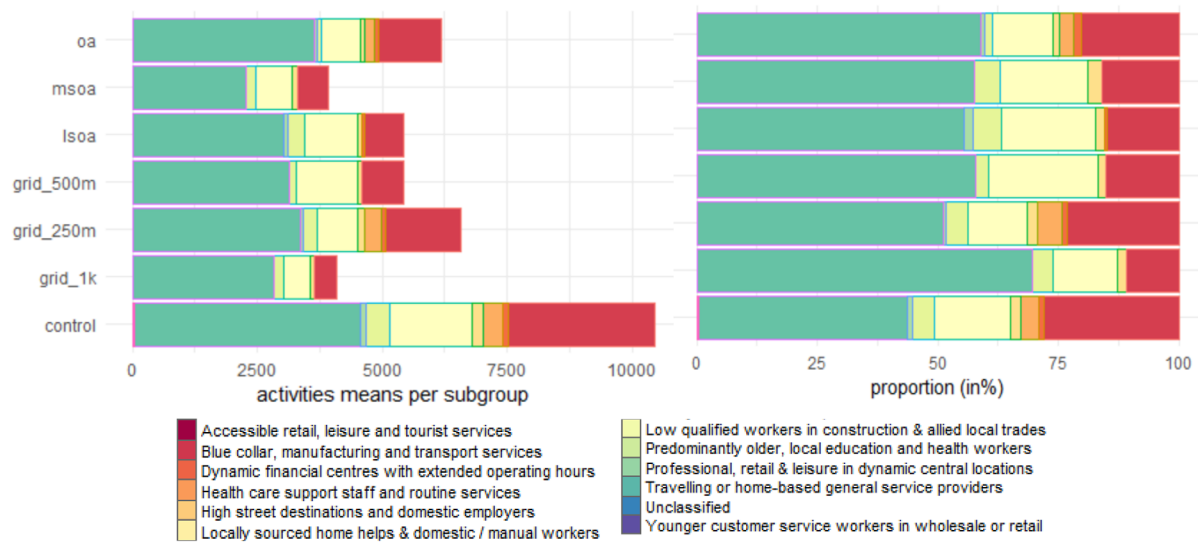


Figure 4. Activity means (left) and subgroup proportion (right).

Importantly, activities proportions per subgroup (right) vary significantly between aggregation types: analysis using the 1km² grids returns a very different impression of Westminster compared to 250m² grids (not the same number, or even distribution of subgroups are visible between the two results).

Lower scales are more granular, but data is omitted due to low counts. However, running a correlation test between the control activity counts per subgroup and the aggregated ones revealed that, despite the data omission resulting from insufficient data events (IDEs) at the low scales (OA and 250m² grid), these smaller scales still present the highest correlation with the control (respectively R=0.96 and 0.98). OAs do not appear to perform better than arbitrary grids cells for day-time analysis concerning WPZs.

In

Figure 5, the 250m² grid WPZ subgroup proportions are compared against the control's, visualising which populations are over-estimated or under-estimated by the aggregated product. Where the colour overfills the bar outline, the subgroup is over-estimated by the aggregate compared to the control, when white remains, it is under-estimated. The 'Professional, retail & leisure in dynamic central locations' group is over-estimated by approximately 10%. Discrepancies could be explained by population densities and dynamism varying between groups. This also shows how the aggregation choices may impact final analysis unequally across space and the dataset.

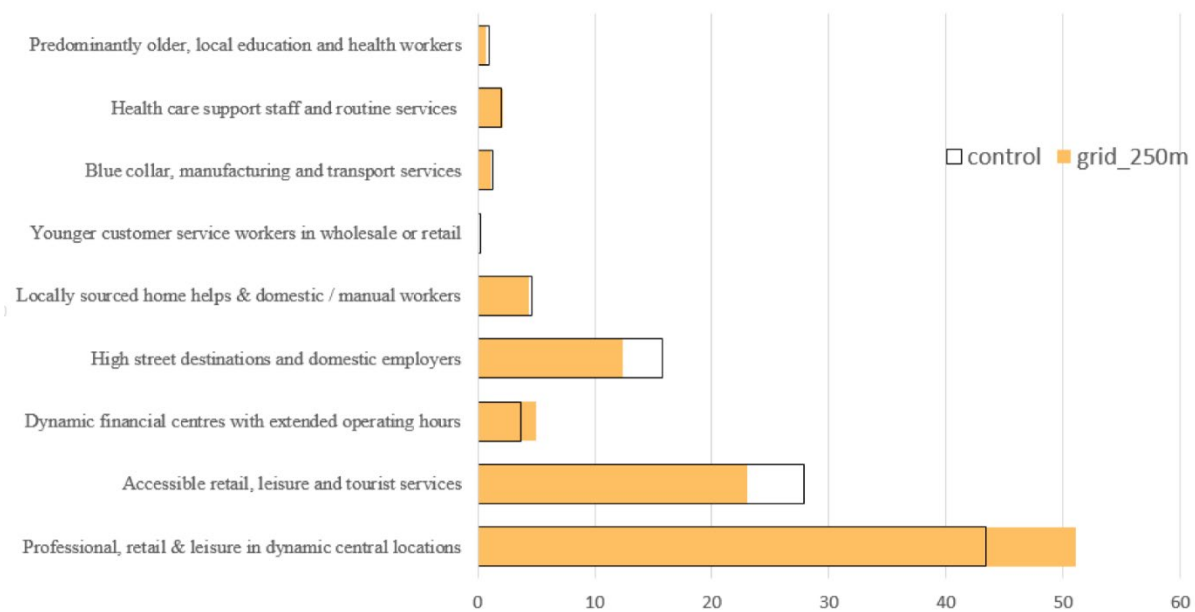


Figure 5. Ratio comparisons between control and 250m².

Though aggregating points to 250m² or OA, for instance, seemed to preserve part of the population profiles obtained by the control, discrepancies in results cannot be neglected, especially as these may deepen when applied to larger scales, or larger amounts of data.

This exploration indicates that scale choice alone changes the distribution of activities and the geodemographic information that may be ascribed to them. It also highlights that the zoning effect also has significant impact on the final analysis, especially when looking at geodemographic classifications which do not necessarily stem from census residential data, such as the WPZ. Here, the grid cells seem to perform better for this “daytime statistic” compared to OA, LSOA and MSOAs at comparable scales. Census geographies were made to be a best fit for residential statistics and thus correlate better with nighttime patterns. Using administrative boundaries such as the OAs may introduce biases to the aggregation process and ‘force’ the dataset into spatial patterns it does not automatically relates to.

This analysis sheds light on this project’s key challenge of using in-app data, identified in Chapters 2 and 3: choosing (or creating) a spatial unit (region) which best represents the data for safe and efficient aggregation. This constitutes a hurdle in the creation of robust, reliable, reproducible, and transparent data subproducts for wider use in research. This assessment was a pivotal point in reframing this thesis’ objective from geodemographic analysis using in-app data towards questions of regionalisation and assessment of MAUP applied to new consumer point datasets.

4.2. Defining the principles and practices for robust regionalisation

Following the results from the previous section, and the discussion sparked, this work proposes the regionalisation of the in-app dataset as a method to mitigate issues of disclosure control and MAUP on analysis. The aim is to create spatial units which would maximise granularity whilst minimising disclosure risk, creating ready-to-use, safe and high-resolution aggregates of the in-app dataset. For this, the principles and important criteria central to the creation of appropriate regions are here introduced and detailed, to direct the regionalisation efforts proposed throughout this work. Then, two established methods of data-driven spatial partitioning are presented, one using quadtree hierarchical algorithms and the second Voronoi tessellations (as touched upon previously in Chapter 2, Section 2.4.3). It is important to note that the resulting regions, (and indeed the need for the approach) are contingent on the data used, and another source of data would yield different results or require different regionalisation adjustments.

To create robust units, a regionalisation methodology must abide to guiding principles (Casado Díaz and Coombes, 2011). These principles inform a set of practices, aiming to obtain units that have specific traits and behaviours answering to stated objectives and requirements. These requirements can be technical, data-led, or informed by the usage intended from the making of the regions. Coombes' work with Eurostat in 1992 defined nine principles which Local Labour Market Areas (LLMA), a form of functional regions, were to meet (Eurostat and Coombes, 1992). They first established the clear definitions of what makes a set of boundaries fit for purpose. These concepts were more recently applied by Casado-Díaz and Coombes (2011) in a critical review of multiple LLMA delineation methods. Borrowing from their structure and vocabulary, the ten in-app data regionalisation principles are separated into 4 main categories: *objective*, *constraints*, *criteria*, and *usability*. Regionalisation tests stemming from this project are later assessed against these principles and practices, with the preferred method abiding to as many of the ten principles as possible. The reproducibility practice (under the *usability* category) was added for this project, building from the discussions in Chapter 2 (Section 2.3.1) regarding transparency of methods and accessibility of tools used in aggregation. Table 4 lists these principles and practices, further detailed under their respective categories below.

Table 4. Principles to guide the definition of in-app data bespoke regions. Adapted from (Casado Díaz and Coombes, 2011).

Principle	Practice
OBJECTIVE 1. <i>Purpose</i> 2. <i>Relevance</i>	To be statistically defined areas appropriate for identifying clusters of activity from in-app data whilst protecting privacy Be temporally relevant, making them relevant for analysis related to temporal data patterns as well as spatial
CONSTRAINTS 3. <i>Partition</i> 4. <i>Contiguity</i>	Every point is in only one area, aiming for >10 devices per area. Each region is to not cross a pre-determined administrative boundary. Each region is to be a contiguous unit.
CRITERIA 5. <i>Homogenous</i> 6. <i>Autonomous</i> 7. <i>Coherence</i>	Terrain homogeneity should be sought after. Underlying terrain is to be accounted for, making the resulting areas recognizable as regions. Outputs aimed to be stable over time for comparative analysis Size range minimised.
USABILITY 8. <i>Conformity</i> 9. <i>Flexibility</i> 10. <i>Reproducibility</i>	Aligned with administrative boundaries (nested within other existing geographies relevant to future analysis). Must perform well for different regions (either cities, or different nested constraints) and comparable point datasets. Must not require specific software or training for use (minimise technical and financial barriers to access).

4.2.1. Objectives

The objectives guide the purpose and relevance behind the regionalisation strategy (Casado Díaz and Coombes, 2011). They seek to justify *why* one should create new units; what *purpose* do they serve, and how are they *relevant* for the research community (what gap in the research toolkit are they filling). As listed in Table 4, this project's objective with regionalisation is to create statistically defined areas which are appropriate for identifying clusters of activity from in-app data, whilst protecting privacy. This is the objective defined throughout this work, both through the literature review (Chapter 2) and the data's exploratory analysis (Chapter 3). The *relevance* principle aims for the regions to be relevant to previous discussions relating to the time-space prism. Concerns of aggregation impacts on analysis are not only spatial, but also temporal. To fulfil the aim of having the regions be *relevant* to the data, as well as *purposeful* of a specific usage, a temporal dimension should be included, making the regions relevant for analysis related to temporal data patterns as well as spatial units.

4.2.2. Constraints

The constraints principle lists the strict limitations the regionalisation process must abide by. The practices under this category tackle questions of *what* the regionalisation should represent. These are restrictions without which the final units could be unusable or not interpretable as regions. These include *partition*, or the idea that every point of data used for defining the region must be present in only one area: no two regions should overlap, and no single data impression should be counted twice as a result of the delineation. Another constraint under the *partition* principle is to aim for more than 10 devices per area. This stems from the aim to control for disclosure risks: without this rule, the regions do not fulfil their key objective. Additionally, each region is to not cross a pre-determined boundary (e.g. administrative), so the future regionalisation methodology can be contextualised geographically within specific study areas.

The contiguity constraint is also key, and in practice means each region is to be its own contiguous unit, and not be interrupted or divided by another region. As explained by Casado-Díaz and Coombes (2011), the contiguity constraint also ensures city centres remain recognizable as such, and distant clusters are not combined into a final region without a geographical relationship.

4.2.3. Criteria

The criteria listed in Table 4 (bullet points 5 to 7) describe desired characteristics for the regions. Criteria have lower priority compared to the objectives and constraints described before. It is difficult to have all nine principles be of equal importance without some overriding the others (Casado Díaz and Coombes, 2011). Criteria thus help differentiate between two methods by privileging whichever method complies with more criteria than the other, as these are selected to make the final units more coherent and practical. *Homogeneity*, as with the census OAs, seeks to maximise similarity within regions and clear separation and heterogeneity between different regions. For instance, homogeneity could be defined as terrain being homogenous within a unit, or as population geodemographic characteristics within a unit to be as similar as possible (Martin, 2010).

Ideally, the regions should also be *autonomous*, insofar as their activity patterns are self-contained. Finally, the units are to be made *coherent* by being stabilised over time (little volatility resulting from the method itself). Coherence describes a region's necessity to be reasonably recognisable (Casado Díaz and Coombes, 2011). A region which drastically changes when comparing different dates of data can lack temporal coherence. Coherence is also addressed spatially in part with the practice of seeking to minimise size range, and obtain regions with comparable sizes, either of terrain or population (Martin, 2010).

4.2.4. Usability

The usability principle encompasses the practices which both ensure that the region methodology is reproducible and flexible, but also guarantee that the output regions are applicable to similar datasets. These are the lowest priority guidelines, but are desirable where they do not negatively affect the objective, constraints, or criteria. *Conformity* requires the output units to be recognisable to users and conform to the known environment. This can be done in multiple ways, such as by nesting the regions within a set of known geographical boundaries (such as MSOAs or boroughs) to ensure the outputs conform to known geographies. In this sense, the conformity principle also assists the *coherence* criteria, helping the regions be recognisable, and the *partition* constraint earlier defined to keep the regions constrained spatially. *Flexibility* is also a concern of usability. Ideally, the units created must perform well in different environments (e.g. different cities or countries) and with comparable point datasets. Applying the chosen method to similar in-app data and coordinates should obtain comparable

results. Lastly, *reproducibility* has been added as a usability principle, to promote an accessible and transparent method, both financially and technically. One of the stakes of aggregating these types of new consumer datasets is to ensure access to a larger number of researchers. Following these goals, the regionalisation methodology should remain transparent and accessible to a larger group. For this, the ideal regionalisation methodology would not require too specific a software, and not rely on high level technical skills to apply and interpret.

4.3. Explorations of existing statistical methods

In Section 2.4.3 of the literature review, two main regionalisation types were discussed: hierarchical and rules-based methods. In the rest of the chapter, two data-driven hierarchical methods are investigated, both of which were applied to point location datasets in the regionalisation literature, whether in the fields of urban studies or physical geography (Dong, 2008; Sevtsuk and Ratti, 2010; Lagonigro et al., 2020; Gu and Shen, 2022). The strategies developed for the LLMA were robust and bespoke regionalisation examples for urban daytime statistics, but relied on crucial census information not present in the in-app data provided, or mostly granular origin-destination travel to work data. Thus, we instead explored data-driven regionalisation methods which allow to identify and represent clusters of static data points rather than flows: quadtree algorithms, and Voronoi-based regions (Cowpertwait, 2011; Molloy and Moeckel, 2017; Gu and Shen, 2022). Both methods are described and tested using a sample of the in-app dataset, and assessed against the key principles listed in Table 4 to identify how they could be improved upon to fulfil the data's regionalisation requirements.

4.3.1. Quadtree based regionalisation.

4.3.1.1. Definition

A quadtree describes a hierarchical subdivision of a plane. It starts with a square containing input data (in this case, in-app data impressions), which is divided into 4 equal squares, each of those 4 subdivided recursively until it satisfies a stopping condition, for instance a threshold number of points is met. Figure 6 demonstrate this process.

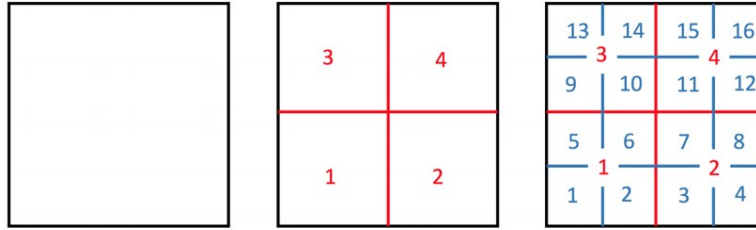


Figure 6. Taken from the Lagonigro et al., *AQuadtree R package description* (See Appendix 2), shows three level quadtree splitting cells. The initial cell is on the left, at the centre is the first subdivision and the second quadtree subdivision is on the right (Lagonigro et al., 2020a).

Quadtrees can be built from various datatypes, including polygons, images, and points. They are used to refine images (pixel levels), but are mostly key in agent-based modelling to identify, aggregate and track agents for model clarifications (D’Angelo, 2016). In GIS, they aim to decompose space into adaptable and easily treatable cells, and have been an established way to optimise spatial searches (Gahegan, 1989; Ebdon, 1992). Transport planning has made use of quadtrees to aggregate mobility data, as demonstrated by Molloy and Moeckel in their 2020 paper (See Chapter 2, Section 2.4.3).

Calling out the “lack of specific methods in R to anonymise spatial data”, Lagonigro et al. (2020a) created an R package which uses a quadtree algorithm to aggregate sensitive point data, such as the in-app data, using a potentially standardised and lightweight method. This echoes this project’s aim of creating transparent strategies for the protection and dissemination of human-generated point data aggregation, using cross-disciplinary tractable methods. The Lagonigro et al. (2020a) R-based quadtree algorithm package was thus tested to assess quadtrees’ performance in aggregating in-app data.

4.3.1.2. Methodology

The open source *AQuadtree* CRAN package contains multiple functions and downloadable test datasets (See *AQuadtree*, Appendix 2). However, the following tests mostly make use of the *AQuadtree* and *plot* functions. The *AQuadtree* function uses a grid coding system following the European Forum for Geography and Statistics (EFGS)’s guidelines for grid datasets (Lagonigro et al., 2020). The function builds an initial grid of a given size, each grid identified and referenced under the EFGS’s coding system. It then recursively subdivides the cells into quadrants, as illustrated by Figure 6 above, if the number of points in each quadrant is above the set anonymity threshold. Not meeting this threshold means the cell is not further divided. In certain cases, disaggregation requires the removal of very low point counts, for which the

algorithm weights the loss rate (how many points must be suppressed) against the lost granularity of the grid that would result from not suppressing (having to keep a bigger cell, and not subdivide where the neighbouring cells would still be above threshold and allow for more division). To minimise the information loss, points that are suppressed are added together, and if this group exceed the anonymity threshold when the divisions are over, they are aggregated into an initial cell and marked as a residual cell. Lagonigro et al., have a more detailed description of this balance work, as well as visualisations of the residual cell process in their package documentation (Lagonigro et al., 2020a) (Appendix 2).

The AQuadtree function takes a point dataset of ‘spatial points data’ format (sp package in R, see *sp* in Appendix 2) and turns it into a quadtree object, which can be plotted by the package’s plot function, colouring the residuals accordingly. AQuadtree has many optional parameters, such as settings for the size of the original grid, the anonymity threshold, the number of subdivisions desired etc. The quadtree objects can be converted as sp or sf objects and saved as such, making them portable and usable in other software (e.g. ArcGIS or QGIS).

To retrieve quadtree objects from the DSH, a conversion to a human-readable format (geojson) is necessary for output check verification processes. After they are retrieved from the secure lab, the products are reconverted to a quadtree object, in order to be able to discern and separate the residual cells, not identified in sf or geojson formats. The AQuadtree package is here applied to the in-app data points to create the maps displayed in the upcoming Results section (Section 4.3.1.3).

Understanding the parameters of the AQuadtree function, as well as their impact on the final output, was important to the creation of robust outputs using this method. The residual cells are a key parameter to set when choosing a threshold level value for the creation of quadtree objects. Different thresholds were tested on a day of data in the City of Westminster to better illustrate the impact of the threshold value on the number of residual cells created. Figure 7 displays three different thresholds (th) for unique devices in Westminster over a week of rush hour in September. The three thresholds compared are 10, 25 and 50 counts.

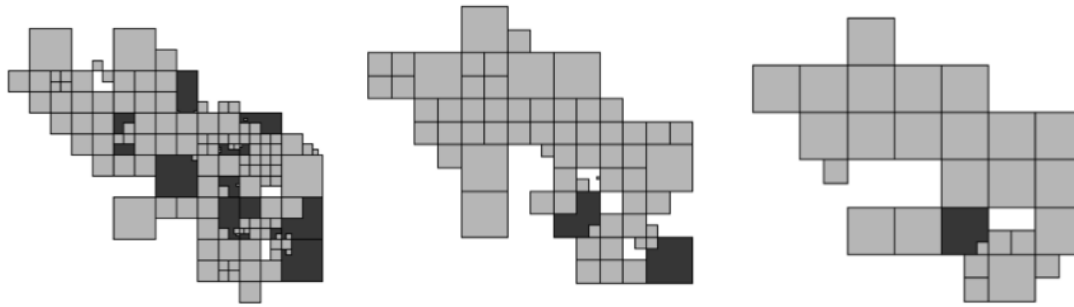


Figure 7. Three quadtree objects created using different thresholds, over the same dataset and period. Th=10 (left), th=25 (centre) and th=50 (right). Black cells are residual cells.

Though threshold 50 only returns one residual cell, its geometries are much less granular than the other values with more residuals. The white square at the centre of the th=25 map (centre), is replaced with a residual cell in th=10 (left). Expectedly, it is seen that lower threshold values return quadtree objects with more residuals than higher thresholds. This is in part owed to lower thresholds (smaller numbers) creating more subdivision of the quadtree, resulting in more potential areas which might leave ‘leftover’ data to fit in a residual cell. Furthermore, a low threshold is easier to reach: it is more likely that the remaining data caused by low counts will be regrouped and accounted for into a residual as their sum is more likely to reach a lower threshold when the process is over. When the remaining data is insufficient, and still below the set threshold, it is discarded rather than turned into a residual cell at the end of the quadtree making process. Higher threshold values thus create quadtrees with more omitted cells which could not be regrouped as a residual. This means higher thresholds not only create less granular geographies, but also omit more of the data than lower, more granular thresholds which create residual cells in place of omitting. In Figure 7, while th=50 creates only one residual, it also removes data altogether in areas that are considered populated with lower threshold values. th=25 and th=10 continue to disaggregate and gives more precision to cells surrounding empty residual space. An increase in the number of residuals is thus not a marker of loss of data or quality in the quadtree outputs: they provide insight that simply omitting counts and treating them as an absence of data would not allow.

For the rest of the methodology, thresholds were thus selected both based on anonymity requirements (<10 devices) and this balance work. A threshold of 100 points was picked to generate the quadtrees made using all device impressions, and quadtrees built from activity counts of unique devices were made with a threshold of 10 devices.

4.3.1.3. Results

The first 2 maps (Figure 8) were made using a subset of the data at the London scale: (left) one day in March 2018 (Monday, 26-03-18) and (right) March 2020 (Monday, 30-03-20). Both days are a Monday, but the 30th of March 2020 contains data collected after the implementation of lockdown rules in the UK (Cabinet Office, 2020). These quadtree objects were made using total counts of events (impressions) with a threshold of 100, rather than unique devices, to display the full activity registered by the unfiltered in-app data on those dates. Black cells represent residual cells and white cells low data coverage. For example, the white patch in southwest London corresponds to Richmond Park, where the data coverage is low.

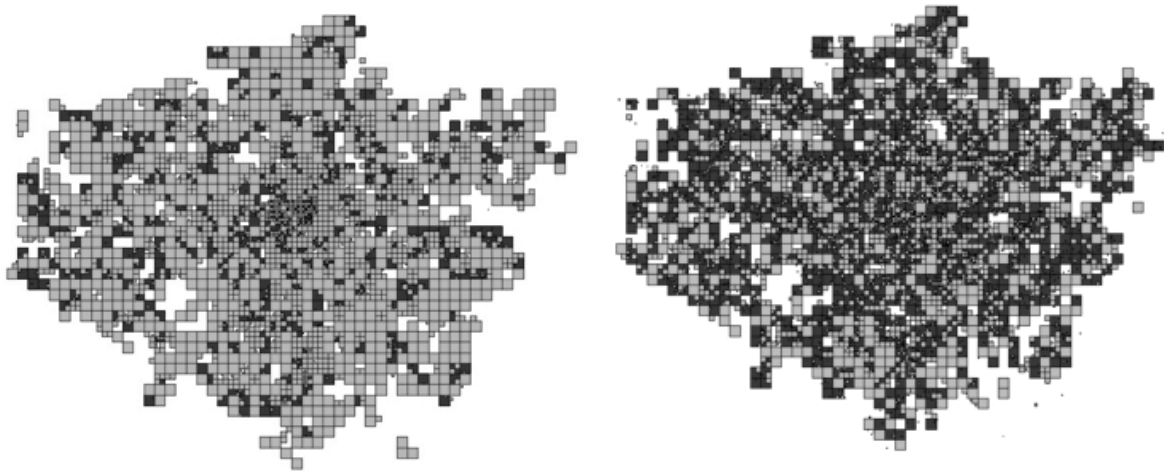


Figure 8. Left - in-app impressions collected across London on the 26th of March 2018. Right - in-app impressions, London, 30th of March 2020. Both maps are plotted using the quadtree hierarchical method with a threshold of 100 impressions per cell minimum.

The 2020 Quadtree (Figure 8, right) presents a larger proportion of residual cells, which are results of low counts or IDEs. By creating and plotting Quadtree objects, we can obtain a quick and non-disclosive glimpse of the data distribution, allowing for rapid comparison between different days and times. Threshold values can also be set to specific variables within a dataset, allowing researchers to efficiently assess the spatial distribution of specific data characteristics (for instance by making quadtrees based on age groups, gender etc. where available).

The next two maps (Figure 9) were made from unique devices activity counts (no repeated impressions per user), at a threshold of minimum 10 devices per square cell. The same dates as Figure 8 are chosen for quick comparison. On March 30th, 2020, London citizens were required to stay home as per COVID-19 regulations, certain businesses were closed and gatherings of

more than two individuals were forbidden (Cabinet Office, 2020). This contextualises the drastic difference between both quadtree outputs, with the 2020 unique devices quadtree (Figure 9, right) being especially sparse. Whether this decreased activity is explained by lockdown measures, or by underlying data-collection processes (changes in devices and apps in the in-app dataset between 2018 and 2020 overall, for instance), the quadtrees provide an overview of a day of data, not requiring other arbitrary decision making than setting the threshold, consistent across the two dates to provide quick comparison.

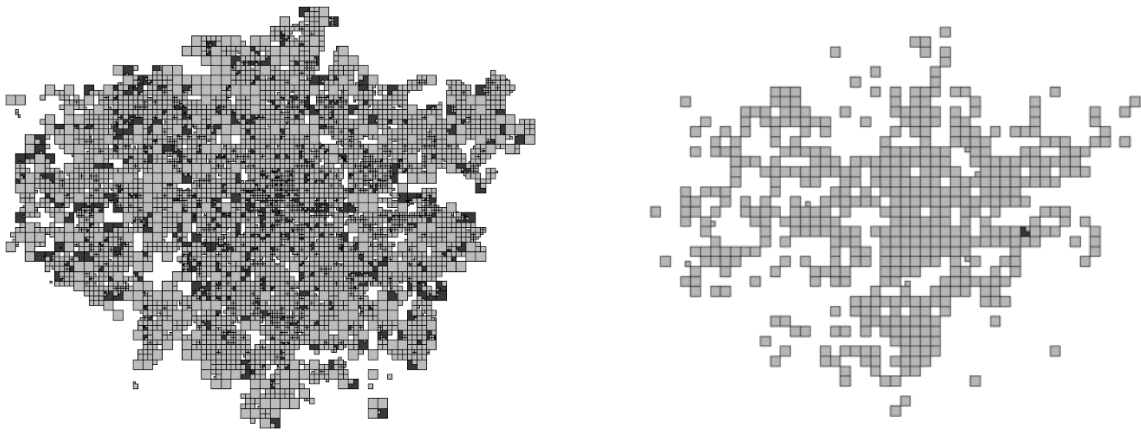


Figure 9. Left - unique devices recorded in London, 26th of March 2018. Right - unique devices, 30th of March 2020. Both maps were plotted using the quadtree hierarchical method, with a threshold of 10 unique devices minimum per cell.

Finally, this quadtree methodology was tested using the same subset of data used in the aggregation analysis of Section 4.1: the week of rush hour data in Westminster (Figure 10). The resulting quadtree regions produce a much more granular level of analysis than seen when using the smallest scale of 250m² grids from the previous assessment. Though inconsistent in size and over time, the cells created with the quadtree algorithm can be much smaller spatially and provide finer detail at the borough scale.

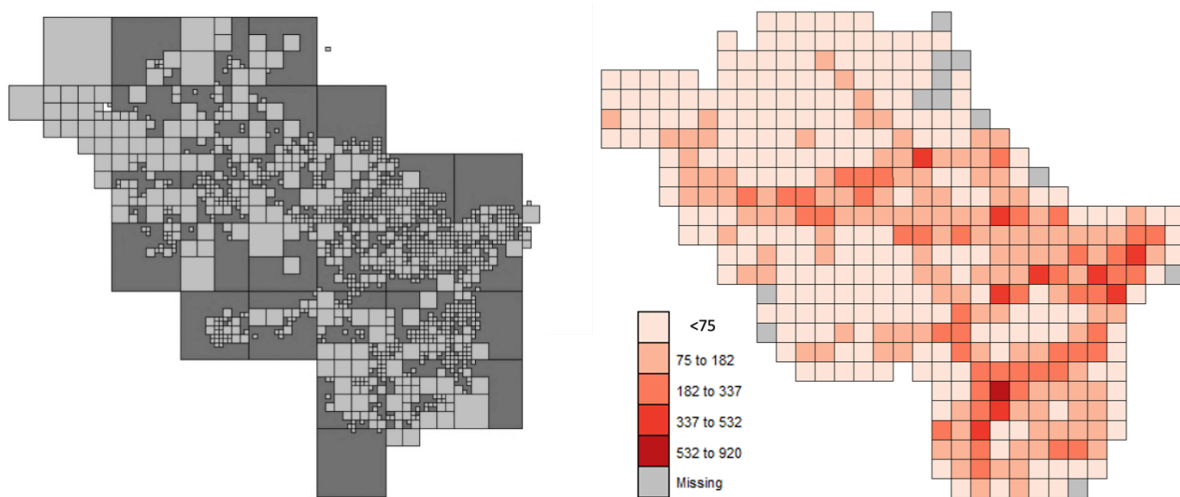


Figure 10. Left - unique devices recorded in Westminster over rush hours (8-10am) over the course of a week in September, aggregated and plotted using the quadtree hierarchical method, with a threshold of 10 devices minimum per cell. Right - activity aggregated to 250m² grid cells over the same data period (Sieg and Cheshire, 2021).

Figure 10 (left) shows how quadtrees help spatially disaggregate high activity hubs differently to traditional grids. Though stylistic choices such as colouring can help distinguish hotspots when using traditional grids, it does not allow for more detail as to where exactly this activity splits within a grid cell itself. This detail is present within the quadtree boundaries themselves. Furthermore, the quadtree object requires less steps than the 250 m² aggregate to make. With pairs of coordinates and the AQuadtree function, a few simple lines of code generate the map, whereas the map from Figure 10 (right) required spatial joins to an ordnance survey grid and coordinate system reconversions where required, and the creation of activity counts per area for colouring. This makes the AQuadtree algorithm a light and efficient regionalisation method for quick spatial inspection of datasets.

4.3.1.4. Discussion

Quadtree algorithms, such as the one provided by Lagonigro et. al (2020) provide an innovative and automated method for anonymising and representing large sensitive datasets. They allow for data-driven spatial aggregation and help inform researchers of the range of scales at which their data can be significant. They have been used extensively for spatial indexing and performing quick operations in GIS and spatial analysis, making them a reasonable candidate for developing rapid data-driven regionalisation methods (Gahegan, 1989; Ebdon, 1992). By creating quadtree maps of London at different dates, we can quickly compare activity levels on

a typical Monday in 2018 with a Monday during the first lockdown of 2020. Removing the need to choose a specific aggregation scale allows the researcher to avoid making assumptions based on previous expectations of the data's behaviour, which is particularly helpful in the context of unforeseen events such as the COVID-19 pandemic. This may also reduce the impact of mislead or arbitrary decision-making in the aggregation process.

However, the quadtree methodology does not answer to all the criteria defined earlier to delineate usable and pertinent regions for in-app data aggregation. To structure this assessment of the quadtree method in answering this project's research aims, we refer back to the principles from Section 4.2, Table 4 point by point:

- (1) *Objectives*. Quadtrees respond appropriately to the objective of identifying clusters of activity from in-app data whilst protecting privacy. They disaggregate the data further than traditional grids can, and allow researchers to display small, granular information alongside areas of lower coverage, without sacrificing privacy. However, their temporal relevance is limited, as they are volatile over time and do not allow for consistent temporal analysis, insofar as the entire regions delineation changes when data is inputted.
- (2) *Constraints*. The partition and contiguity constraints are respected, with the notable exception of the residual cells. The residual cells may not be contiguous: a smaller quadtree cell could be contained inside a residual. As it concerns the residuals, this would disproportionately impact areas of low coverage, making the quality and reliability of the regions' contiguity inconsistent over the spatial distribution of the data.
- (3) *Criteria*. This section of the principles is where most of the issues with the quadtree method appears. None of these criteria are respected by this methodology. The terrain is non-homogenous and the final regions are not easily recognizable as logical divisions of familiar spaces. Molloy and Moeckel (2017) proposed a solution to this by having the quadtree regionalisation process be nested inside administrative boundaries, which would also help respect the partition constraint. However, this does not account for terrain, nor does it make the final areas recognizable as such, since squares cannot be perfectly nested inside irregular shapes. The outputs are also

not stable for comparative analysis. Comparing two different days of data with this method essentially consists of comparing two separate sets of regions altogether, which is much more difficult to calculate and interpret than the comparison of counts aggregated to identical regions. The size range is also not minimised: something that is a perk of the quadtree (having high and low coverage areas be both represented thanks to variance in region size) is here a drawback for region coherence.

- (4) *Usability*. Finally, the quadtrees are usable. They perform well at different scales with varying activities (London-scale or borough-scale as demonstrated above), but also for different types of data, with examples from literature applying them to transport planning and data visualisation (Gahegan, 1989; D’Angelo, 2016; Lagonigro *et al.*, 2020b). The AQuadtree package does not require specific software or training, minimising the barriers to usage and remaining accessible to researchers with rudimentary R knowledge. However, it remains difficult to quantify the impact of the algorithmic processes on results: the nature of the method being hierarchical, the initial groupings of the data into the starting cells also impact all subsequent processes and subdivisions, something difficult to account for in a final output.

Overall, the AQuadtree package is straightforward, lightweight, and open source. It allows for quick insights into large human generated datasets, whilst promoting a more standardised aggregation method. Their use can thus be recommended in preliminary and exploratory analysis, as they are an efficient and simple tool to help inform project-specific choices on scale and regionalisation. However, they may not be a viable solution for disseminating data sample or outputting and communicating results, as they are complex to interpret. They are also not suited for use in geodemographic and mobility analysis, as the quadtree products are dynamic and highly changeable with new data inputs or when created from different temporal scales.

4.3.2. Voronoi

After testing the quadtree methodology, Voronoi tessellation were assessed, as they have been a common method for packaging CDR data, but also a proposed regionalisation method in physical geography and mobile computing (Megerian *et al.*, 2005; Dong, 2008; Sevtsuk and Ratti, 2010; Cowpertwait, 2011; Jiang *et al.*, 2019). This method is here assessed as in-app data is often described as complementary to CDR, and the consideration in disseminating and

aggregating these datasets might have useful overlaps (Bwambale *et al.*, 2020; Kishore *et al.*, 2020). This section describes Voronoi diagrams, and how they may be used to create data-driven regions with the in-app data. The assessment follows the same structure as the quadtree assessment: describing the methodology, providing example regions generated from the in-app data, and assessing these results against the principles defined in Section 4.2.

4.3.2.1. Definition

A Voronoi diagram, also called Thiessen polygons or Dirichlet tessellation, is a partition of a plane into segments close to each of a given set of an object (such as a point). Mathematically a Voronoi cell corresponds to all the points of the plane closest to that object than to any other (Voronoi, 1908; Aurenhammer *et al.*, 2013). Figure 11 shows an example Voronoi diagram made with 8 points.

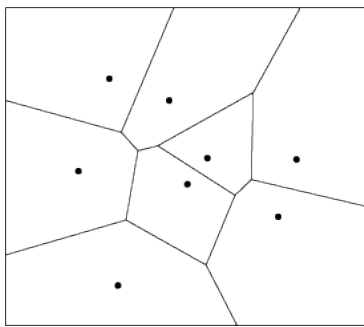


Figure 11. Example Voronoi diagram.

Voronoi diagrams, often observed in natural phenomena, have a wide range of applications: in meteorology they are used to estimate rainfall, in natural sciences they can help predict forest fires, biological tissue development etc. (Longley, 2005; Bock *et al.*, 2010). In spatial computation, they help speed up searches for nearest neighbours, or generalise vector databases (Longley, 2005).

In geography, they have often been used by physical geographers to estimate areas around points of interest, and have recently been applied to estimate the trade areas around retail stores or centres (Longley *et al.*, 2015). Over the past two decades, Voronoi diagrams were also the typical solution for partitioning mobile phone coverage from cell towers (Megerian *et al.*, 2005; Jiang *et al.*, 2019). Regions can be created by grouping points into contiguous and non-overlapping regions defined by these Voronoi tessellations (Cowpertwait, 2011). Building from other research which applied this computational geometry technique to mobile phone data, the

following section assess its potential in generating comprehensive regions for in-app data (Hagen, et al., 2022).

4.3.2.2. Methodology

Before creating the Voronoi regions, the dataset must be clustered to obtain cluster centroids: these centroids become the object around which the Voronoi are constructed (Hagen, et al., 2022). This is both due to large amounts of data making it impossible to tessellate around individual points, and because creating regions with one device per region would be both impractical and highly disclosive, and miss the mark of the regionalisation purpose. For this, sensitivity analysis of the different parameters concerned by the clustering procedure (DBSCAN) is conducted to choose the most appropriate settings for the dataset. From the cluster centroid, Voronoi regions are then generated to aggregate the data to. The resulting Voronoi regions are then compared to 250m² OSGB grid and OA, assessing the counts of data preserved by each aggregate.

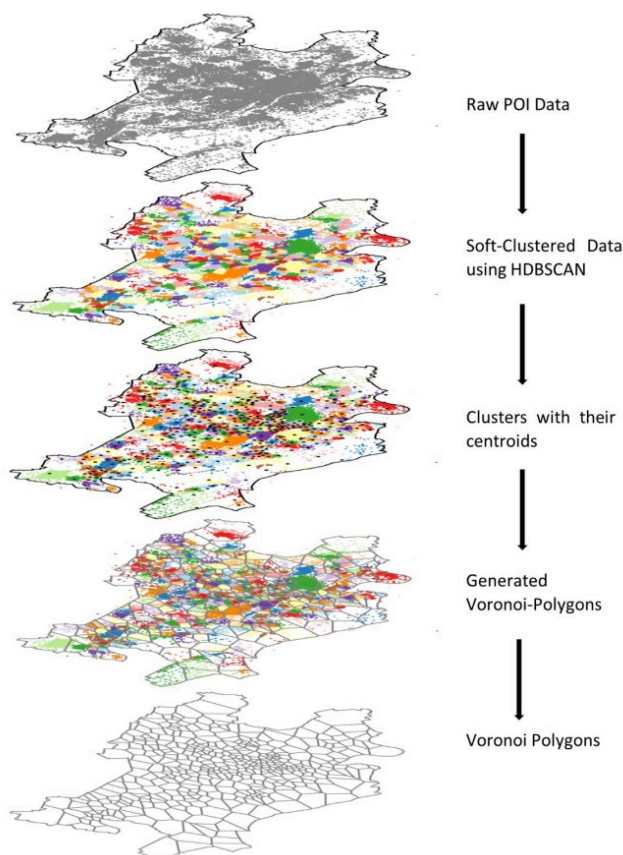


Figure 12. “Procedure for generating Voronoi polygons using HDBSCAN for POI” extracted from Hagen et al., 2022, p.7.

4.3.2.2.1. Clustering the dataset

As the original dataset contains close to 7 billion points, the regions generated for this study are made using a smaller sample of the data. A subset of the points is thus collected in the London Borough of Camden over the month of September 2019. To assess the stability of the outputs over time, one day of data is tested (2nd of September 2019) as well as a stabilised combined week of points (2nd of September to the 8th, Monday to Monday) to compare the resulting changes in regions.

The subset dataset is clustered using the DBSCAN clustering algorithm (Figure 12, step 2). DBSCAN was demonstrated to be a more appropriate clustering technique to detect clusters of atypical shapes, such as along roads for traffic, or islands of data within other clusters (Diker and Nasibov, 2012; Mai *et al.*, 2019). As the in-app data tends to be concentrated along streets, its capacity to distinguish rectilinear clusters makes DBSCAN a field-appropriate choice. Additionally, Schubert *et al.* offer valuable analysis demonstrating the competitive performance of DBSCAN for use in geographic datasets, emphasizing the importance of parameter choice in ensuring the good use of the algorithm (Schubert *et al.*, 2017).

The data is clustered using the `dbscan` function (see *dbscan*, Appendix 2), varying the two main parameters identified by Schubert *et al.*, (2017) in a sensitivity analysis (Hahsler and Piekenbrock, 2022).

- (1) Epsilon (ϵ) sets the maximum distance for two points to be considered neighbours.
- (2) MinPts describes the minimum number of points necessary to make a cluster.

Though both parameters can be informed by field knowledge (for example, MinPts =10, corresponding to 10 devices to cross the anonymity threshold), it is still best practice to conduct a sensitivity analysis to inform the parameter choice (Schubert *et al.*, 2017; Hahsler and Piekenbrock, 2022). 10 would be the most appropriate MinPts value if we were clustering unique devices. However, the dataset contains all events (impressions) for the chosen period at this stage of the process, and one device may be generating multiple points and creating a cluster of its own. ϵ values ranging from 5 to 100 and MinPts value from 10 to 100 are thus tested. The resulting numbers of clusters and noise points (points which are not considered part of a cluster and thus discarded later) are listed in Table 5 of the results section (Section 4.3.2.3.1). The example clusters of the dataset cannot be displayed at this stage due to their disclosive nature, but step 2 of Figure 12 illustrates an example from Hagen *et al.*'s comparable work (2022).

4.3.2.2.2. Obtaining cluster centroids

The cluster centroids are obtained by calculating the mean longitude and latitude for all points of a same cluster. These centroids are used later as the basis for the Voronoi regions (Figure 12, step 3). However, summarising a cluster with a single centroid point means losing some of the cluster's interesting features and shapes, and results in some of the cluster's points not necessarily being considered in the centroid's Voronoi polygon. Having one point to base the tessellation on is crucial for the generation of Voronoi regions, but is a notable disadvantage of this methodology, which is discussed further in the coming assessment (Section 4.3.2.4).

4.3.2.2.3. Voronoi regions

The Voronoi region boundaries are based on the distance between the cluster centroids. They are generated using the R terra package (see Appendix 2), and the resulting tessellation is clipped using the Camden borough boundary. This final clipping step can be done in QGIS or in R, and ensures the regions are nested at the borough level. Applying this method iteratively for each borough, or LSOA could create data-specific regions which are nested within existing geographies.

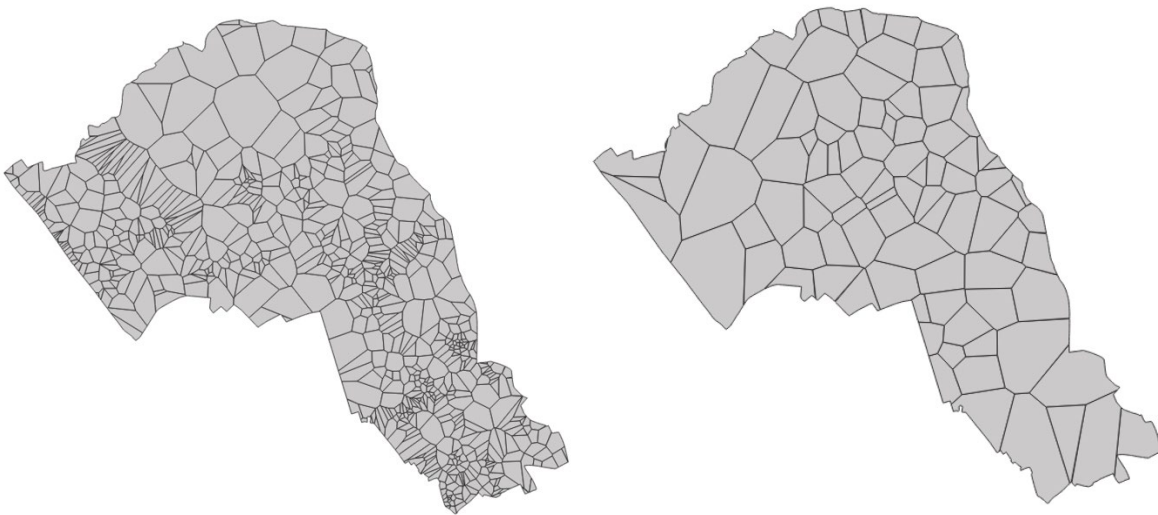


Figure 13. Voronoi boundaries made with cluster centroid. Left - Clustering parameters are $e = 10$ $MinPts = 10$. Right - parameters: $e = 50$, $MinPts = 10$. A smaller value of e when clustering the dataset results in more compact Voronoi regions.

4.3.2.3. Results

4.3.2.3.1. Sensitivity analysis

The number of clusters and noise points generated for varying epsilon and MinPts (mpts) value are compared, both with one day of data (68,270 points) and one week of summed data (864,470 points). The aim was to maximise the number of clusters created (to obtain a maximum of centroid resulting in more compact Voronoi regions), but minimise the loss of data resulting from the discarded noise points. For the day of data, e10mpts10 had the most clusters and e50mpts10 the least noise. e20mpts10, displayed the best cluster to noise point ratio (see Figure 14). This makes e20mpts10 pertinent for the rest of the analysis consisting in assessing the amount of data omitted when aggregating points using these regions. For the week of data, e5mpts10, e10mpts10 and e10mpts20 are selected based on similar reasoning (see Figure 15).

*Table 5. Sensitivity analysis results. *(e=epsilon, MinPts = minimum points). Shaded in yellow, the parameters kept for testing day of data Voronoi, blue those selected for the week of data, green is shared by both.*

*PARAMETERS	1 day of data (68,270 points)	1 week of data (864,470 points)
e 5 mPts 10	657 clusters, 45844 noise points	4803 cl, 162696 npts
e 5 mPts 20	249 cl, 53215 npts	3001 cl, 278510 npts
e 5 mPts 100	23 cl, 63477 npts	684 cl, 628542 npts
e 10 mPts 10	720 cl, 28855 npts	2053 cl, 53237 npts
e 10 mPts 20	345 cl, 41979 npts	1461 cl, 101621 npts
e 20 mPts 10	399 cl, 11498 npts	465 cl, 10341 npts
e 50 mPts 10	91 cl, 1662 npts	12 cl, 319 npts
e 50 mPts 20	83 cl, 4393 npts	24 cl, 1088 npts
e 100 mPts 100	26 cl, 6198 npts	6 cl, 1228 npts

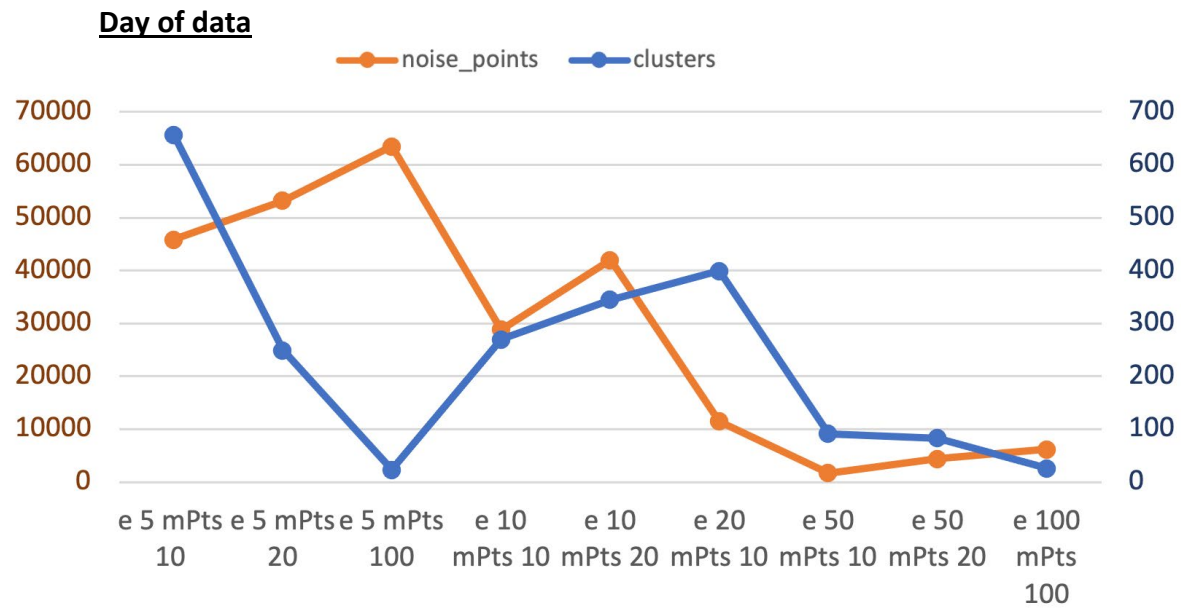


Figure 14. Parameter impacts for the day of data. e20mPts10 is where the line of noise points drops below the number of clusters generated. The aim is to have a maximum number of clusters for a minimum amount of noise points.

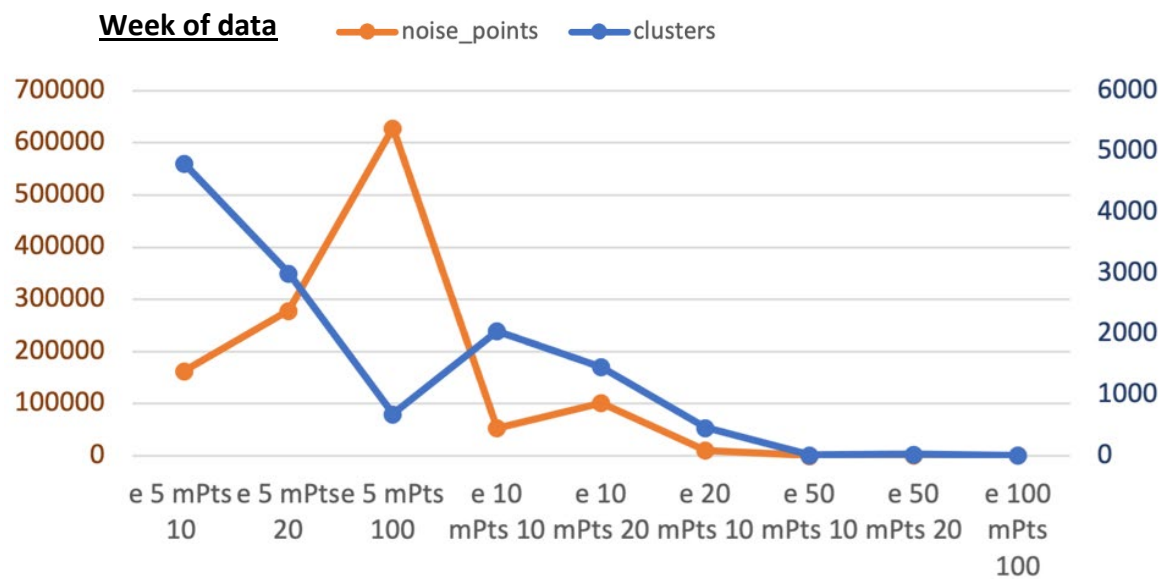


Figure 15. Parameter impacts for the week of data. e5mPts100 presents surprising results with large amounts of discarded noise points and few final clusters. e5mPts10 is the most attractive option, with close to 5000 clusters (which would result in a similar number of regions), and less noise points than the following two options. e10mPts10 and e10mPts20 are still tested for their lower number of noise points.

4.3.2.3.2. Comparison with OSGB 250m² and OAs

From the results of the sensitivity analysis, the clustering parameters which maximise the number of clusters for more compact regionalisation, and minimises the number of discarded noise points are kept (to reduce arbitrary data loss). Out of the 18 parameter combination tested, 6 were retained (3 for the day of data and 3 for the week), for the making of Voronoi regions (coloured in Table 5). The resulting Voronoi regions are used in a comparative analysis against Camden OAs and OSGB 250m² grids. For this, a day of data was aggregated using the 3 types of zones (Voronoi, OSGB, OA). Activity counts per cell were generated, counting the number of unique devices per cell for that day of data. To assess the granularity retained by each spatial unit, the *remaining data* for each is calculated. This number is obtained by adding all activity from the cells which fall under the anonymity threshold of 10 devices per cell and constitute an IDE (*omitted counts*), and subtracting them to the *total counts* for each aggregation scale.

$$\text{Remaining data} = \text{total counts} - \text{omitted counts}$$

This helps assess which regionalisation method preserves a maximum of points post aggregation (a trade-off between the smallest possible scale whilst preventing too many omitted points resulting from IDEs). These notions of *remaining data* and *omitted counts* are reused in Chapter 5, and here serve as a metric to track the Voronoi's performance. The e10mpts10(day) Voronoi regions performed the best in preserving the largest amount of data. Of the 13918 activities, 1028 were omitted, resulting in 12,890 significant counts. e5mpts10(week) comes second. OA aggregation is 5th, with 30% less data than e10mpts10(day), and OSGB 250m² comes 7th, with a 40% data loss compared to the first ranked Voronoi. This means that OSGB 250m² grid preserves less data than the least suited Voronoi made with imperfect parameters.

Table 6. Comparing the granularity and omission counts for the Voronoi regions, OSGB250 and OA.

	Total points (Granularity) (T)	Omitted counts (all counts <10) (O)	Remaining data (T-O)
e 10 mPts 10 day	13918	1028	12,890
e 20 mPts 10 day	4556	91	4,466
e 50 mPts 10 day	9160	643	8,517
e 5 mPts 10 week	21330	9882	11,448
e 10 mPts 10 week	14285	4354	9,940
e 10 mPts 20 week	13629	3485	10,144
OSGB250	8677	732	7,945
OA	10513	1585	8,928

The area sizes of e10mpts10(day) are compared with the OSGB 250m² grids. One OSGB 250m² cell has an area of 62500 m², whereas e10mpts10(day)'s smallest cell is 855 m², with a mean cell area of 30218 m² overall. This means that, on average, the Voronoi regions for these parameters are 30% smaller than the OSGB 250m² grid, and preserves 40% more points. This makes the e10mPts10day Voronoi region the most performant of the zones tested based on granularity criteria (point preservation and small areas). However, the large differences in size between the Voronoi cells come into the way of one of the previously mentioned principles aiming for comparable sizes in outputted regions. This is further discussed below.

4.3.2.4. Discussion

As showed by the sensitivity analysis, DBSCAN parameters greatly influence the output of the Voronoi regions. As such, this methodology is not as data driven as it is driven by parametric choices and good understanding of the clustering methodologies at play. DBSCAN is also one of many options when it comes to finding cluster centroids for the making of Voronoi regions. This adds more complexity and uncertainty to the final results, as they can be heavily influenced by the decision of which clustering method to use. The clip of the Voronoi required to make them fit inside administrative boundaries also creates uncertainty. The boundary effect resulting from the clip means that certain points which should have been part of a larger cross-boundary Voronoi end up being segmented after the creation of regions in ways that are difficult to control for (Fotheringham and Rogerson, 1993; ESRI, 2020).

Clustering the dataset before creating Voronoi is also arguably a form of aggregation. This thus means that this method relies on two or three data-altering steps, something data-driven bespoke regions should be hoping to reduce. Conceptually, data is also lost when it is filtered down to a cluster centroid. As such, though the Voronoi provide a useful context for the creation of regions, especially as the resulting regions performed better than traditional grids in preserving data counts, they do not provide an innovative approach to reducing the MAUP issues described in the literature review and through previous assessments earlier in this chapter.

Going back to the adapted Casado-Díaz and Coombes (2011) principles, we here list the ways in which Voronoi regions do not address many of the important features and specific research needs:

- (1) *Objectives.* The Voronoi do produce statistically defined areas, and help identify clusters. However, their relevance is limited: they are hardly comparable to their surroundings, and do not have a defined temporal dimension. However, the Voronoi can be made with an average of a week of data, for instance, an advantage over the quadtree methodology. This means Voronoi regions could be made stable over time if made with an average sample.
- (2) *Constraints.* Every point is in only one area, respecting the partition constraint. However, it is possible that not every point is in the Voronoi belonging to its cluster's centroid post-aggregation, as that depends on the compactness of clusters and their spread from centroid. The Voronoi regions cross pre-existing administrative boundaries. There is a possibility to clip Voronoi to respect specific boundaries, but this happens after the creation of the regions, not as a parameter in the making, which exacerbates the boundary effect, and makes it hard to control for region size after clipping occurs (Fotheringham and Rogerson, 1993). Voronoi regions are contiguous.
- (3) *Criteria.* The Voronoi regions do not have an inputted metric of homogeneity and do not correspond to underlying terrain. This thus does not fulfil the criteria of creating regions which are recognizable as part of their surroundings. In Figure 13, Finchley Road was composed of multiple horizontal Voronoi polygons going along the road: this does not correspond to what one may generally recognises as a road, geographically (a single long object, in this case vertical, rather than a series of stacked horizontal objects.). Outputs are more stable over time than they were with the quadtree method, which constitutes an improvement in that regard, especially as they can be made from an average count, unlike quadtrees.
- (4) *Usability.* As they are not aligned with administrative boundaries, they are hard to link to existing geographies for analysis. The method is reproducible thanks to readily available packages for clustering. However, many steps are difficult to make transparent and review externally, due to the sensitive nature of the dataset. The method is translatable to other types of point datasets, though the sensitivity analysis and parameter choice will differ from study to study.

4.4. Summary

The quadtree and Voronoi case studies proposed two ways of segmenting point data into areal regions. They proved more performant than OA and OSGB in preserving data counts, or in displaying hotspots of activity at a glimpse. That said, based on the set principles of regionalisation, the statistical solutions provided by either quadtrees or Voronoi do not address some larger issues identified with current regionalisation efforts seeking to address MAUP issues. This is due to their instability over time and the difficulties of using them in parallel to existing boundaries. The temporal dimension, present in both the region objectives and criteria (see Table 4), is also largely unanswered by these methods. Stability over time relies on the creation of 'baseline' regions, where aggregated products could be compared to one another over the same set of regions. Thus, data-driven methods which cannot be generated from a baseline of the data create region too volatile to be stable over time.

Simply segmenting space efficiently does not suffice in making what has been defined as “usable and useful” regions. Segmenting space in ways that maximise homogeneity within each unit, allows for stability over time, and respects underlying terrain and known features, would provide better outputs for aggregation and data dissemination. The tests performed in this chapter allowed useful exploration of some of the key technical challenges. Both quadtree and Voronoi displayed some of the desired characteristics of usable and efficient bespoke regionalisation methodologies. Their main perk resides in their easy access and reproducibility, largely due to portable R packages and straightforward functions. Moving forward from those tests, the aim is thus to create a new methodology which would emulate much of what quadtree and Voronoi functions did well, but find a way to stabilise the output over time, and make it much more representative of underlying terrain, as well as linkable to other geographies. The OA methodology detailed in the literature review (Chapter 2), a rules-based method, answers a lot of these concerns. They start from atomic units (building blocks) that are repackaged based on sets of considerations (Cockings *et al.*, 2013). The making of a similar initial atomic unit for in-app data could also help create average counts of activity per small-scaled unit, helping in making stable outputs over chosen periods of time. The delineation of OAs was also developed with a specific dataset in mind: where the hierarchical methods demonstrated in this chapter are data-driven, the specific methodologies were designed to be general, and thus not in-app data specific. To create bespoke regions, a bespoke rules-based methodology should be considered.

5. Operationalising H3: development of a bespoke regionalisation methodology

The previous chapter outlined guiding principles for the making of useful bespoke regions. It tested some existing tractable data-driven methods and assessed their good fit to these key principles. Drawing from the results of the quadtree and Voronoi regionalisation tests, this chapter seeks to propose a rules-based method, inspired from the OA delineation methodologies, which better answers to the objectives, constraints and criteria outlined. The regionalisation developed and tested in this chapter seeks to emulate the positive traits of the previous methods proposed: both quadtree and Voronoi were easy to implement thanks to reusable functions and packages, were data driven, and statistically relevant. Additionally, the issues highlighted by both methods are here addressed, by proposing an output which is stable over time, is linkable to other geographies without clipping constraints, and is more representative of underlying geographical features.

Thus, this chapter first outlines the methodology for the making of these bespoke regions, starting at the choice of an appropriate atomic unit (comparable to the building blocks for the OA delineation method) (Cockings *et al.*, 2013). After detailing the regionalisation algorithm, the regions are compared against common geographies used for aggregating data, to assess their capacity to preserve data counts and geographic granularity. A sensitivity analysis is then conducted to evaluate the methodology's volatility and the impacts of parameters and data inputs on the final outlines. Finally, going back to the adapted Casado Díaz and Coombes (2011) principles, we address the important criteria aimed to be met by this methodology, beyond count preservation alone.

5.1. Methodology

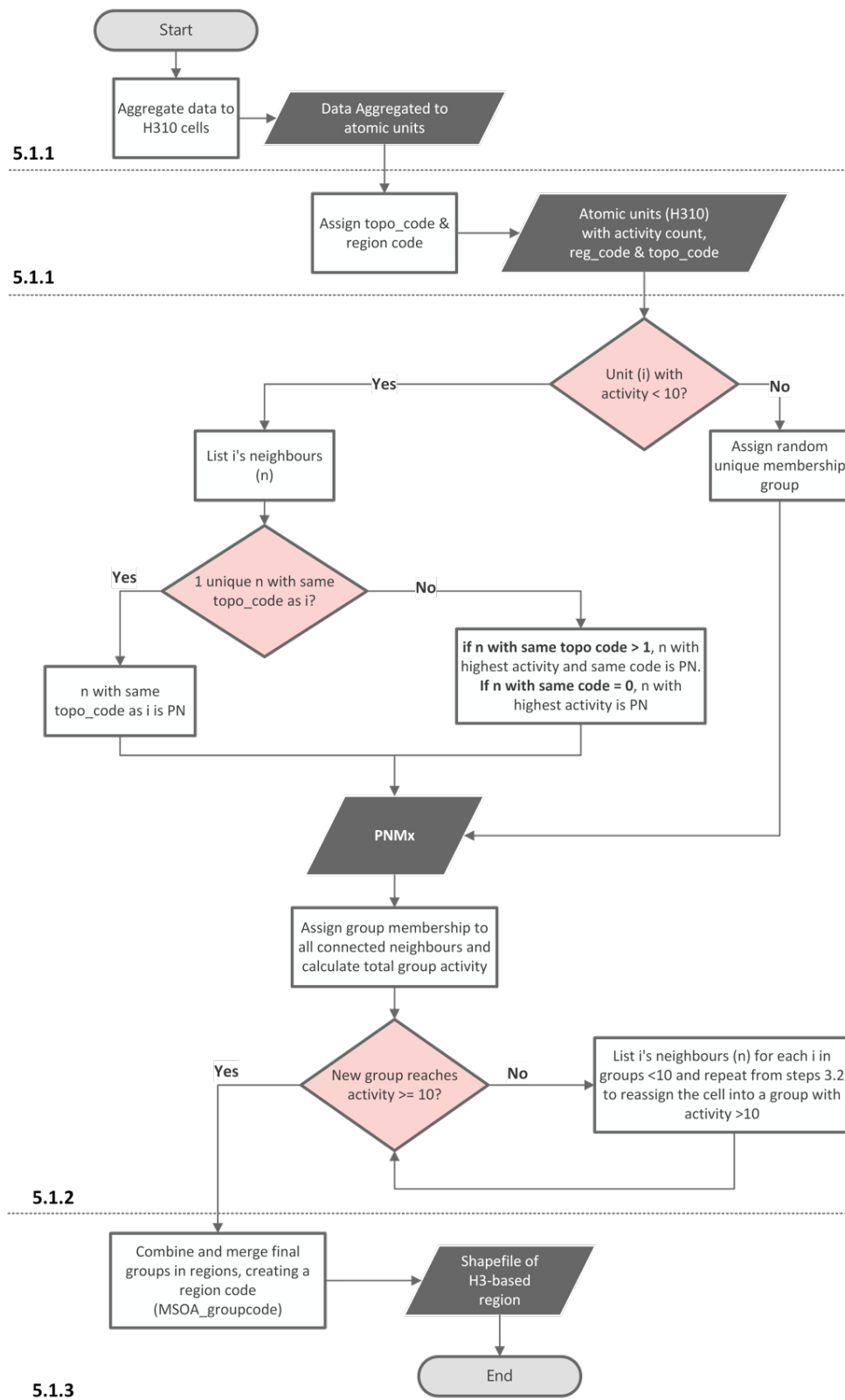


Figure 1. Flowchart of the regionalisation algorithm. PNMx signifies Preferred Neighbour Matrix. The divided parts (dotted lines) correspond to the methodology sections of the same number for further detail. This process is repeated for groups of hexagons within pre-selected merge boundaries (MSOAs) to obtain outputs nested in existing geometries.

The methodology consists of first creating atomic units containing in-app data, attributing contextual information to each unit, and combining areas of low (disclosive) counts into larger homogenous regions. Figure 1 provides an overview of this process, with each step detailed further throughout this whole methodology section.

5.1.1. Making of the atomic units

To meaningfully aggregate location data, it is useful to have an atomic unit which becomes the building block for larger regions (Cockings *et al.*, 2013). This is the strategy used in the making of the OA. To reproduce this using in-app data, we seek to aggregate the data points to a small-scale unit to later further recombine. All tests in this methodology are made on a subset of the in-app data, only including events recorded in the Greater London area over the month of September 2019. The Greater London study area was selected due to data availability and consistent coverage throughout the data period (See Chapter 3), and for consistency with the analysis conducted in previous chapters, which also focused on Greater London.

For the making of atomic unit, the H3 hierarchical spatial index is chosen (first introduced in Chapter 2). H3 tiles the globe into hexagonal areas at a range of resolutions (Brodsky and Uber Technologies Inc., 2015; Bondaruk *et al.*, 2019). The highest resolution (15) provides hexagons of an area of 0.895m² (~0.58 metres edge length), with the lowest resolution hexagons (0) being larger than 4 million km² (~1281 km edge length). H3 was here chosen for its versatility and speed in indexing large amounts of geometries, and for its flexibility in the choice of scale. Hexagons also allow for easy radius approximation. Uber technologies motivate using hexagons to “bucket” their app events by highlighting that polygonal zones around areas, such as postal codes are subject to change in way unrelated to their own data (Brodsky and Uber Technologies Inc., 2015; Stevens, 2006). Bespoke zones similar to postcodes and street blocks might require frequent updating as cities change rapidly. They also present most advantages when aggregating census data (Cockings *et al.*, 2013). Grid systems such as the H3 do not align to streets or neighbourhoods by default like postcode do, but they can efficiently represent neighbourhoods when clustered (Brodsky and Uber Technologies). The h3 R package that supports the geometry also provides documented functions for working with the hexagons’ indexes, which for instance enables the rapid extraction of a list of each hexagon’s neighbours or the calculation of ring distances to other hexagons (Appendix 2).

5.1.1.1. Assessing H3 resolutions

For the making of the atomic units, we first determine which scale of H3 hexagons to aggregate the in-app data to. Every pair of coordinates from the in-app data can be assigned the H3 index corresponding to the location, making for a rapid attribution of the many location points to hexagons. This is done using the `geo_to_h3` function of the `h3` package presented above (see *h3*, Appendix 2). Table 1 provides the H3 resolutions and the average areas and edge lengths of their hexagons, from largest (0) to smallest (15), as extracted from Uber’s H3 documentation (Brodsky, 2018).

Table 1. H3 resolution statistics: average area and edge length of hexagons per resolution. Source: Brodsky, 2018.

H3 Resolution	Average Hexagon Area (km ²)	Average Hexagon Edge Length (km)
0	4,250,546.8477000	1,107.712591000
1	607,220.9782429	418.676005500
2	86,745.8540347	158.244655800
3	12,392.2648621	59.810857940
4	1,770.3235517	22.606379400
5	252.9033645	8.544408276
6	36.1290521	3.229482772
7	5.1611923	1.220629759
8	0.7373276	0.461354684
9	0.1053325	0.174375668
10	0.0150478	0.065907807
11	0.0021496	0.024910561
12	0.0003071	0.009415526
13	0.0000439	0.00359893
14	0.0000063	0.001348575
15	0.0000009	0.000509713

To compare the performance of each H3 resolution and determine the appropriate atomic unit for the regionalisation methodology, we first define what we understand by “data preservation” to quantify the comparison. The definitions of *total counts*, *remaining counts* and *omitted counts*, earlier introduced in Chapter 4 (Section 4.3.2.3.2), are thus further detailed and developed into a comprehensive metric below.

Smaller units collect more data overall, as unique counts may be registered in more areas as users cross over boundaries: over a day of data, one device would be registered only once if we aggregated the activities at the London scale. However, splitting space in more areas results in more *total counts* as the same device may travel through different regions over the period and

be counted as a “unique” device more times. As a static device does not cross region borders, it is counted only once regardless of spatial splitting, resulting in the increasing number of total counts being used as a potential proxy for recording movement through the city.

We progressively divided the study period of London with increasing H3 resolutions, and counted the unique devices registered, assessing the increase of total counts based on areal subdivisions. Figure 2 illustrates this at the London scale for H35 (hexagon size $\sim 252.9 \text{ km}^2$) through to H311 ($\sim 2.1 \text{ km}^2$). The in-app data not being accurate at the metre level (see Chapter 3, Section 3.1.2), we could not pretend to operate at the precision level of resolutions 12 or below (9 metre edges). Resolution 5 (8km+ edge) is the highest resolution acceptable for a significant assessment. Anything above H3 size 5 would not qualify as aggregating the data, as one resolution 5 tile can already be roughly the size of an entire county (Norfolk, for instance, is smaller than a single H3 5 cell) and only divides London in a handful of hexagons (see Figure 2)

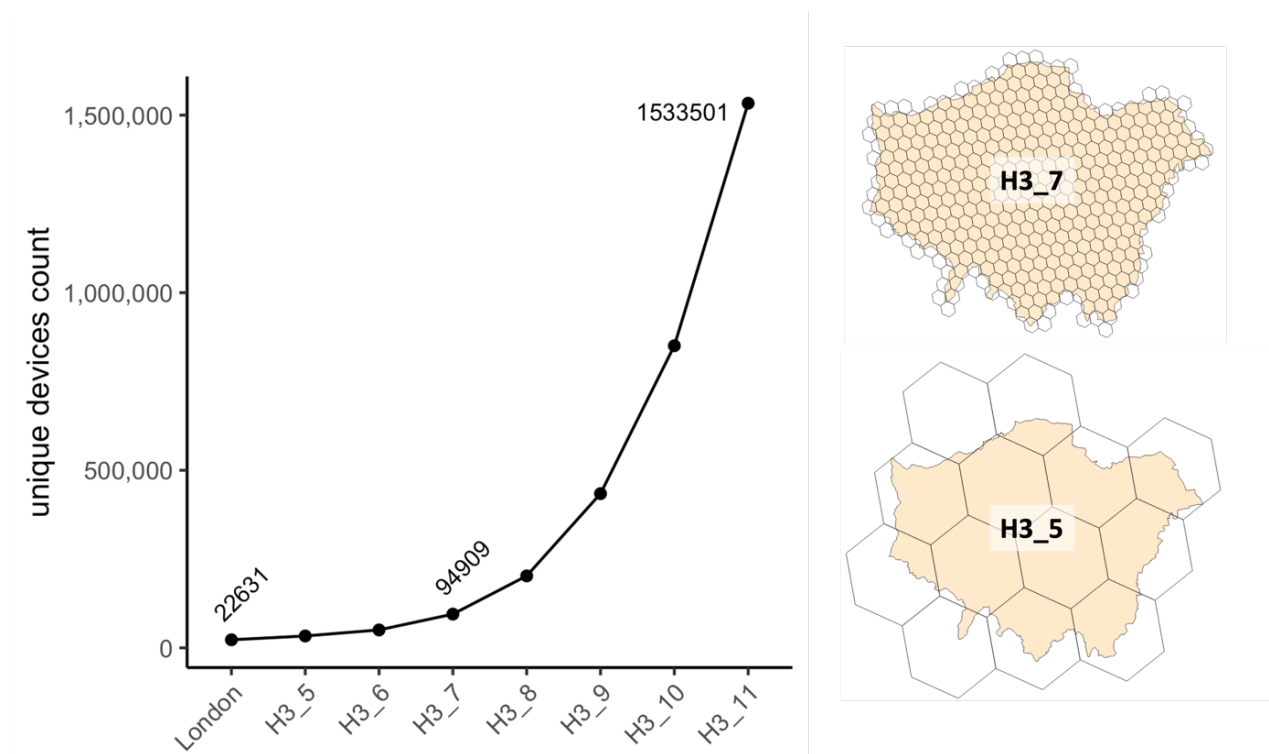


Figure 2. Increase of unique devices (total counts) by scale, with examples of spatial divisions of London for H35 and H37.

To pair with the assessment of total counts, we assessed the number of counts which would be omitted at each resolution, due to IDE. This helps make a trade-off between having the smallest size possible, which would maximise the number of points but have very few points per cell,

and make sense of the atomic unit by having fewer counts but more counts acceptable at the atomic unit. Using higher resolutions also implies more computational overhead, something which can be assessed by performing this comparative exercise against resolutions.

In order to assess resolutions 5 to 11 throughout the dataset for all points, function was written which, for each point of data, assigns it the corresponding resolution 11 index, counts the number of activities per H3 cell (number of unique devices in the tile) and returns a data frame with the activity per cell and the H3 resolution number. The function then uses the `parent_to_h3` function from the `h3` package to find the corresponding resolution 12 tiles for each line and counts the activity for this resolution. This goes on until activity is counted at resolution 5. At each iteration, a file with the activity levels for the given H3 resolution is written, before moving onto the next resolution.

The second step of the H3 resolution assessment process loops through the intermediate files, sums the total activity for that day of data across all H3 of the same resolution, and adds all the counts of units below an activity count of 10. Any hexagon which contains a count below 10 would thus need to be censored, and its content would be considered as omitted data. However, the bigger the tiles, the less unique devices are registered in total, reducing the final count (*remaining counts* post omission). Figure 3 shows the resulting counts for resolutions 5 to 14. Here, the total unique counts are expressed as a percentage of the maximum number of aggregated counts (considered as the *total counts* at resolution 14), and the *remaining counts* are a percentage of the resolution's total count. For example, resolution 6 keeps almost all its counts (very few are omitted due to IDE), but it only picks up on 5% of the counts registered by resolution 14. The intercept is between resolution 10 and 11, implying a shift between keeping a large number of points, with low granularity, and having high granularity but a significant amount of points omitted due to low counts.

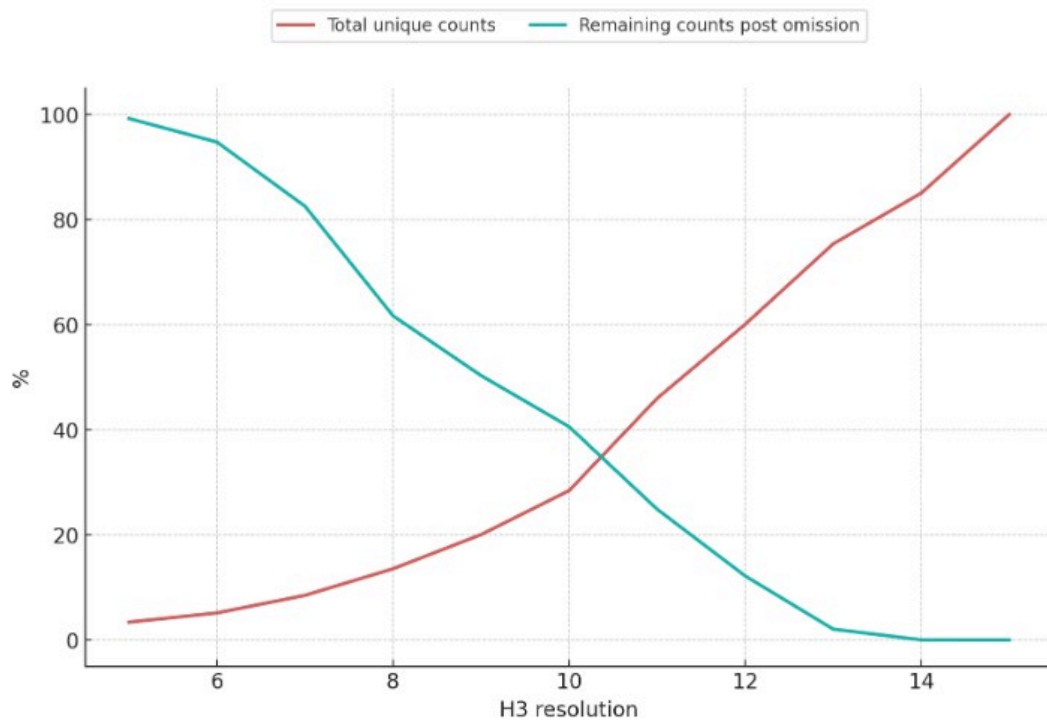


Figure 3. Proportion of total counts (pre-omission) for each H3 resolution compared to resolution 14 total counts, plotted with the proportion of counts remaining post omission (due to IDE) at each resolution.

It is notable that the increase of *total counts* with higher resolutions is likely the result of the same device being duplicated, as it travels through different areas and is considered unique by each. The duplication of the same device may or may not be desirable depending on the aim of each analysis, and there is no consensus about the number of “desirable” duplicates. However, as stated in the regionalisation aims, the hope is to maximise the number of points detected to get a sense of movement over the course of the day and to demonstrate the aggregation methodology’s potential in provided aggregates as close as possible to the original data. Thus, resolutions 10 and 11 appear to be good candidates, both due to their relatively small sizes making them acceptable atomic units (such as building blocks were for postcodes and OAs), and due to the statistical assessments demonstrated (Figure 3).

Resolutions 11 and 10 were then assessed for computational performance. At the London scale tiling 116,555 hexagons to cover the region with H310 takes seconds. This takes 10 minutes for 699,330 H311 hexagons (on a single threaded Intel 64 processor). H310 is thus selected, partly due to computational performance (processing at higher resolutions being very costly, with an exponential increase in cells) (Brodsky, 2018). With an edge of roughly 76 metres and an average area of 15,047m², resolution H310 remains very granular and outperforms the others at the trade-off of efficiency and granularity.

5.1.1.2. Aggregating data to the atomic unit

After having selected the atomic unit scale and filtered a sample of the dataset, we package our data points in H310 hexagons to create the atomic units. For this, we consider a device as analogous to an individual and count the number of unique devices per H310 cell for each day in September 2019. This returns an activity count per H310 cell per day in September 2019. September 2019 was picked as it was the busiest month for the year, with the least internal skew (the median activity and mean activity throughout September were the closest than for any other month, with only a 1% difference). 2019, as discussed in Chapter 3, was also the most stable year for the specific in-app dataset used: after the disruptive changes visible in 2018, and before any potential 2020 COVID19 events.

We calculate the mean activity count for all days, obtaining the average activity per H310 hexagon per day in September 2019, across Greater London. The activity count thus represents the total number of unique devices counted in each area for a typical day, removing duplicates: a person who may visit the same area twice over the studied time-period (here, a day) only features once in the activity count for this area and this time. Activity count is then not a proxy for dwell time, but a count of individual visits to an area over a day. These steps correspond to the very first stage of the methodology flowchart (Figure 1). Figure 4 depicts the distribution of the data subset at the atomic unit level, at the Greater London scale. Its uneven distribution emphasises the need for bespoke regions, and highlights areas of low counts which would benefit from further aggregation. Roads are heavily represented in Figure 4, as many individuals may pass through H310 cells over roads and add up to their activity count as dwell time is not accounted for. The metric of activity count seems thus to be more of a proxy for movement throughout the city throughout the day rather than representative of specific hotspots.

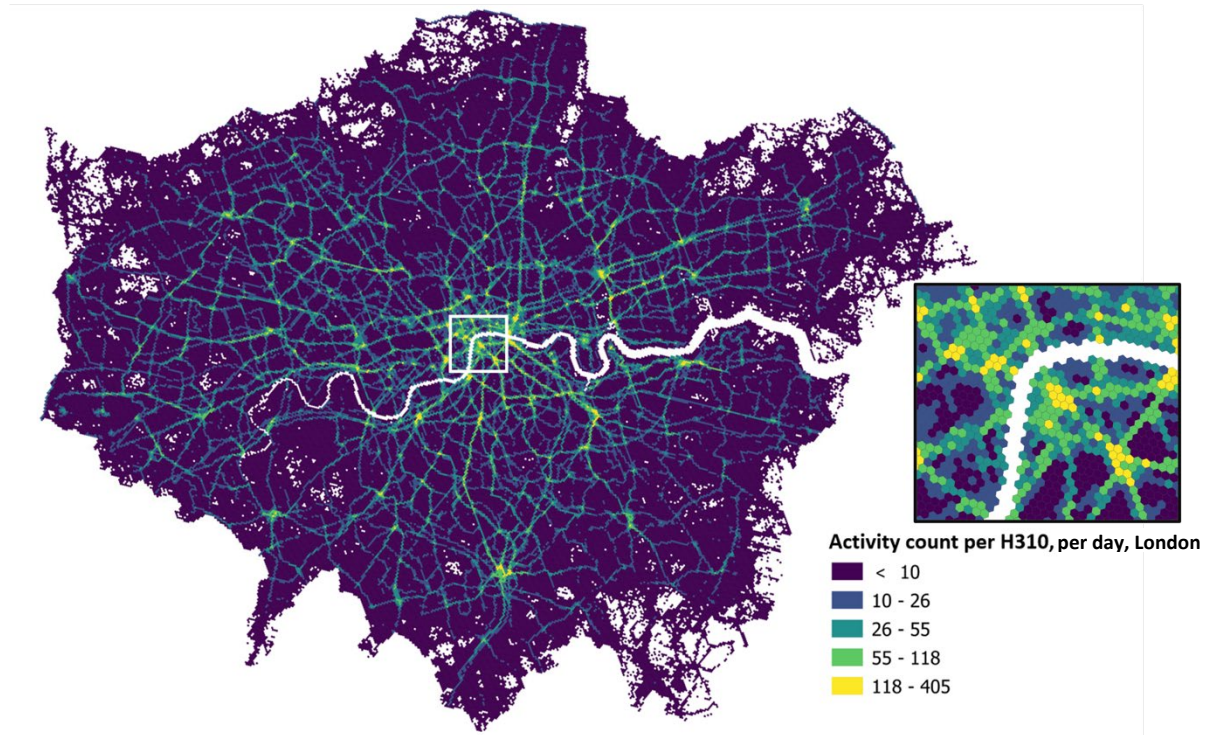


Figure 4. In-app data distribution at the London scale, September 2019. The Activity count per H310 per day corresponds to the number of unique devices per area. The Thames is white as the H310 are nested inside London MSOAs which are typically clipped alongside the river. Other white hexagons correspond to H310 which never contained in-app data throughout the sample period.

5.1.1.3. Assigning contextual information to the atomic units

At this stage, we have successfully created specific atomic units to further combine into bespoke regions. We now need to assign contextual information to each unit to inform their combination into larger homogenous regions. Where census regionalisation methods include inputs such as demographic and economic layers to inform a homogeneity metric, the focus of this work is to propose a flexible approach, which later can be adapted to include more input layers depending on the nature of the data used and aims for the regions. We thus propose land-use as a case example for rule-based homogeneity, and demonstrate how to nest the regions into administrative boundaries to promote linkage. The following preliminary steps are thus needed to inform the regionalisation:

- (1) Assigning administrative information to ensure the regions are nested within existing geographies (MSOAs) (Section 5.1.1.3.1).
- (2) Assigning terrain information to homogenise the final regions (Section 5.1.1.3.2).

These steps help inform the rest of the process to create regions that account for terrain and are nested into existing geographies. Using this contextual information, we later assign each cell to a *preferred* neighbour based on these characteristics: sharing the same terrain and administrative profiles, and those with the highest population. Hexagons with low data (counts below 10) are then merged with their preferred neighbour to create regions of higher, non-disclosive counts, an iterative process further detailed in Section 5.1.2 (Preferred Neighbour Assignment).

5.1.1.3.1. Assigning administrative boundary information

5.1.1.3.1.1. LSOA and MSOA scale comparison

As discussed earlier, new regions are most useful if they can be linked to existing geographies for analysis (Walford and Hayles, 2012). To generate regions that can be nested within census geographies, we select zoning merge boundaries that ensure only H310 cells within the same census geography are combined. MSOAs and LSOAs are census geographies frequently used for sharing census data in England and Wales, and their widespread use make them valuable regions for nesting. They could, for example, allow the aggregated data to be compared to demographic statistics or used in policy analysis, amongst others (Martin, 2010). To choose the most appropriate scale, we compare the in-app data distribution within MSOAs and LSOAs. If the final regions are to be constrained to specific geographies, the question is asked of whether there is enough data per LSOA or MSOA to repackage data within them. We thus use the same methodology as described in the selection of the atomic units (Section 5.1.1.1) to aggregate our data to LSOAs and MSOAs and generate activity counts for each, counting the average number of unique devices over our study period (September 2019). Table 2 summarises the statistics of this preliminary analysis, and Figure 5 shows the densities of average counts per L/MSOA across the study period.

Table 2. Summary statistics of LSOA and MSOA activity counts

	LSOA	MSOA
<i>Number of areas in London</i>	4835	985
<i>Number of Areas with count < 10</i>	57	0
<i>Mean count per area</i>	75	197
<i>Mean number of H310 hexes per area</i>	17.5	118.6
<i>Median count</i>	59	132
<i>Count range across areas</i>	3 - 1225	30 - 1432

In London, 57 LSOAs display activity counts below 10, which would not allow opportunity for non-disclosive subdivision, as is it best practice to omit counts below 10 to prevent risks of re-identification and disclosure. Furthermore, all MSOAs are above a count of 10, and the minimum activity counts is of 30 unique devices, meaning even the least active MSA could theoretically be divided into three sub-regions. We thus select MSOAs as the merge boundaries.

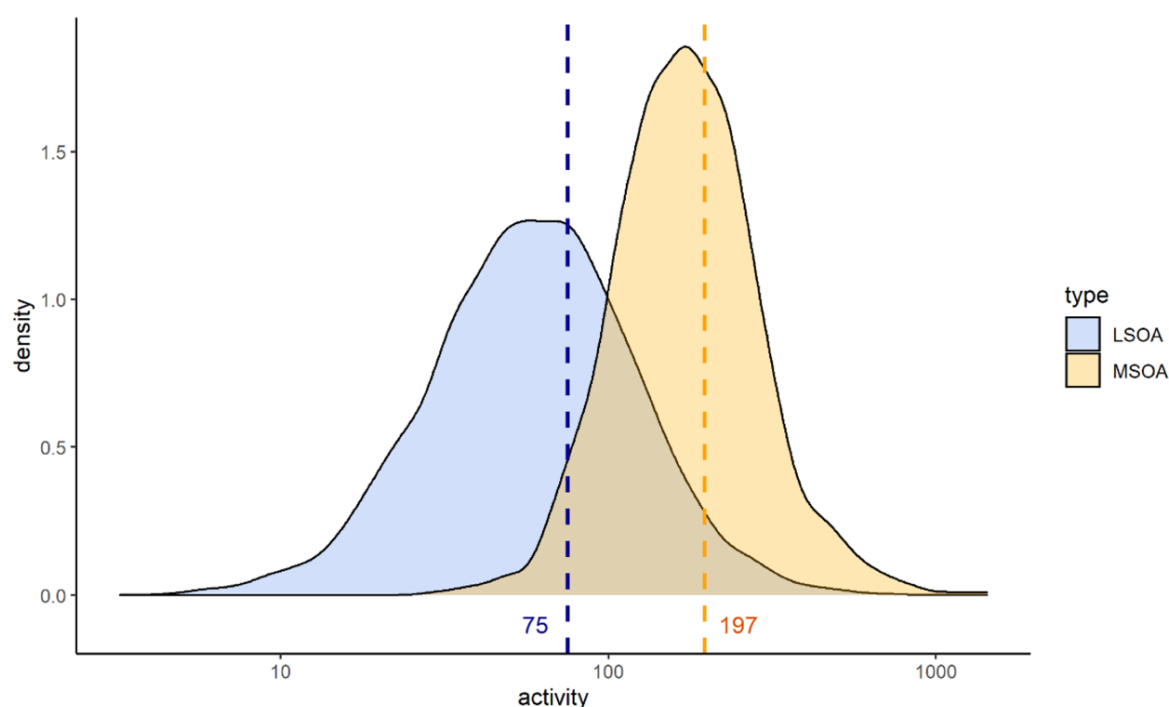


Figure 5. Densities of average unique counts per LSOA (blue) and MSA (yellow), across London.

5.1.1.3.1.2. Linkage and assignment

Each H310 is assigned the code of the MSA it falls within. We find the list of hexagons contained in each MSA polygon using the `polyfill` function of the `h3` R package. Where a hexagon overlaps more than one MSA, it is fully assigned to the one with the majority area overlap. Figure 6 illustrates this assignment process.

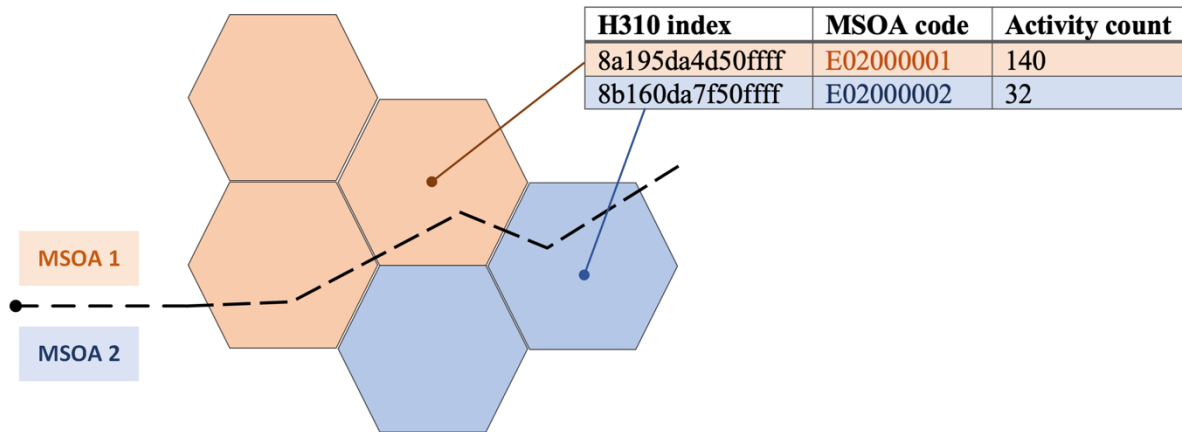


Figure 6. Diagrammatic illustration of the MSOA code assignment, with a synthetic sample of the data after this step.

5.1.1.3.2. Land use assignment

5.1.1.3.2.1. Creating the land-use and terrain profile (topo_code)

The aim is to inform the grouping of the units by weighting them based on data characteristics where available, and contextual information to make homogenous regions. One such piece of contextual information can be underlying terrain. We decide to homogenise our regions based on the terrain to promote regions which respect underlying land features unlike arbitrary tiles which cut them indiscriminately (Brunsdon and Comber, 2020).

We thus use a raster-based dataset describing land-use and terrain in London, provided for free and updated weekly by Geofabrik (Geofabrik, 2022). We download rasters for each of the following features: water, road, greenspace, residential area, commerce and retail area, and “others” (industrial and military areas). As some of these areas overlap (e.g. water in greenspace, road in residential area) we add the layers in a raster mosaic, creating different codes for overlapping areas. Figure 7 shows this raster data and corresponding codes for Greater London. The raster codes are designed to reflect the different layers that compose the mosaic.

Table 3 describes the resulting topo_codes assigned to each type of land use.

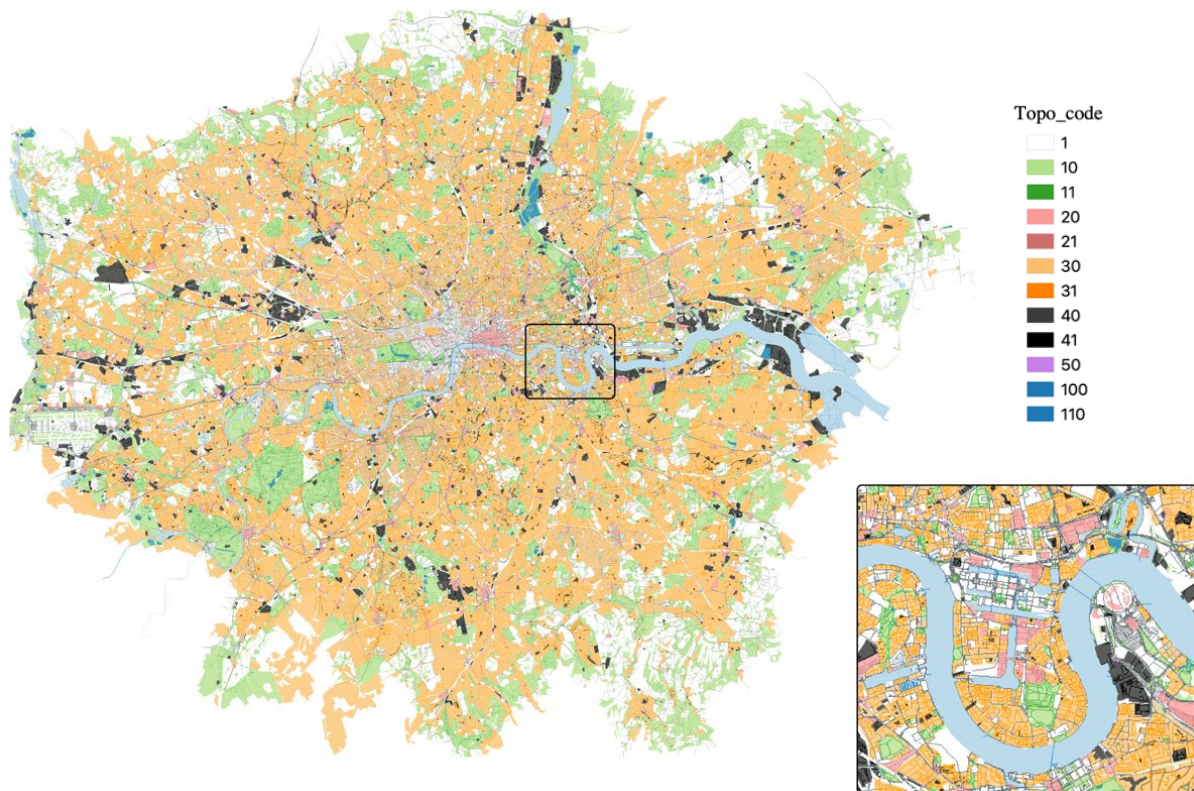


Figure 7. Raster mosaic of terrain information, Greater London.

Table 3. topo_code assignment to their corresponding land use. The red entries illustrate the purpose of having decomposable numbers where layers overlap for ease of interpretation.

Land use	Topo_code assigned
Road	1
Water	100
Greenspace	10
Commerce and retail	20
Residential	30
Other	40
Road in residential area	31 (30+1)
Water in greenspace	110 (100+10)

5.1.1.3.2.2. Linkage and assignment

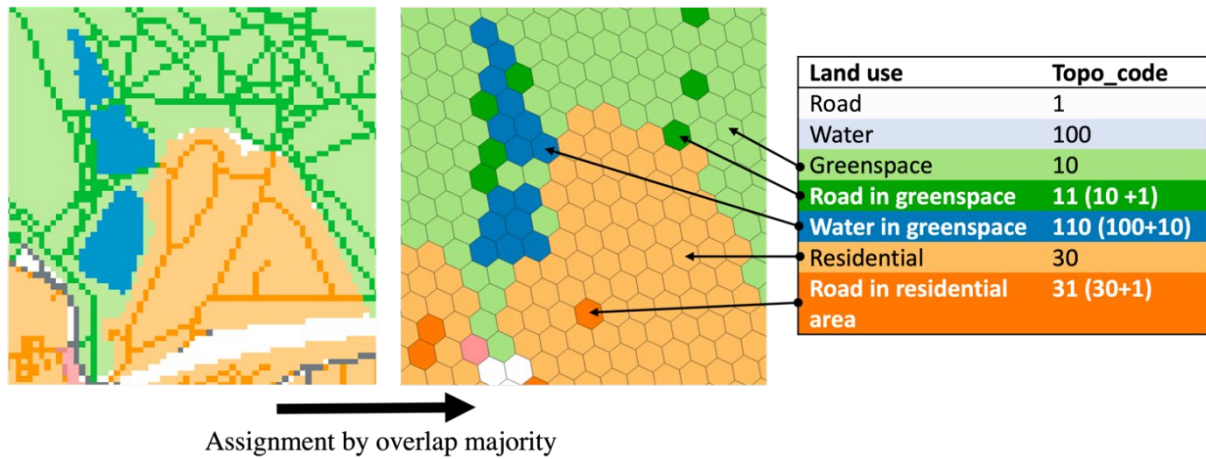


Figure 8. Illustration of the topo_code assignment process with codes designed for land use overlaps.

We then attribute topo_codes to the hexagons, assigning the one that overlaps each hexagon in majority. This is performed using zonal statistics between raster and vector layers: the code that is present in majority is calculated across the rasters cells within the zone defined by each hexagon vector object (Mearns, 2015, p.54). Figure 8 illustrates this assignment process. Table 4 displays a synthetic sample of the dataset after all contextual information has been assigned.

Table 4. Synthetic sample of processed dataset at atomic unit level to be used for regionalisation.

H310 index	MSOA code	Activity count	Topo_code
8a195da4d50ffff	E02000001	140	31

5.1.2. Preferred neighbour assignment (Making of the PNMx)

Having successfully created units and attributed characteristics to each, we now combine hexagons containing insufficient counts into larger non-disclosive regions. We want each unit with low activity counts (below 10 unique devices) to combine with a neighbouring cell with similar characteristics to surpass the disclosure threshold where possible. We thus use the cell's topo_code, MSOA code, and activity count to determine which of its neighbours it should combine with. This step of the process goes as follows:

- (1) Filtering and assigning neighbours: filter all the cells under a count of 10, retrieve each cell's 6 adjacent neighbours and rank each neighbour based on their characteristics. Retrieve the preferred (most similar) neighbour (Section 5.1.2 in Figure 1).

- (2) Group membership and merging: identify all cells with a common preferred neighbour and assign them a group membership, then combine and merge all the cells belonging to the same group into larger homogenous regions (Section 5.1.3 in Figure 1).
- (3) Calculate the new groups' count and repeat the process for the hexagons in groups below the threshold.

The ordering of the filtering and assigning of cells is crucial to the regionalisation method. The key is to combine regions to surpass a count of 10, so the very first step is filtering hexagons below this count, highlighting this as the first priority. By iterating until the threshold is reached, we ensure this key condition is met. The whole process is performed for each MSOA separately, ensuring no resulting regions cross MSOA boundaries (nesting the output regions inside MSOAs). Topo_code homogeneity is thus the third factor, used to inform the combination of the hexes to reach the threshold, rather than a strict condition. The following parts describe these steps in more detail, explaining the preferred neighbour's assignment process and the identification and merging of the groups.

5.1.2.1. Filtering IDE units and ranking their neighbours

We first filter the H310 cells with an activity count of less than 10. For each cell filtered (i), we apply the `k_ring` function from `h3 R` to retrieve its 6 direct neighbours (n). The neighbours are then written in a data frame with the original cell and their respective topo_codes (topo_i and topo_n) as well as their activity counts. The neighbours are assigned a 1 if their topo_code is the same as i's, and 0 if not. They are then ranked, privileging neighbours with the same topo_code (topo_n=1), and highest activity count as a second ordering factor. This means that when there is more than one (or zero) neighbours with the same topo_code, the highest ranked neighbour will be the one with the highest activity instead. The top-ranking neighbours for each cell below 10 are listed next to it in the PNMx. This is illustrated in Figure 9. Merging with the neighbour with highest counts both increases the potential of reaching the threshold in the first iteration (and with the smallest area), and reduces the likelihood that the total activity of the combined group is reduced to a number below 10 by duplicates (the same device being in cell A and cell B, resulting in a count reduction when A and B combine, as that device will only be counted once).

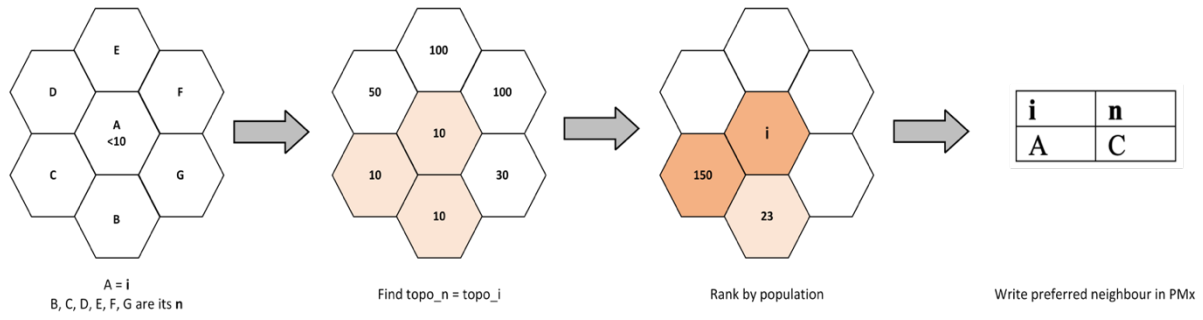


Figure 9. Example process of assignment of preferred neighbour.

5.1.2.2. Graph theory: finding connected nodes

Once we have all the low-count hexagons' preferred neighbours written in the PNMx, we want to combine them into larger regions. However, the merges are often interconnected, with multiple hexagons merging with linked neighbours. If this process is done iteratively, there is an added complexity to determine where the process starts and ends. This would also require that a "merged object" is still recognised as being a cell's potential preferred neighbour, and risks outputting different regions based on the start location of the process. To prevent these issues, we first find all interconnected cells and merge all the final groups at once, rather than merge iteratively.

For this, we borrow algorithms and concepts from the discipline of graph theory in mathematics, by converting the database into a graphical object for community detection (Holmes and Haggett, 1977; Fortunato, 2010; Liang *et al.*, 2022). Graph theory has been employed to model a wide variety of relationships and processes, in fields ranging from linguistic and computer science to biology and social science (Diestel, 2000). Though it has been applied to detect flows of mobility as well, we here use it in a simpler way to efficiently detect all connected H310 cells without needing to convert the almost 1 billion H310 cells into computationally costly polygonal spatial objects, and detect their neighbours geographically. Graphical objects can conveniently represent matrices with nodes and edges, making the computation of the relationships much more efficient (Holmes and Haggett, 1977; Gibbons, 1985; Fortunato, 2010). In our case, nodes are the H310 cells, and edges the relationship between them (where they are each other's preferred neighbour, or their neighbour's neighbour).

Using the PNMx, we convert all cells into nodes with their connections being directional edges (pointing towards their preferred neighbour). We use the igraph package in R to convert the

PNMx into a directed graphical object (igraph Core Team, 2022) (Appendix 2). The components function of this package retrieves all the linked components of the graph by listing their membership to a group. The directed graph is then converted to an undirected graph, to identify linked hexagons regardless of distance: those that are neighbours, or neighbours of neighbours. We then list all cells which would be part of the same final region and attribute a group membership or group ID to each. For instance, if A's preferred neighbour is F, and B's is also F, A-B-F would have the same group membership and merge into the same polygon. This is described visually in Figure 10 and corresponds to step 5.1.3 in Figure 1.

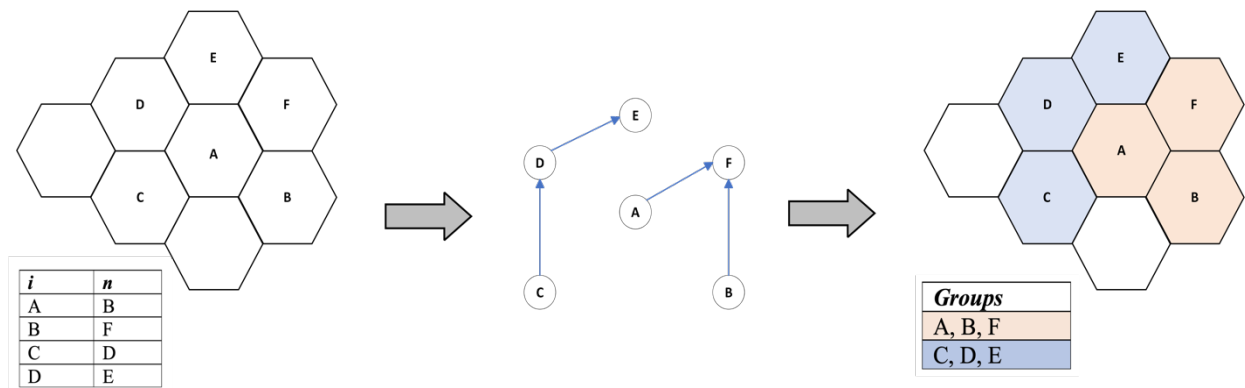


Figure 10. Example of membership detection through graphing method and group generation.

Once group memberships have been identified, cells that belong to groups with activities below the threshold go through a new iteration of the PNMx process, with slight variation. Their neighbours are relisted, but this time ranked based on their new group's activity, after topo_code. The low-count hexagon will seek to regroup with a neighbour that belongs to a group already above threshold, to maximise the chances of reaching the threshold with this new iteration. Thus, the groups under the threshold are broken down, and the cells that make them up are reassigned to neighbouring groups to meet the threshold. If the loop cannot reassign a group, it stops after 5 iterations with no changes, and keeps exceptional below threshold groups (for instance, if the whole MSOA were under a count of 10 and there are no more neighbours to combine with as the process does not cross MSOA borders). Group IDs are assigned at the end of all iterations, built on the group membership and MSOA codes.

5.1.3. Merging

Finally, the cells are merged based on their group ID, creating regions of combined H310 hexagons. The entire process of assigning neighbours and merging takes approximately 15

minutes on a single threaded Intel 64 processor, for the Greater London area, with 116,555 hexagons, requiring few spatial operations outside of the final conversion and merge, and refraining from recalculating each neighbourhood as it iterates (only the ones below 10). This is a clear advantage of sorting and attributing preferred neighbours using h3 functions and graph methods, as they greatly reduces costly spatial operations. In the resulting regions, cells with above-threshold counts that are not attributed a neighbourhood through the PNMx remain unmerged and thus at the H310 scale (edge length 75.8 metres).

5.2. Outputs

5.2.1. Region shapefile

The output of the algorithm is a shapefile with region codes and their polygon geometries. These regions are called the H310-based regions, or H3BRs, throughout the rest of this thesis. Figure 11 is a map of the resulting H3BRs, coloured by MSOA to demonstrate the successful nest within census boundaries. The MSOA code is retraceable from the region codes, as they are made following the format [MSOA_code]_[group membership number] (for example, E02000017_14). If certain MSOAs had total counts of fewer than 10, it follows that the hex-based regions within them will also have counts fewer than 10 as the regionalisation process prioritises nesting within the boundaries of the chosen larger spatial unit. This is an important feature of the algorithm: though it seeks to maximise the number of regions surpassing counts of 10, it will not do so to the detriment of other important regionalisation factors, such as the spatial unit constraints deemed essential for linkage to contextual data.

Using another source of data, with different activity counts and distributions for instance, would yield different region outlines. This is at the core of data-driven methodologies and helps fit the results to the original dataset and analytical intentions. These variations, and their contingency on the dataset, are discussed in more detail in Section 5.3 of this chapter (sensitivity analysis) as well as throughout Chapter 6.

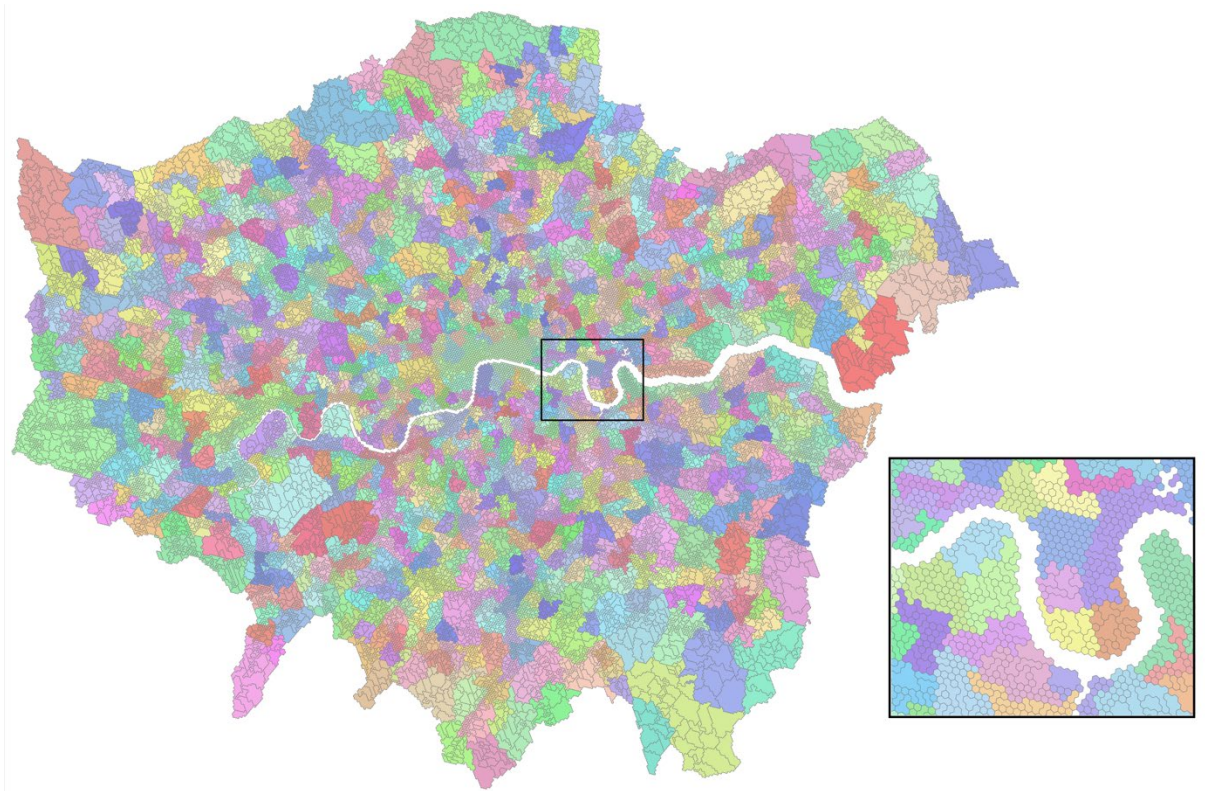


Figure 11. H3BRs at resolution 10, nested in and coloured by MSOA. The map shows areas of varying density of activities, with cells remaining unmerged next to other region made of bigger merged groups. The tidal stretch of the River Thames is excluded from the output regions as census units, such as MSOAs, are conventionally excluded from it.

5.2.2. Functions

One aim of this regionalisation methodology was to propose a tractable and reproducible algorithm. We thus detail the code behind the key functions developed for the workflow. This is provided here in the form of pseudocode describing the core processes (Figure 13 and Figure 14), contextualised within a summary chart of the loop (Figure 12). Throughout all code examples, i refers to the hexagons which is being assigned neighbours, n represents the neighbours. The final output of the process is a table of H310 IDs with their final group id, maximising groupings over 10. The functions and latest versions of the algorithm can be accessed on GitHub (<https://github.com/lousieg/H3BR>).

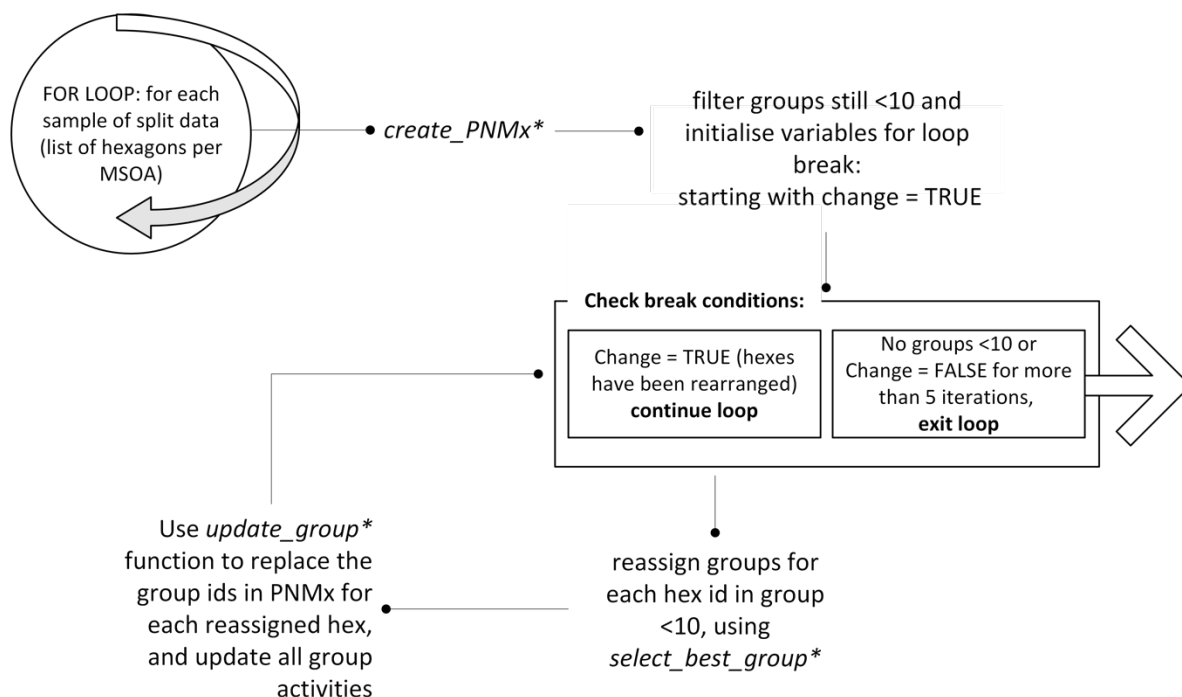


Figure 12. Process contextualising the use of the functions provided for the regionalisation.
* = functions.

The first and most central function is the one which reads in activity data per H310 and returns each H310 index with their region group number and group activity. It is the function which generates the Preferred Neighbour Matrix for one iteration (*create_PNMx*).

```

FUNCTION create_PNMx(M) :

    list10 = FILTER(h310 id WHERE activity < 10)

    FOR EACH i IN list10:
        neighbours(n) = k_ring(i, radius=1)#gets the neighbours
        n = REMOVE(i FROM n)#removes I from its own list of
        neighbours
        df = SET COLUMN i TO i, COLUMN n TO n
        df = CONVERT TO DATAFRAME(df)

    END FOR

    SET COLUMN NAMES(df, ["h3_10_char", "neighbours"])
    data = LEFT JOIN(df, M)#join with original data to retrieve
    contextual information
    SET COLUMN NAMES(data, ["i", "n", "topo_i", "activity_i",
    "topo_n", "activity_n"])
    data = REMOVE ROWS WITH NA VALUES FROM data

    data = ADD COLUMN same_topo
    IF topo_i == topo_n THEN same_topo == 1 ELSE same_topo == 0
    SORT BY i, same_topo DESC, activity_n DESC
  
```

```

PNMx = REMOVE DUPLICATE ROWS FROM data BASED ON i
PNMx = SELECT COLUMNS [1, 2] FROM PNMx

PNG = CREATE GRAPH FROM DATAFRAME(PNMx, DIRECTED=TRUE)
Groups.membership = FIND CONNECTED COMPONENTS IN GRAPH(PNG)

group = CONVERT TO DATAFRAME(Groups.membership)
group = ADD ROW NAMES AS COLUMN "i" TO group
SET COLUMN NAMES(group, ["i", "group"])
group.group = CONVERT TO CHARACTER(group.group)

RETURN group
END FUNCTION

```

Figure 13. Pseudocode snippet for the create_PNMx function.

This function takes data which has been split, in our case by MSOAs, and returns their PNMx ID with group activity count. Then, we use these groups and their counts to identify the groups which may need to be recombined in a second iteration (group activity < 10). At this step, we initialise variables to track iterations and changes to ensure the cells are being reattributed to a new group. At the start of the loop, we identify cells with group activity below 10 and recombine them using the following functions (Figure 14)

```

FUNCTION select_best_group(hex_id, df):
    current_hex = FILTER(df WHERE i == hex_id)
    n_details = get_n_details(hex_id, df)*
    new_groups = FILTER(n_details WHERE group !=
current_hex.group) #find neighbouring hexes in different groups
    IF COUNT(new_groups) == 0 THEN
        RETURN NULL
    ELSE
        best_group = new_groups
        ADD COLUMN same_topo = (topo_code == current_hex.topo_code)
        ADD COLUMN priority = IF same_topo THEN 1 ELSE 1
        SORT BY priority DESC, group_activity DESC
        SELECT FIRST ROW
        RETURN group id from new_group

FUNCTION update_groups(df):
    df = GROUP BY group id
    df = MUTATE(group_activity = SUM(activity))
    RETURN df

```

```

* FUNCTION get_n_details(hex_id, df):
    neighbours(n) = k_ring(hex_id, radius=1)
    neighbours = EXTRACT(hex_id FROM neighbours)
    n_details = FILTER(df WHERE i IN neighbors)
    RETURN n_details

```

Figure 14. Pseudocode snippet for functions "select_best_group" and "update_groups" with "get_n_details" being an important element of the first: the one which extracts the characteristics of i's neighbours once more

Select_best_group identifies each hexagon's neighbour's and its group information (with *get_n_details*) to recombine the hexagons from groups <10 into neighbouring groups, ranking these new groups again based on *topo_code* and highest count. It excludes the neighbours which are in the same group as *i* to ensure the hexagons are redistributed and not reassigned to the same initial low count region. Group activities are recounted to see if further reassignments are required.

If no hexagons' group counts are below 10, we exit the loop and keep the PNMx at this stage. If the count of groups under 10 does not change at all over 10 iterations, we exit the loop (this can happen if the entire MSOA is under a count of 10, though this is rare as shows in the LSOA vs MSOA assessment above). If the count of groups under 10 changes from the last iteration whilst remaining above zero, the iteration count is reset and the loop goes for another iteration, until there are no groups under 10. The final step after exiting the loop consists in turning the H310 IDs into polygons using the *h3_to_polygon* function (from the *h3* package) and merge all shapes with the same ID with the *sf* R package (Appendix 2).

These functions are detailed here as an output of the methodology, as a major focus of these explorations was not only to provide regions for this specific dataset, but also propose the reusable functions to future users.

5.3. Validation and Sensitivity Analysis

5.3.1. Validation: comparison with traditional aggregation methods

To determine how much data is preserved by using our methodology and to compare the differences in areas omitted between aggregation methods, we assess our regions against common ways of aggregating these types of datasets: grids and administrative boundaries. To conduct this comparison, we use the notion of *remaining counts* used previously in Chapter 4 and Section 5.1.1.1 of this chapter. As discussed earlier, this methodology seeks to preserve a maximum number of counts and keep the data as close to the original as possible. Thus, the following comparison assesses the *total counts* and the *omitted counts*. The hope is to produce areas for which the *remaining counts*, obtained by subtracting the omitted counts from the total count, are as high as possible.

We use one day of raw in-app data from September (Monday the 2nd) across London and aggregate it separately using the H3BRs, OAs, LSOAs and OSGB250. For this, we use the

H3BRs created earlier from the average month of data per H310, rather than remake specific September 2nd regions, to reflect a realistic usage of the regions and compare their performance with other pre-existing shapes. This is why certain regions may be under a count of 10 for the H3BRs in the following examples, as the results were made from aggregating a different data sample from a shorter time period. This aims to demonstrate how the use of a stabilised H3BR (made from an average of the dataset over a longer duration) fares compared to other methods, rather than making new regions specifically for this daily sample.

We assess the data registered and the level of granularity (*remaining counts*, described above) achieved by each aggregation method. We thus calculate the *omission count* by adding all activities contained in areas of IDEs. We subtract these lost counts from the total count, which returns the *remaining counts* (non-omitted data). The aim is thus to find a balance between generating the smallest possible areas whilst preventing too many cells to fall under the omission threshold of 10 devices. Table 5 lists the comparison between H3BRs, OAs and LSOAs. LSOAs see less data omission due to their larger size (only 4,835 LSOAs across London against more than 25,000 for both OAs and H3BRs) but display less counts overall due to low granularity.

Table 5. Comparing data omission when aggregating to H3BRs against OA and LSOA aggregation.

	H3BR	OA	LSOA
Total geographical areas	30,730	25,053	4835
Total activity counts	764,326	557,231	329,776
Omitted activity counts	30,896	46,381	740
Remaining activity counts	733,430	510,850	329,036
Comparing with H3BR:		30% less remaining counts	55% less remaining counts

Table 6 lists the percentage of data and cells omitted for H3BRs, H310 (pre-regionalisation), OA and OSGB250 aggregates at the London scale. The four geographies compared here have similar area averages (OSGB250 average size of 61,068² m², OAs 62,830 and H3BRs 51,205). Here, we show that making bespoke regions has advantages over simply aggregating to a smaller scale, as H310s pre-regionalisation have an average area of 15,047 m² but results in

² less than 62500 m², which is 250 metres x 250, because some squares are cut to the outline of Greater London.

large spatial and data omissions (23% of data omitted, and 78.8% of hexagons omitted). The H3BRs provide close to double the counts compared to OAs and OSGB250s. This exemplifies both the multiplicity of the data thanks to smaller areas, and the reduced concern of large omissions made possible through specifically packaged areas. Despite the areas being larger in size than H310 alone, we notice here the importance of zoning, with the H3BRs preserving more data than its more granular counterpart. Zoning is also highlighted by the H310 based regions omitting fewer low counts than the larger OAs and OSGB250: the larger regions, usually containing more counts per region due to their size (and thus more regions above 10), should theoretically have lower percentages of omitted data, but in fact have higher omissions as their zoning does not fit this data.

Table 6. Remaining data points post omission of low counts for aggregates of similar scale but different zoning.

	% Of counts omitted due to low counts	% Of cells omitted	Remaining counts (T-O)
<i>OSGB250</i>	8%	42%	474,499
<i>OA</i>	8%	51%	510,850
<i>H310</i>	23%	78.8%	675, 188
<i>H3BR</i>	4%	6.9%	733,430

Where the OA aggregation incurs a loss of more than half its areas, the H3BRs only omit 6.9% of their areas, particularly in places of low device coverage, rather than arbitrarily across London. This is shown by the maps in Figure 15, where the data omission for H3BRs is concentrated in areas with typically low data, such as the very outskirts of London or zones of low mobile and internet coverage (Richmond Park), whereas OA omission is scattered with a more random pattern not underlined by the data's distribution. OSGB250 cells resemble the H3BRs more closely than OAs, being more neutral in nature, and see less omitted cells than the OA aggregate. However, they do not consider land use or data distribution, resulting in twice as many counts being omitted than for the H3BRs, and making them difficult to merge with other geographies.

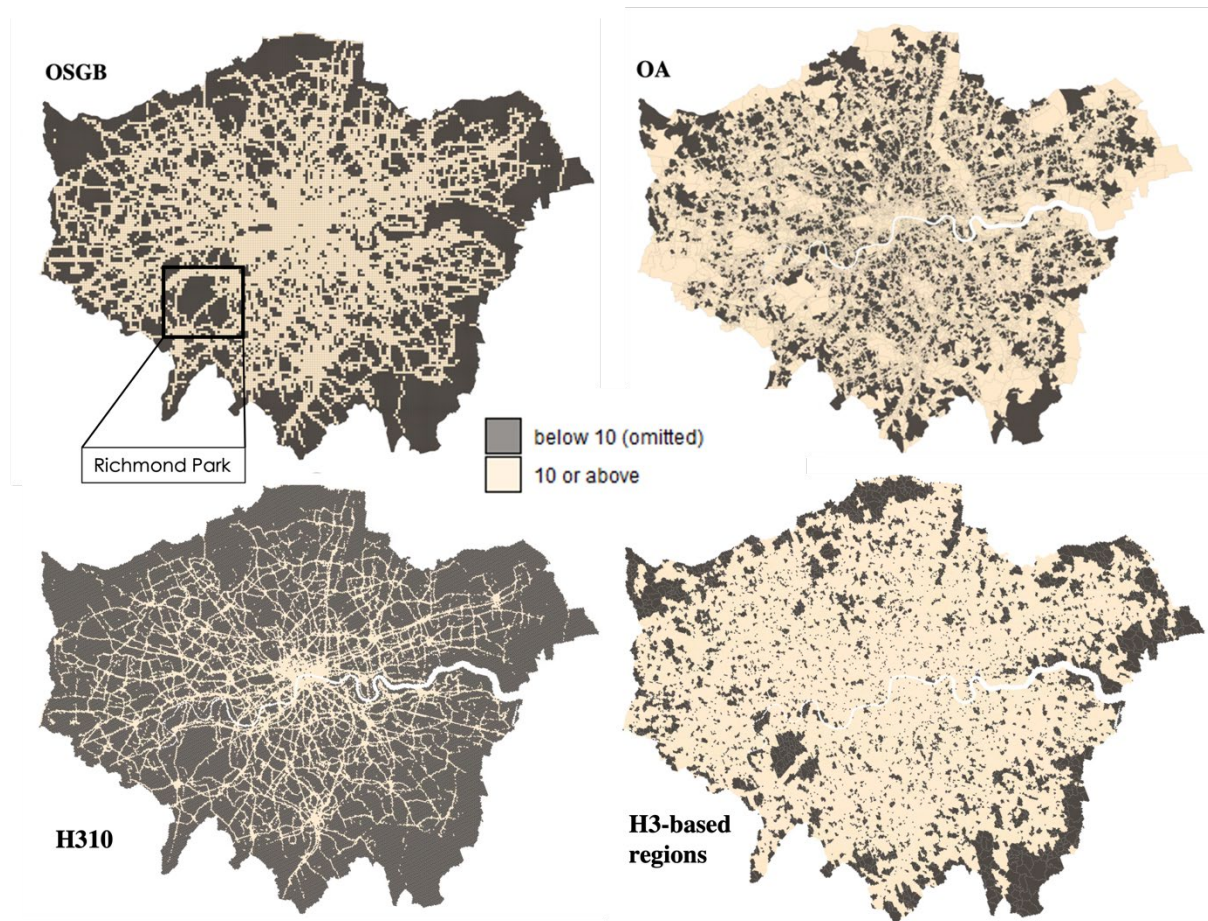


Figure 15. Maps plotting the cells omitted due to low count for OSGB250, OAs, H310 and H3BRs.

Road networks are also particularly noticeable for gridded aggregates (OSGB and H310). This can be both due to there being the most datapoints alongside road networks, and due to the movement of devices counting as “unique devices” as they move through the grids which make up the road. By regionalising the data, the roads remain largely unchanged compared to the H310, as their high activity counts often makes them exempt from having to recombine. Rather, regionalising enables the inclusion of the adjacent lower-activity areas as their own distinct units. For example, in the borough of Camden shown in Figure 16, H310 hexagons are not recombined in the South of the region where there is the most activity and busy intersections (Such as Kings Cross and St Pancras Stations), whereas the North of Camden recombines to allow a better use of its lower data counts.

Figure 16 also compares the OA and H3BR for the London Borough of Camden, along with the percentage of omitted cells for each. Our dataset has good coverage for Camden, a busy leisure and retail area. These maps demonstrate that despite good coverage for this borough, the OAs still incur significant data loss. Furthermore, omissions observed in the H3BR example

more logically concentrate around parks and low coverage zones (Hampstead Heath, situated to the North). This makes a case for the use of bespoke regions, regardless of data availability or data scarcity, when trying to promote granular analysis.

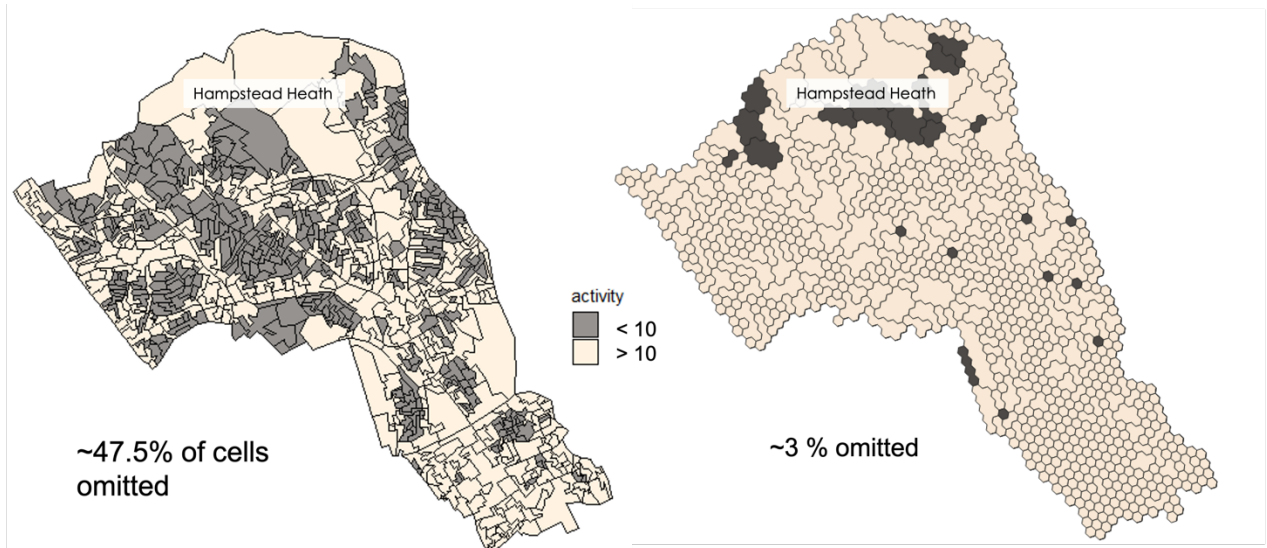


Figure 16. OA census geographies omitted due to low data counts (left) compared to area omitted from the H3BRs (right), when aggregating the same day of data points to each.

5.3.2. Sensitivity analysis using the Jaccard index

Comparing the H3BR as done above makes a compelling argument for the method, and seems to validate the techniques proposed. However, attesting for the stability and reproducibility of the method is necessary. A sensitivity analysis of the regionalisation method was thus conducted, with the aim of identifying the impact of both data and parametric changes (such as changing thresholds) on the final regional outcomes. The changes are compared by calculating the Jaccard coefficient of various outputted regions (Jaccard, 1912).

5.3.2.1. Definition

The Jaccard coefficient (or Jaccard index) is a measure of similarity and overlap between two features. It is summarised by the size of the area of intersection between two objects divided by the size of their union, and returns a value between 0 and 1 (with 1 describing two identical object) (Jaccard, 1912).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

(With J describing the Jaccard coefficient, A as object 1 and B as object 2.)

It was originally developed by Paul Jaccard, a botanist who sought to compare the distribution of alpine flora in three different districts. It was thus originally a spatial concept, as Jaccard measured the similarity in areas of spread (Jaccard, 1912). Since then, it has been used extensively, notably in comparing datasets (Fletcher and Islam, 2018; Costa, 2021).

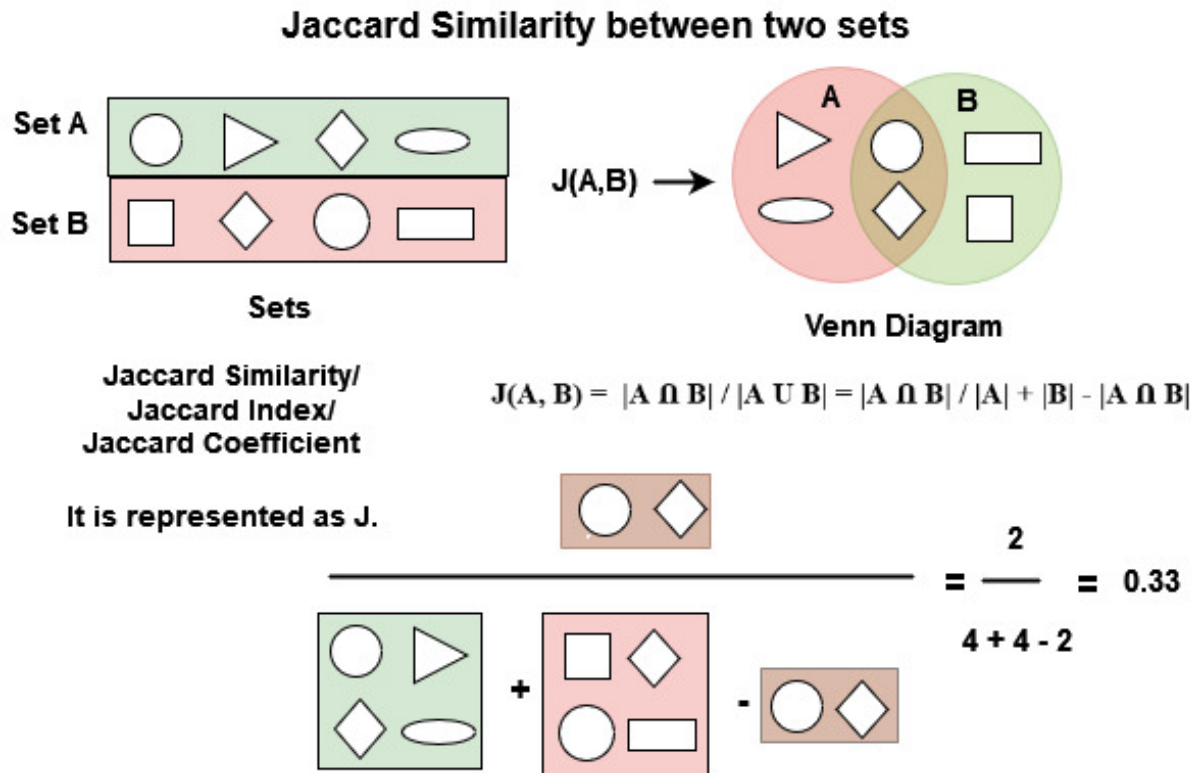


Figure 17. Diagrammatic explanation of Jaccard similarity calculations.

Source: <https://www.geeksforgeeks.org/>

Since its development, the Jaccard index has been commonly applied to spatial research, particularly in the fields of ecology and biodiversity, thanks to its ease of use for comparing different ecosystems. It is often employed to represent the proportion of total species pool that are shared by communities compared (Lapin and Barnes, 1995; Giraudel and Lek, 2001; Azaele *et al.*, 2009). In the case of the H3BR methodology, it can help assess how different two regions would be if they were made with varying thresholds and criteria in mind. In short, we are interested in the proportion of total regional units that are shared by the output regions compared.

5.3.2.2. Applying Jaccard to assess volatility

Comparing the Jaccard index of the regions constructed by changing one parameter at a time, we can see how they alter the outcome regions, which can help quantify the method's volatility. Here, we look for the stability of the outcome based on relatively small differences at the starting point. If large changes in the data bring about large differences, this says little about the volatility of the methodology and more about data instability. However, if small changes, whether in the data or the model's parameters, bring large differences in outcome, a measure of stability could be introduced to improve the prototype and reduce its volatility.

5.3.2.2.1. Changes in data

Using the H3BR making algorithm, we generate sets of regions from each of 7 consecutive days in London (Monday 02-09-2019 to Sunday 08-09-2019 of September) and compare the Jaccard relationship of the outputted shapefiles. Figure 18 provides an example of what the Jaccard index represents in this context: the coloured regions are those that remain the same for both region sets compared.



Figure 18. Intersect of the regions generated for Monday and Sunday. The areas in white are those that differ between the two regions generated for each day. Everything else remains identical between both. Monday and Sunday have Jaccard coefficient of ~ 0.25 .

We also compare each day to the “average day”, the average count of activity for each cell across the week. We use activity counts and a disclosure threshold of 10 for all the tests. Figure 19 displays the resulting Jaccard indexes between all days’ regions. We notice that the weekend (especially Sunday 8th) is the most dissimilar to other days, with Jaccard indexes of less than 0.28 between it and other days. Most days are closer to the week’s “average day” than to any other day (Jaccard indexes of above 0.32). Furthermore, comparing regions created from weekly averages returns Jaccard indexes values of around 0.8, with variations in indexes not straying further than 0.05 points, indicating a stability of the method from week to week. This suggests that selecting a larger timeframe than the one to be studied and averaging the counts generates a more stable region overtime. The similarity in results for weekdays in contrast to weekends follow predictable patterns of activity, reflected by the outputted regions. This highlights the regions’ potential in relating to and being representative of the data inputted in their making.

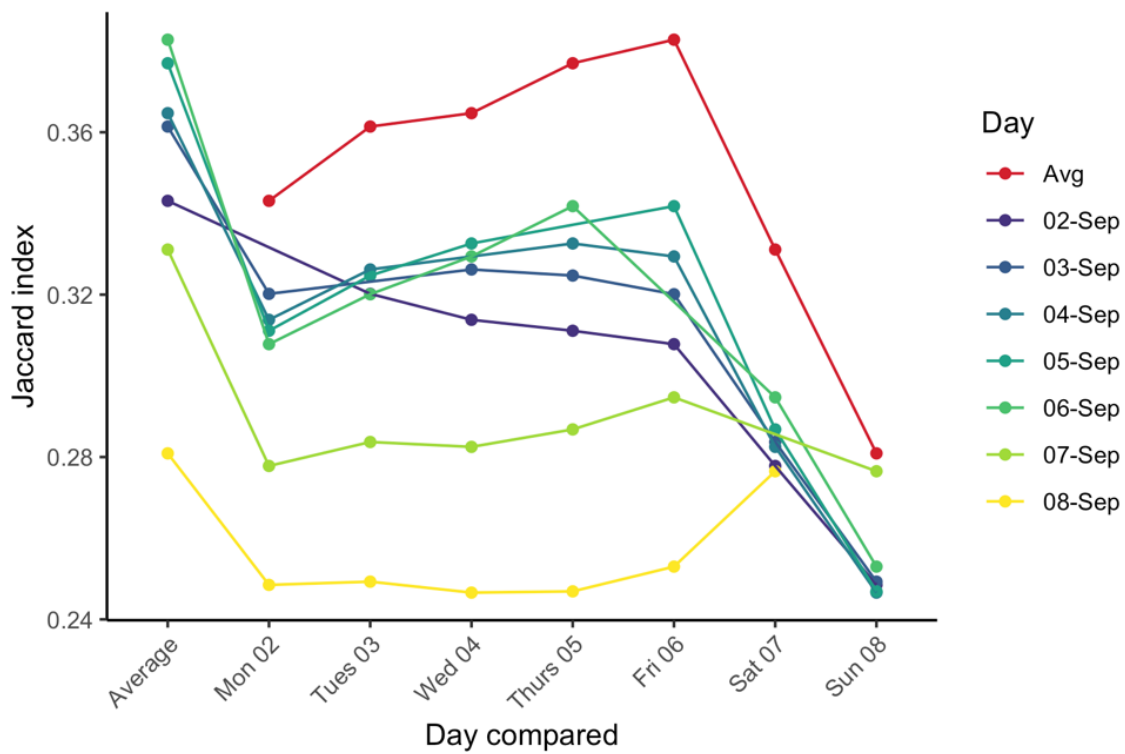


Figure 19. Comparison of Jaccard index between regions made with data from different days of the week ('Avg' corresponding to the 'average day' across September – mean of all days). Higher values on the y axis correspond to a greater overlap with the regions on the x axis.

5.3.2.2.2. Changes in thresholds

The threshold determines the number below which a cell must find a neighbour to merge with. Changes in threshold will impact the number and necessity of merges in the region-making process. We perform all threshold tests on the same dataset: the average day count for the month of September in London, and generate multiple regions with incremental changes in thresholds. After testing thresholds between 10 and 100 with increments of 10, Figure 20 displays the results of varying thresholds either side of 20 and 50 with much smaller increments, showing that small changes in thresholds have a minimal impact on the outcome regions.

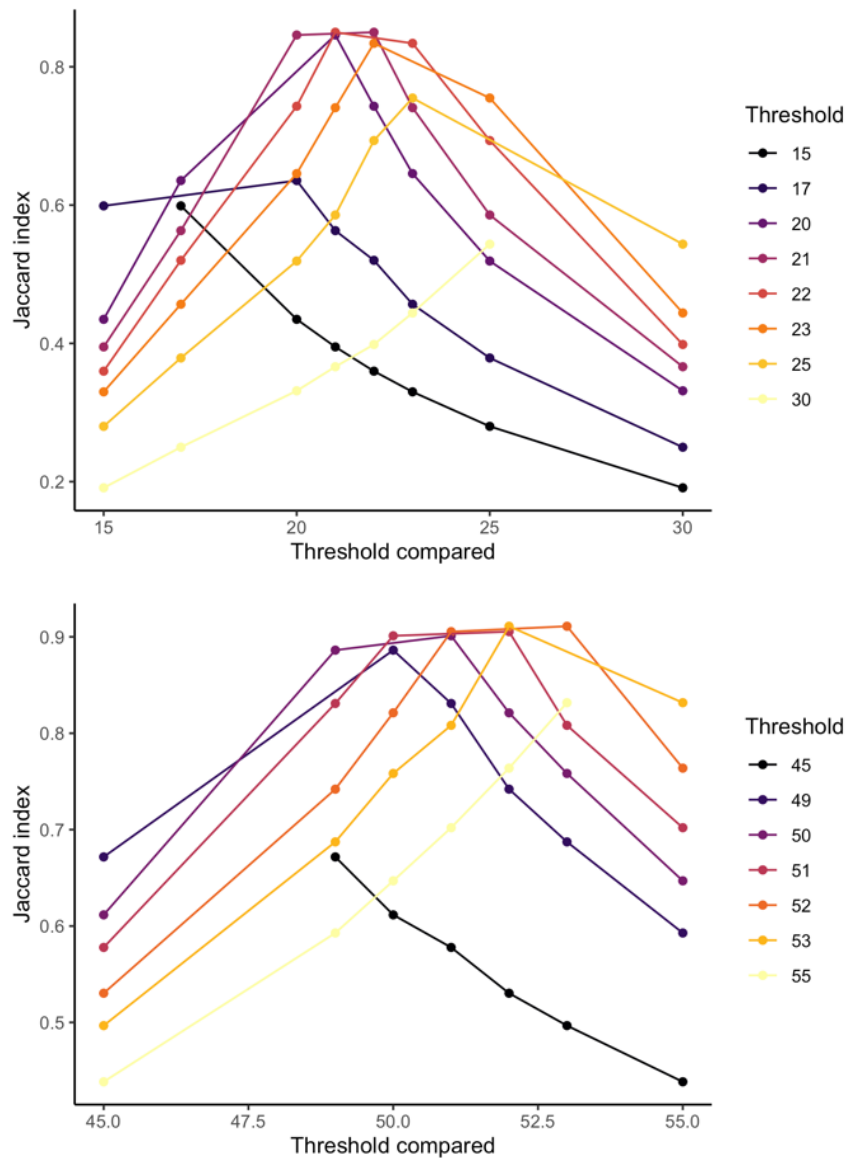


Figure 20. Comparison of Jaccard index between regions made with varying thresholds. Higher values are more similar to other regions than lower values (for example 45 is has less overlap with other regions than 51)

Small changes in threshold (steps of 1 to 3) seem to have little impact on the final regions, with Jaccard similarity indexes being of 0.8 and above (80% of the regions are unchanged with a small change of threshold). However, the change is not the same when happening at higher or lower values. Comparing regions outputted with threshold 50 and 100 creates regions with a 0.6 Jaccard index, whereas 50 and 20 (a smaller step, but towards a smaller threshold), creates regions with no relationship. Figure 20 also shows that smaller threshold values (15-30) are more volatile, with larger variations overall than between 45-55. This can be explained by the data amounts and distribution. There are few cells with activities above 50 in this test data, whereas in comparison there are triple the amount above 50 and 6 times more above 10 than

above 50. This signifies that increasing the threshold above 50 incurs less change in output regions than changes in lower threshold values. Lowering the threshold results in fewer cells being required to merge overall, changing the outputted regions much more.

However, these differences tend to be geographically consistent, with busier areas (thus smaller cells) being more impacted by changes in threshold. At the London scale, this change is most visible along road networks. On the one hand, this signifies that busier areas with more data are the most prone to repackaging due to parametric changes, indicating inconsistent volatility in the region-making strategy. Yet, this can also be mitigated by a set of considerations when selecting a threshold for regionalisation, most likely based on field knowledge and understanding of the research question. In this case, despite potential volatility in lower thresholds, a threshold of 10 devices is appropriated, informed by disclosure control guidelines rather than the data's central tendencies. The option of making regions with a more stable sample of data (here average counts over a longer period rather than a single day of data) also helps mitigate these effects.

5.3.2.3. Discussion

The H310-based regionalisation method thus proves to preserve counts significantly when compared to other geographies. Additionally, the space is also preserved, with only 7% of the area omitted, the bespoke regions provide insight into entire spatial areas which would otherwise be lost to the aggregation process. Indeed, it is both data and geography which are preserved by best fitting delineation. The addition of land use and MSOA to the process promotes the combination of H310s to be more representative of land features and administrative boundaries than otherwise arbitrary grids.

A sensitivity analysis was also conducted to assess the volatility of the method. Changes in outputted regions were observed when the threshold condition for merging changed. However, the changes observed from a change of data input is encouraging, as they confirmed predictable patterns of mobility (namely, that weekdays are more similar to one another than weekends). This means the data driven methodology captures the underlying data movements, reflected in the output regions.

This regionalisation methodology is also particular in its blend of rules-based and hierarchical methods. Overall, it leans more towards a rules-based approach compared to fully hierarchical

data-driven solutions like Quadtrees or Voronoi regionalisation. While techniques like Quadtrees and Voronoi diagrams dynamically adapt regions based on data distribution (creating flexible, data-driven boundaries that reflect proximity or density) the H3BR incorporates predefined rules in its initial stages. These rules, such as filtering cells with fewer than 10 points and segmenting regions based on land-use, guide the regionalisation process before any further aggregation occurs. While H3BRs allow for data-driven adjustments, particularly in the hierarchical aggregation step, it is more structured and rule-bound in its approach, contrasting with the previously assessed methods (Chapter 4). Thus, the H3BR methodology could be considered as hybrid, but its reliance on initial rules places it closer to a rules-based approach.

As the quadtree methodology is hierarchical, and the Voronoi delineation method required HDBSCAN prior to regionalisation, they are difficult to compare quantitatively with the H3BR. We demonstrated previously that these two options performed better against OSGB and administrative boundaries, when only considering count preservation, though this assessment itself was flawed as it is difficult to address the significant amount of data alterations implied by the clustering steps necessary to these methods. However, as the H3BRs are also performant in preserving counts, our interest lies more towards their qualitative assessment against Voronoi and quadtree. Namely, can this new method outperform them in creating more logical, reusable and interpretable regions? This is discussed in more detail below, with an assessment of the H3BRs against the regionalisation principles used to evaluate the quadtree and Voronoi tests in Chapter 4.

5.4. Assessment against regionalisation principles and aims

This section goes back to each principle outlined in Section 4.2 of Chapter 4 to determine the fit of this methodology to the specific objectives set out. *Objective* and *Constraints* are key elements which should be prioritised, whereas *Criteria* and *Usability* describe desired traits which will help differentiate otherwise equivalent outputs.

5.4.1. Objectives

The objective principle is composed of the *purpose* (1) of the regionalisation, and its *relevance* (2) (referring to Table 4, Chapter 4):

- (1) The H3BRs respond appropriately to the objective of defining statistical areas appropriate for identifying clusters of activity whilst protecting privacy. This has been particularly demonstrated with the comparison to other geographies. The regions also allow for densely active areas to remain at the very granular level of H310s, and be displayed alongside areas of lower activities and lower granularity, highlighting the data hotspots rather than redistributing them.
- (2) The temporal dimension is still missing; there were no specific time constraints proposed for this regionalisation. However, they were made with a stable sample of the data in an attempt to make the output regions stable over time for wider use.

5.4.2. Constraints

Constraints include the strict limitations of *partition* (3) of the region and *contiguity* (4) of the units:

- (3) Every point is part of a single final H3BR. Contrarily to Voronoi, there are no cluster centres which could be outside of the final delineation, as such, every point is guaranteed to be inside the region it has been assigned to. The regions are assigned MSOA labels, and the algorithm proceeds independently for each MSOA, ensuring no region crosses the pre-determined administrative boundary. The choice of MSOA could be easily changed for other study-appropriate administrative boundaries.
- (4) The H3BRs are contiguous, and none are divided by another region with the proposed solution. Distant clusters cannot be combined into a same final region, as only units that have a relationship of adjacency, at the very least through neighbour relationship, can be assigned the same group ID.

5.4.3. Criteria

The desired characteristics listed by the criteria principle list that the regions should be *homogenous* (5), *autonomous* (6) and *coherent* (7).

- (5) Terrain homogeneity is included in this regionalisation methodology. The final outputs show this through the distinction of parks and roads in the region shapes, and the fact that no regions cross the River Thames or other key boundaries. This is a significant improvement from quadtree and Voronoi, which had no land use or

terrain input. However, this homogeneity trait is only a secondary requirement, after the condition of reaching disclosure-appropriate counts.

- (6) The homogeneity criteria mentioned above goes hand in hand with the autonomous character of the regions. The H3BRs are hard to recognise as regions as such to the naked eye (particularly the denser areas, which remained at the H310 level). However, MSOAs are recognisable thanks to the nesting process, and the most significant land features (see River Thames, parks etc. above) are noticeable. Thus, there is a notion of the regions being recognisable as regions which partition familiar space.
- (7) The output is coherent insofar as it is stable over time. The sensitivity analysis showed that most days were closer in resemblance to an average day than any other day, meaning it was an appropriate decision to make the regions from a stable sample's average. However, they were no size constraints, meaning the size range of the regions were not minimised. Apart from the guarantee they could not be larger than MSOAs, nor smaller than H310s, regions size can range in a non-negligible way for this criterion.

5.4.4. Usability

Finally, the usability principles, introduced as part of this research, include *conformity* (8) *flexibility* (9) and *reproducibility* (10):

- (8) As mentioned, the regions are aligned with a predetermined administrative boundary, to promote linkage and usability
- (9) They perform appropriately at different scales, for Greater London or selected boroughs. They were not tested on different point datasets, but on different samples of the in-app data provided, and the results were consistent as presented in the sensitivity analysis. However, more should be done to assess their flexibility with data points of different nature, and in perhaps sparser, less activity-dense areas than London to see if the outputs remain consistent and interesting for such analysis.
- (10) The algorithm is free, and its functions detailed in this chapter. Though it was here implemented in R, it is translatable to other programming languages and does not require further software or training. The workflow has been detailed and documented to ensure it could be reproducible by interested parties.

Table 7 lists these criteria and whether they have been addressed by the three data-driven methodologies discussed over the last two chapters of this work.

Table 7. Comparison matrix assessing if the principles are met by the regionalisation methods

Principle	H3BR	Voronoi	Quadtree
Purpose	√	√	√
Relevance	±	⊗	⊗
Partition	√	±	√
Contiguity	√	√	±
Homogenous	±	⊗	⊗
Autonomous	√	⊗	⊗
Coherence	±	⊗	⊗
Conformity	√	⊗	⊗
Flexibility	√	√	√
Reproducibility	√	±	±

√	Criteria met
±	Partial
⊗	Not met

5.4.5. Summary

This chapter has demonstrated that using bespoke regions when aggregating in-app data performs better than using existing geographies at the same scale when the aim is the balance disclosure control and data preservation. An iterative regionalisation method was developed and tested based on the classification and merging of H3 hexagons. Various H3 resolutions were assessed for the selection of H310 as the atomic unit, and LSOA and MSOAs were compared to settle on MSOAs as the nesting boundary. The algorithm uses a sample of the in-app data (average counts per H310 over the month of September 2019) to identify H310 cells of low counts and merges them with a neighbour with similar land use characteristics where possible. The outputted geographies thus respect underlying land use and are nested into a set of existing geographies (MSOAs for the prototype presented here). The outcome regions were tested against OAs and OSGB250 grids at similar scale, demonstrating that the H3BRs preserve more analytical completeness, both thanks to higher granularity and a reduced amount of IDEs. The resulting maps illustrated the geographic differences between omitted cells, highlighting the regionalisation method's capacity to better represent the data patterns than its OA,

OSGB250 and H310 counterparts. This closer fit to the data can help prevent a loss of analytical validity, reducing the impact of aggregation on subsequent analysis.

Unlike census data, in-app datasets are not always internally comparable: some points may be generated by routing apps alongside roads, while others may record a single moment in time (a purchase for instance). The nature of the actions, and the frequency of data they generate, may be different within the dataset. Without zoning, activities appear mostly concentrated alongside roads, and suppression due to low counts may remove valuable information these datasets could offer. The zoning strategy developed here seeks to maximise a use of all types of data proposed by in-app datasets by reducing suppression as much as possible, regardless of the activity type. It could however be applied to a filtered dataset, where home locations can be extracted for example, or a specific type of behaviour highlighted. Furthermore, making the regions from an average, and counting data activities rather than unique devices during the H310 combination process may result in some regions remaining below a count of 10 when aggregating other days of data to the resulting regions. As previously discussed, disclosure control might not be the only consideration when developing regionalisation methods: nesting the regions inside geographies, keeping duplicates as a proxy for movement, or creating stable regions over time by using data averages may be desirable features of the regions which might result in imperfect counts of 10. Users could use the algorithm, thanks to the functions provided, with stricter conditions on the disclosure control threshold to ensure it is met every time, but this might result in larger, less precise regions. This choice can be made by users based on their region's intended usage.

A potential shortcoming of this methodology could be its oversimplicity, as the merge algorithm only considers three conditions in its decision making (activity count, land use and MSOA). Land use information was selected as the homogeneity metric, more conditions could therefore be added to the merging algorithm, perhaps in line with specific research question or future insights into the data's geodemographic characteristics. Using land use in parts was motivated by the hopes it could help factor into a metric for the creation of more compact regions, or further emphasise intra-unit homogeneity and extra-unit heterogeneity.

Previous work in this field has developed solutions to representing aggregated data distribution by using Voronoi polygons or spatial clustering methods (Sevtsuk and Ratti, 2010; Jiang *et al.*, 2019; Wang *et al.*, 2022). Both quadtree and Voronoi, previously assessed, meet fewer key

criteria than the H310 based regions. These regionalisation strategies do not encompass the representation of underlying land use and the possibility to link to pre-existing geographies. This new method sought to address these considerations, by not only focusing on a trade-off between granularity and disclosure. However, improvements are required during the merging process to generate more compact and more homogenous zones. Further work could also investigate how varying H3 scales and atomic unit types impacts resulting data counts and consider the value of having more data duplicates as a result of spatial subdivision (Cockings *et al.*, 2013). Additionally, this work focuses on the Greater London region due to data availability. It would be valuable to assess whether the concepts discussed are applicable to areas of lower coverage than Greater London, especially as this method aims to preserve areas of traditionally low counts.

A key missing component of these regions is the time dimension. The regionalisation principles set out to create regions which would have a temporal component, which would both be novel and acknowledge the importance of temporal aggregation. The regions would be different if made from data corresponding to a week of activity or an hour of activity. There are risks that aggregating shorter time periods using the current proposed regions would mean counts are still significantly omitted. This can be addressed by making regions with different data time frames, similarly to making regions with other datasets or for other, potentially less dense areas. The next part of this work will delve into this question of temporal impact, by aiming to quantify MTUP effects on the H3BR-making methodology. Proposing this assessment, as well as creating H3BRs for varying timescales, could help address this lack of temporal dimension in the current version of the H3BRs.

Nonetheless, the approach here creates output regions that minimise the impact of aggregation on analytical completeness and validity and provides non-disclosive, consistent counts that account for the context of the built environment and are linkable to other pre-established geographic units. The methodology is tractable and flexible, and removes some of the computational challenges of working with large spatial datasets. It thus offers further opportunities for accessing sensitive datasets for research purposes.

6. MTUP and times low data: assessment of the H3-based regions

So far, the research conducted throughout this thesis has been fundamentally spatial. However, over the past 40 years of geographic research, more and more data has also become temporal, this trend only accelerating with the increasing amount of passively generated location data (Peuquet, 1994; Kraak, 2000; Kitchin, 2013). For that reason, time coherence was added to the list of principles driving this thesis' regionalisation efforts (Chapter 4). Though the H3BRs do not currently have a temporal consideration built into their algorithm, this chapter seeks to address the all too important questions of temporal granularity, and MTUP effects, on regionalisation and aggregation.

This chapter thus has two key aims. The first aim is to assess whether H3BRs remain stable and usable with temporal changes. This hopes to find that *times* of low data can be preserved alongside *places* of low data through the use of bespoke regionalisation. The second aim is to harness the strengths of the H3BR reusable algorithm to provide key explorative attempts at visualising and quantifying some core MTUP effects, through incremental changes of temporal dimensions.

To do so, the chapter first presents MTUP and its effects, transposing them to the MAUP effects of scale and zone for ease of use throughout the subsequent analysis. Then, the current H3BRs are assessed at varying temporal scales and zones and compared to OSGB250, to investigate whether the advantages of the method seen previously for a day of data (Chapter 5) are maintained regardless of temporal dimension. Building from this, sets of H3BRs are created for specific hours of data (including typical nighttime hourly activity and daytime activity), to see if varying the temporal input still creates coherent and useful regions for aggregating hourly data. These new regions are tested, reusing methods employed and developed throughout previous chapters for gauging region performance. Namely, they are assessed using the Jaccard similarity index, comparison of data omission, and through assessing the differences in results obtained from conducting spatial analysis with different regions sets. For this final comparison,

points of interest (POI) data is assigned to each new hourly region set, and the resulting compositions of the space are compared, showing that different hourly regions present different pictures of otherwise identical, fixed space. Finally, the summary and discussion section of this chapter rounds up these findings, emphasising the stakes of considering MTUP with equal measure to MAUP when selecting aggregation and regionalisation scales. The hope is that this chapter both provides crucial recommendation for best practice usage of the H3BRs, and demonstrates the potential of this algorithm to be used as a tool to explore questions of MAUP and MTUP, informing future efforts in the making of a more complete space-time tessellation.

6.1. MTUP: definitions and implications for the H3BRs

6.1.1. Defining MTUP

So far, this work has focused on mitigating the effect of MAUP on the aggregation process through regionalisation. However, considering and preserving time granularity is a key element of data preservation in the case of in-app data and mobility analysis. Research in geographic information science has considered time as a fundamental element even before time geography defined it as such (Hägerstrand, 1970; Thrift, 1977; Peuquet, 1994). As discussed in Chapter 2, places are considered in time geography as a spatial and temporal intersection for daily tasks, and the cyclic habits and routines of those who use those spaces are informative of individual behaviours and characteristics (Axhausen, 1995; Kwan, 1999; Crang, 2001). Each process, each activity has its own temporal range and resolution (e.g. some cycles spanning hours, days or months) (Meentemeyer, 1989). When studying human activities in space, the temporal dimension is thus critical to consider (Kraak, 2000). In fact, it has been demonstrated that some behaviours can only be observed at certain resolutions of time (Harrower *et al.*, 2000). Changing the temporal scale of analysis can thus result in a significant loss of analytical completeness, as a whole range of behaviours occurring at these resolutions are lost to the researcher (Purdam and Elliot, 2007). Thus, we understand that, when it comes to data with a temporal element, such as in-app data, or studies with a temporal scale, ‘selecting the appropriate level of detail for a task is essential to study the right phenomena’ (Hornsby and Egenhofer, 2002; Cöltekin *et al.*, 2011, p. 1).

Though time was widely formulated as a fundamental element in spatial thinking for decades, the term Modifiable Temporal Unit Problem (MTUP) was only first formally coined by Cöltekin *et al.* in a 2011 workshop proceeding on persistent problems in geographic

visualisation. In this position paper, they identify key temporal scale effects which impact analysis and suggest categorising them under the label of MTUP (Cöltekin *et al.*, 2011). By drawing the first explicit parallel between issues of temporal scale and the famous MAUP, Cöltekin *et al.* formalise the idea that analysts must be equally aware of issues pertaining to temporal scales as they are aware of MAUP's impact on spatial analysis. They also help identify the MTUP's effect in analogy to the MAUP's zoning and scaling effects (introduced in detail in Chapters 2 and 3).

They list three main temporal effects: “*duration* (how long), *temporal resolution* (how often) and the *point in time* (when).”(Cöltekin *et al.*, 2011, p. 1). Other efforts in formulating the MTUP effects included concepts of temporal boundaries (when the time frame starts and ends), but this chapter will centre around Cöltekin *et al.*'s initial definition (Cheng and Adepeju, 2014). We find that the scaling and zoning effects which define the MAUP can be helpfully transposed to the MTUP to help quantify the impacts of these changes of temporal measurements in the context of aggregation:

Scale: the temporal scale here corresponds to the temporal resolution and duration defined by Cöltekin *et al.* Namely the ‘size’ of a temporal unit. Aggregating an hour of data rather than a day implies a change of temporal scale during aggregation, analogous to aggregating to different spatial scales.

Zone: the temporal zone corresponds to the point in time. For example, Monday and Tuesday have the same scale (a 24-hour day), but correspond to different ‘zones’ (different points in time, different days of the week). Thus, aggregating different days of data corresponds to a change in ‘temporal zones’ with this analogy.

This simplification of the MTUP allows for a more quantifiable understanding of changes of temporal dimensions in the context of the H3BR. This chapter seeks to illustrate these effects on the way H3BR are generated, and highlight why considering temporal scales is vital for the method to be generalisable to multiple uses cases and datasets. Below, we show how changing the scale and zone of the data aggregated to H3BR equates to a loss in the regions’ relevance where temporal differences are not considered, and propose ways to approach temporal dimensions when generating these bespoke regions.

6.1.2. MTUP impacts on the H3BR's performance

6.1.2.1. H3BRs: data-driven, for a day

The H3BRs detailed in Chapter 5 are made with a day of data. They are thus made at the *temporal scale* of a day, and the *temporal zone* is stabilised by using the average activity of multiple days. Thus, these bespoke regions are made to best fit a typical day's worth of in-app data. In Chapter 5, it was demonstrated that, for any day of data, the H3BR outperformed other forms of aggregation in terms of data preservation. Figure 1 summarises this finding.

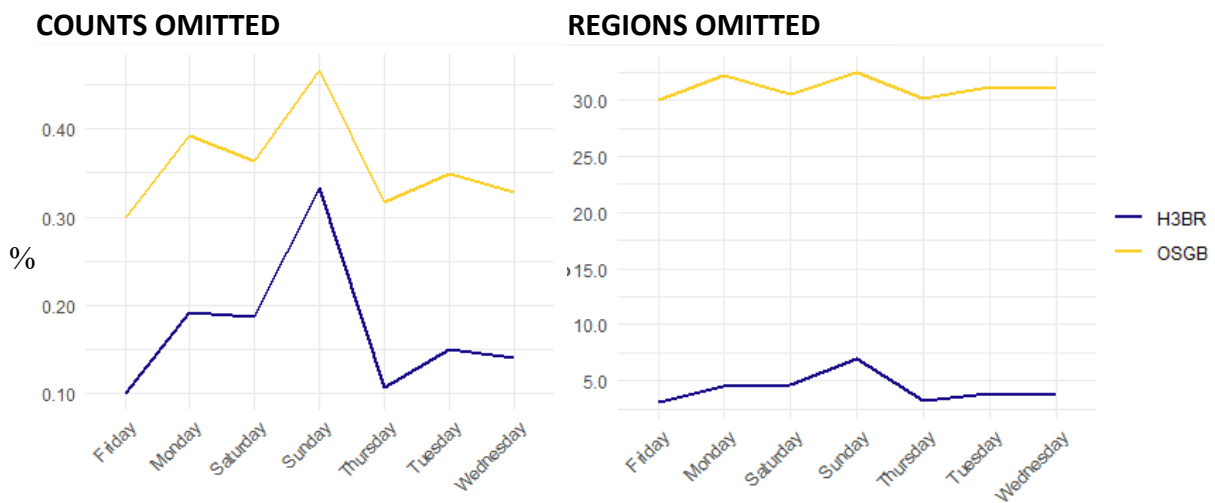


Figure 1. Comparison of proportions of counts and regions omitted for different days of data (temporal zones), for H3BR and OSGB.

When changing the temporal zone but keeping the scale stable (aggregating different days of data) we see the counts omitted are not always identical. This is expected as the regions are made based on an average day, and the omissions are due to certain days' activity being below this average day's. In fact, this highlights the relationship between a day's omission and its total contribution to the week's activity. For H3BR, that relationship displays a correlation of -0.99. Days which contribute the least activity when compared to the whole are the ones with the highest percentage of regions omitted. Though the relationship is still confirmed for OSGB, the number is lower (-0.83), indicating that more regions may be omitted at random.

Regardless, the data aggregated to H3BR experience significantly lower omissions than when aggregated using the OSGB250 grids. For OSGB, a third of regions are omitted day by day, against less than 7% on any given day for the H3BR. Sunday appears to be the day with the highest omission (2% more regions omitted on Sunday than any other day for the H3BR),

though the H3BR aggregate made based on a day of data preserves significant amounts data on this day nonetheless.

What can be expected if the H3BRs are used to aggregate and share the same in-app data at different temporal scales and varying temporal zones? In the following sections, we seek to identify whether bespoke regions made with a certain time frame of data perform better than arbitrary grids through the changes of temporal dimensions. We find that the H3BR, as they are currently defined, are only viable for the specific temporal scale (a day of data) they were built from and for.

6.1.2.2. Changes of temporal scale

First, different temporal scales are compared in decreasing order to visualise the MTUP's impact on the regions' relevance. Here, *relevance* is understood as the H3BR's core purpose of best fitting the data and preserving counts as much as possible, especially when compared to traditional aggregates. Throughout these sections, we continue comparing the H3BR to the OSGB250, which were the arbitrary units closest in size and performance to H3BR over the last chapters of comparisons related to MAUP (Chapters 4 and 5).

To assess changes of scale, we compare the data lost as we reduce the quantity of time and aggregate the data to H3BR and OSGB. For this, the activity counts (number of unique devices per area *per time*) were calculated using incrementally decreasing time blocks, starting with 24hrs of data (the activity count for each region over a day) and ending with 1hr (the activity count for an average hour of the same day). We then obtain the activity count per hour block per region, rather than the activity per day per region. Some examples of these 'hour blocks', and how they are calculated, are illustrated by Figure 2 below. This was performed using the raw in-app data points over the month of September 2019, the same data sample used to create the H3BR and much of the prior analysis conducted throughout Chapters 3 to 5.

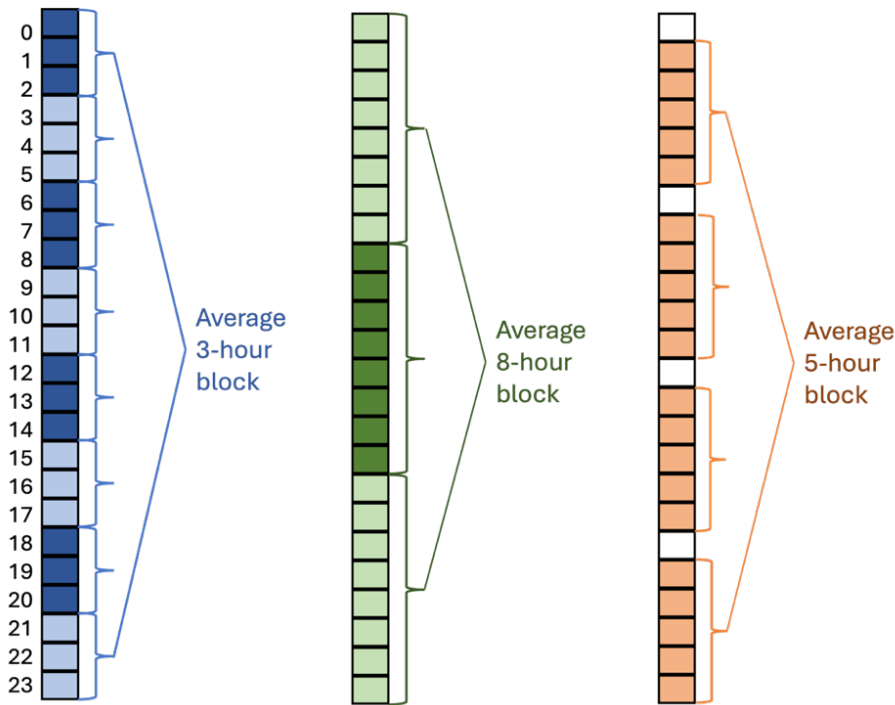


Figure 2. Illustration of the making of the hour blocks. In order to capture significant blocks of time, a rolling average is kept for the hour blocks where required. For example, for the average 3-hour block, the activity count is calculated for all consecutive 3-hour blocks starting at midnight (0-2, 3-5 and so on) and all activity across these blocks are averaged to get the ‘typical 3-hour block activity’. Where the block number is not a fraction of 24, a random hour is exempted (such as for the 5-hour block).

The final hour blocks thus capture the average (or typical) activity for their temporal scale (e.g. the average activity of any three consecutive hours). This is in the hopes of mitigating the temporal zoning effect implied in this exercise. Taking only one 3-hour block or one 12-hour block to represent their full temporal scales meant that outlier hours, and their potentially different activity levels, could be driving the underlying activity gain or loss, rather than the quantity of time itself.

Figure 3 displays the percentage of regions omitted by hour blocks (i.e. temporal scale) for H3BR and OSGB250. As demonstrated earlier, H3BRs preserve much more space when they are used to aggregate a day of data (see Figure 1 and Figure 3) (0.3% of regions omitted against 33% for OSGB). However, this advantage is progressively lost as the temporal scale decreases (shorter hour-blocks). At the 4-hour block scale, the H3BR’s trend intersects with the OSGB’s (Figure 3). At this point, H3BRs do not preserve more areas than using arbitrary grids, and thus do not maintain their relevance as an optimal aggregate for the in-app data. We can note that the variations around the 8 to the 4-hour blocks can potentially be explained by the moving averages, and those blocks potentially containing more mixes of day and night hours.

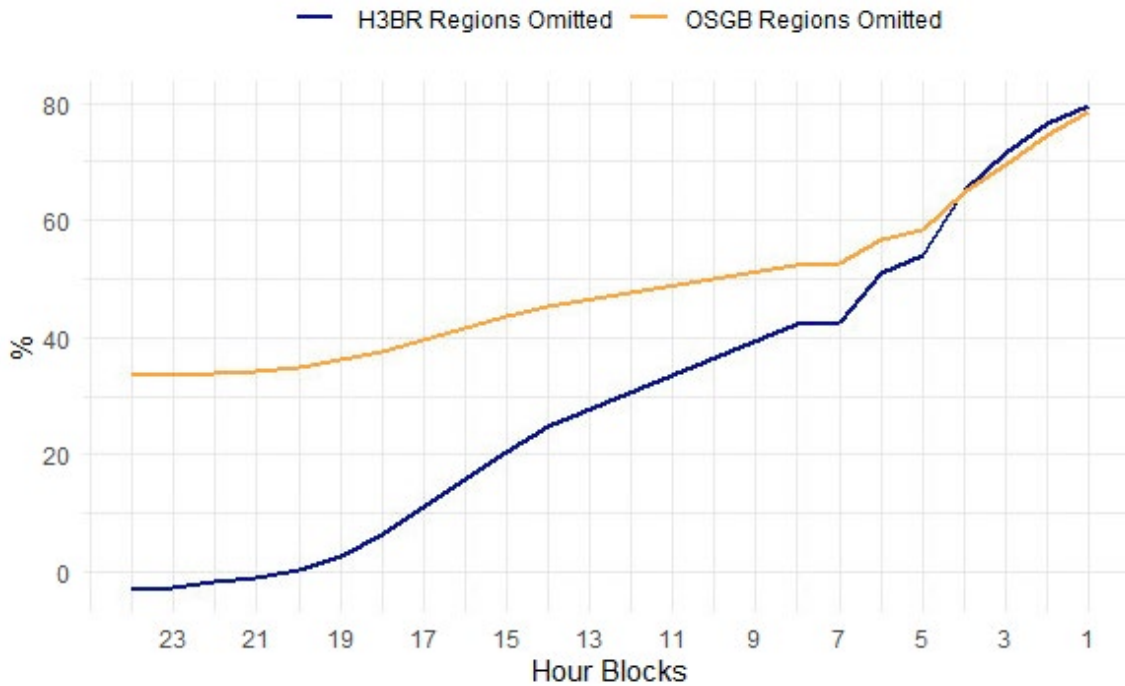


Figure 3. Percentage of regions omitted by hour-block for H3BR and OSGB250. 24 on the x axis corresponds to a full day of data.

This demonstrates a key point of the MTUP issue described earlier. A bespoke region is only the best fit at the temporal scale it is created from and for. The H3BRs’ data-driven aspects do not confer them an advantage and a guarantee to perform better than arbitrary grids at all temporal scales. Ultimately, a change of temporal scale renders them arbitrary as well, as they do not capture the data’s distribution at these scales and are made to meet a threshold of 10 for a *day* of data. However, before investigating the possibility of making hourly H3BRs, we turn towards investigating how a change of temporal zones at these scales affects the H3BR’s performance.

6.1.2.3. Changes of temporal zones

To assess the impact of changing temporal zones, a similar comparison as described above is conducted by aggregating varying temporal zones of data to the H3BR and OSGB and assessing the resulting omissions. For this, the average activity for each hour, averaged across a week is calculated. For example, we averaged the activity counts at 3 am for each day of the week, in order to control for ad-hoc events which may impact a specific day and hour’s activity. Figure 4 shows the percentage of counts and regions omitted for each hour, when aggregated to OSGB250 and H3BR.

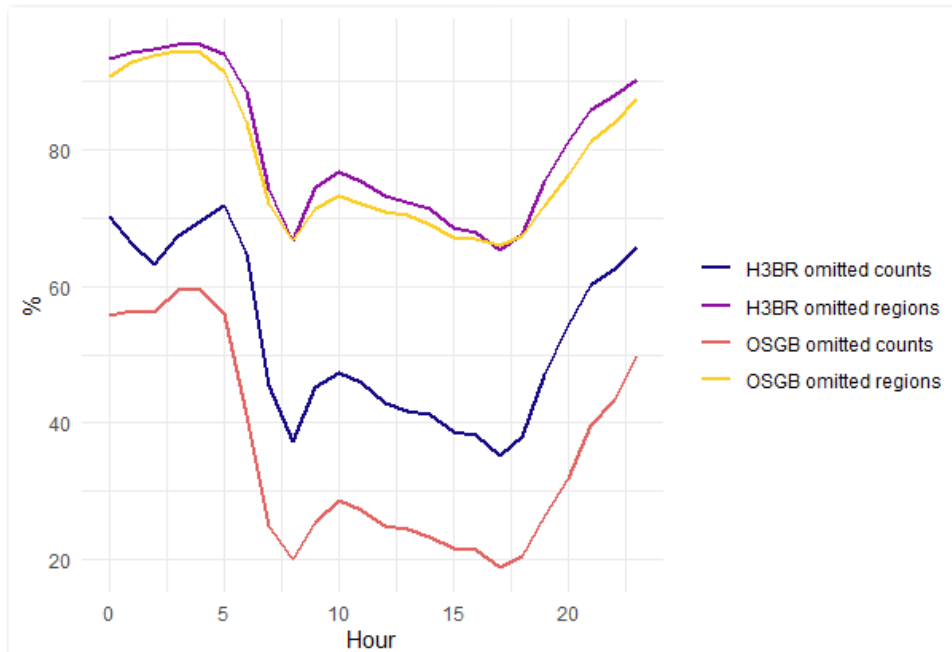


Figure 4. Percentage of counts and regions omitted per hour for OSGB250 and H3BR

The percentage of regions omitted per hour are similar for both sets of aggregates, and seem to be the reverse of the activity per hour (activity peaks expected at 8am and 5pm incur the least omission, especially when compared to the omissions of low activity nighttime hours. See later figure of activity per hour, Figure 5). However more counts are omitted for the H3BR aggregate, across all hours. This could be due to the H3BR's size. At an average area of 51,205 m², they are smaller than the OSGB250's average area of 61,068 m². Their smaller size could explain why they are less likely to meet the threshold of 10 devices per hour compared with their grid counterparts. This adds to the earlier temporal scales findings to confirm that the H3BRs, which perform so well for a day of data, are worse than arbitrary grids for aggregating hourly data. We can suppose that the omitted region proportion remains similar because there are more H3BRs than OSGB250.

The difference between hourly activity from day to day would be very difficult to analyse spatially with the regions as they are. In fact, the overwhelming majority of the data is lost at the hour scale, making it difficult for sharing any meaningful hourly data outside of a secure lab using the regions as they are. We thus cannot at this stage observe if, spatially, noon activity on Sunday is different to noon activity on Thursday, for example. This is true for both OSGB and H3BR, and at this stage, in order to analyse significant hourly data, we must either reshape the H3BR or perform analysis in controlled environment.

6.1.3. Implications

These first sections of the chapter have introduced the key terms and interpretations of MTUP used throughout the subsequent analysis. At this stage, it was demonstrated that the H3BRs, though performant for a day of data, suffer greatly from changes of temporal scales. They do not confer any advantages for aggregating data below blocks of 4 hours, and perform poorly for hourly aggregates. These exploratory analyses focused on hourly temporal zones and scales as the aim of this chapter is to investigate whether the H3BRs can perform with increased time granularity. Thus, the impact of changing temporal zones and scales by aggregating weekly or monthly averages was not analysed here. It is safe to say that MTUP has impacts in those directions as well. However, the initial regions were created with the hopes of creating smaller, more precise spatial units which would meet the threshold of 10 devices. This is less necessary for a month of data, for example, where the threshold is much more likely to be met at the H310 scale pre-regionalisation. Weekly, monthly and yearly aggregates do not necessarily require the level of regionalisation needed here for data preservation, and the MTUP effects we take interest in here uniquely concerns delineation and data omission for granular temporal scales.

When conducting analysis using external sources of data (e.g. consumer data, in-app data etc.) the decision of the temporal scale is often dependent on data availability alone (Cöltekin *et al.*, 2011). When the choice is available, it often relies on ‘trial and error’ approach, as the quantifiable impacts of the MTUP are less formalised than for the MAUP (Cheng and Adepeju, 2014). This adds further dimensions to the MAUP issues discussed throughout this thesis, as both of these problems are intrinsically linked when dealing with spatial areas built around temporal dimensions. However, owing to an access to the original in-app data points, this research has an opportunity to reflect upon the MTUP’s impact on the methodology, and to attempt to visualise and quantify recommendations of use of the H3BR based on the temporal effects described. For certain MTUP effect (particularly a change of temporal scales), there is a possibility to conduct temporal analyses of geographic processes (such a regionalisation) using H3BR as a method where controlled temporal input is possible.

The following sections thus first seek to assess the MTUP’s impact on the H3BR methodology’s volatility. Then, they investigate whether the H3BRs can be adapted for, and extrapolated to, hourly analyses. Finally, the key aim is to make analysts aware of the MTUP as much as the MAUP in the context of H3BR applications. Thus, suggestions are provided to adapt the regions and promote their most realistic and informative use.

6.2. Hour-based H3BR

Hourly data provides unique insights in time geography, rhythm analysis and mobility analysis, as detailed throughout Chapter 3. However, before using in-app data at the hourly scale, we can wonder whether the data sample is sufficient to do so, and whether, practically speaking, it is possible to make H3BRs with hourly data at the basis. First, *times of low data* are identified, investigating which hours of the day capture the least in-app data activity. These times of low data can then be explored spatially to investigate whether there are geographically identifiable and thus relevant to regionalise. Then an hourly H3BR is made for comparison with the current daily version to see whether the times of low data can be analysed through a more appropriate, time-scaled version of the regions.

6.2.1. Identifying times of low data

An initial exploration of hourly distribution is conducted to identify times (hours) of low data based on the same data sample which was used for the regionalisation process and exploratory analysis above: the month of September 2019. Figure 5 shows the average percentage activity per hour and the average percentage activity per day across London (Sieg and Cheshire, 2023).

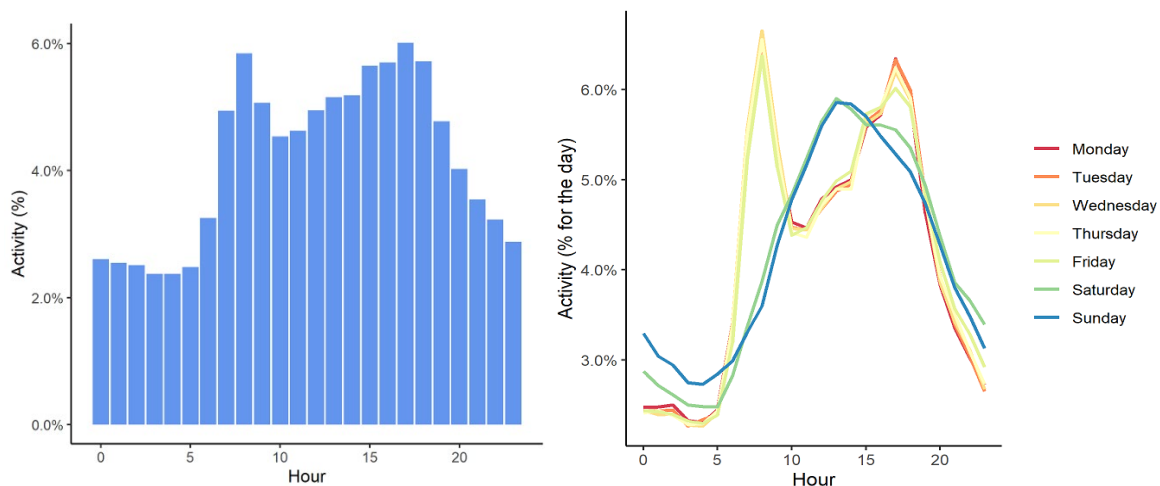


Figure 5. Proportions of hourly activity of a typical September day in London (left) and hourly activity for a typical week (right)

The second chart highlights two distinctive daily patterns: the weekday pattern (from Monday to Friday), which displays two main peaks of activity around rush hours, and the weekend pattern, which is unimodal and increases until midday. These patterns, though expected, help illustrate the data's behaviour on the London scale and show that nighttime hours are expected

to contribute the least activity regardless of weekday. However, adding together the nighttime hours (6pm to 6am), we notice their total contribution to a day's worth of data amounts to more than 40% (20% if considering only 11pm to 5am). This is a non-negligible proportion of activity which could be driven by different types of behaviours, and be differently spatially distributed. These potentially defining traits can only be investigated through careful choices of aggregation which preserve these times of low data.

In order to assess whether there are spatial differences between nighttime, daytime, weekday and weekend activities, binary classes were created based on Figure 5 results for a first inspection. Regions were attributed their peak time: if their peak activity happened during a weekday, they were categorised as a weekday unit. If it was during a weekend, they became a weekend unit. The same logic followed to determine whether a region was to be considered daytime or nighttime driven. Figure 6 maps these binary classes, with a focus on the London Borough of Camden. Most green space activity (Hamstead Heath, Primrose Hill, Regent's Park) is classified as occurring over the weekend, during the day. On the other hand, Soho is a weekend night hub. The Bloomsbury and Fitzrovia areas, mostly student and working neighbourhoods, are here dominated by weekday activities, and Camden High Street (above Primrose Hill) appears most active at night.

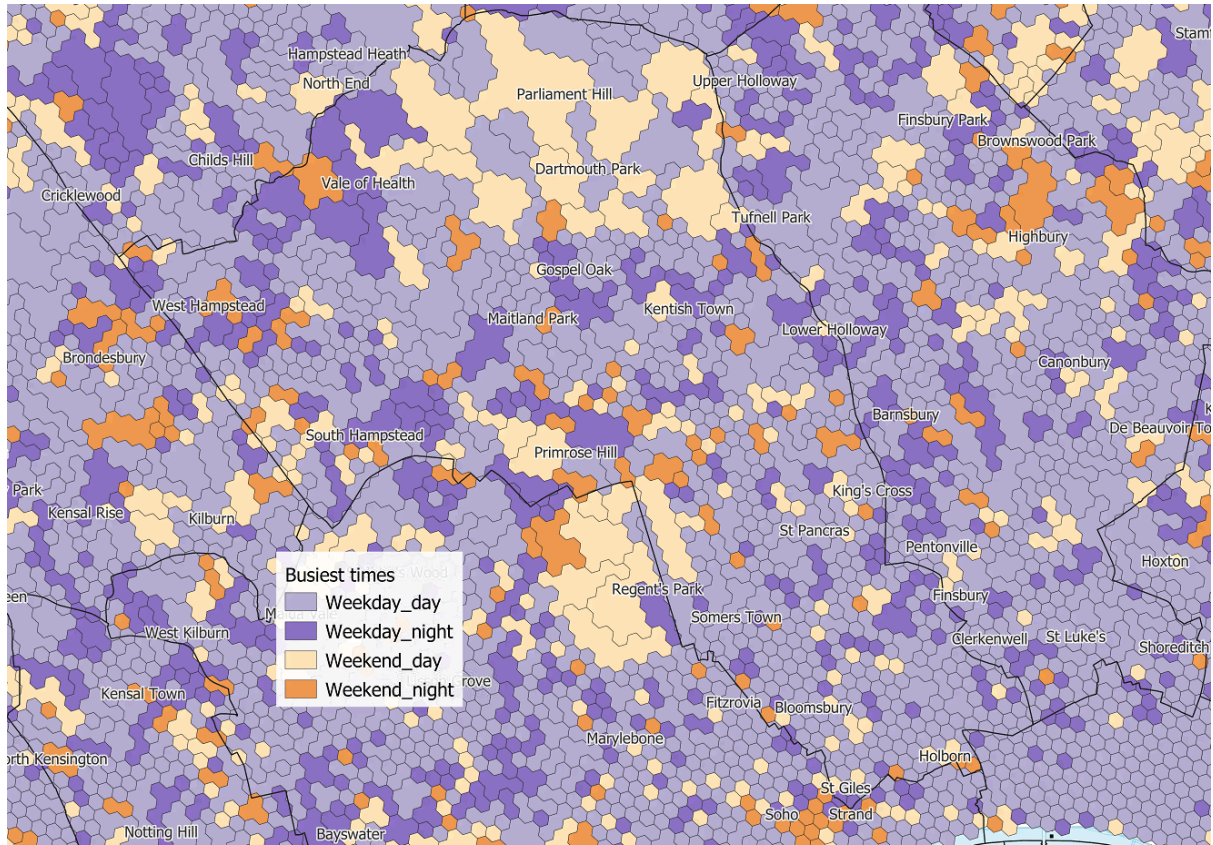


Figure 6. Map of H3BR regions coloured by time attributes. Centred on the London Borough of Camden.

The darker colours in Figure 6 indicate that the region's peak activity occurs at night. Though this type of activity does not contribute to a majority of the dataset, it is noticed that it can contribute to specific areas' principal activity. This indicates there might be some merit in investigating these times of data closer, to see what may differentiate them from others. This exploration thus confirms a purpose in making an hour-based H3BR for exploring these times of low data, or, more fundamentally, being able to delve deeper into hourly-patterns as a whole and not be constricted to daily analysis due to IDE.

6.2.2. Performance of an Hour-based H3BR

6.2.2.1. Can H3BR regions be made at the hour scale?

Throughout the rest of this chapter, the H3BRs will be referred to as *H3BR_D* for the version made based on the average day of data (developed in Chapter 5) and *H3BR_H* for the ones made based on an hour-scaled sample. A key thing to consider before making *H3BR_H*, is whether the current methodology and choices are applicable to the temporal scale of an hour.

Before generating the previous H3BR_D, an administrative boundary was chosen to contain the iterative algorithm but also ensure linkage with other datasets. The choice of MSOAs as this container was informed by the in-app data counts within MSOAs. If a large number of MSOAs had not met the threshold of 10, it would have been impossible to subdivide them into smaller regions, as these would certainly have not met the threshold (as detailed in Chapter 5, Section 5.1.1.3.).

Thus, for H3BR_H, a similar exercise is conducted, to confirm whether the current H3BR-making algorithm can be applied to an hourly sample of in-app data. This is done by aggregating the same hourly data as in Section 6.1.2.3 (exploring changes of hourly zones) to MSOAs instead of H3BR and OSGB, and counting the omitted MSOAs per hour. Figure 7 displays the results.

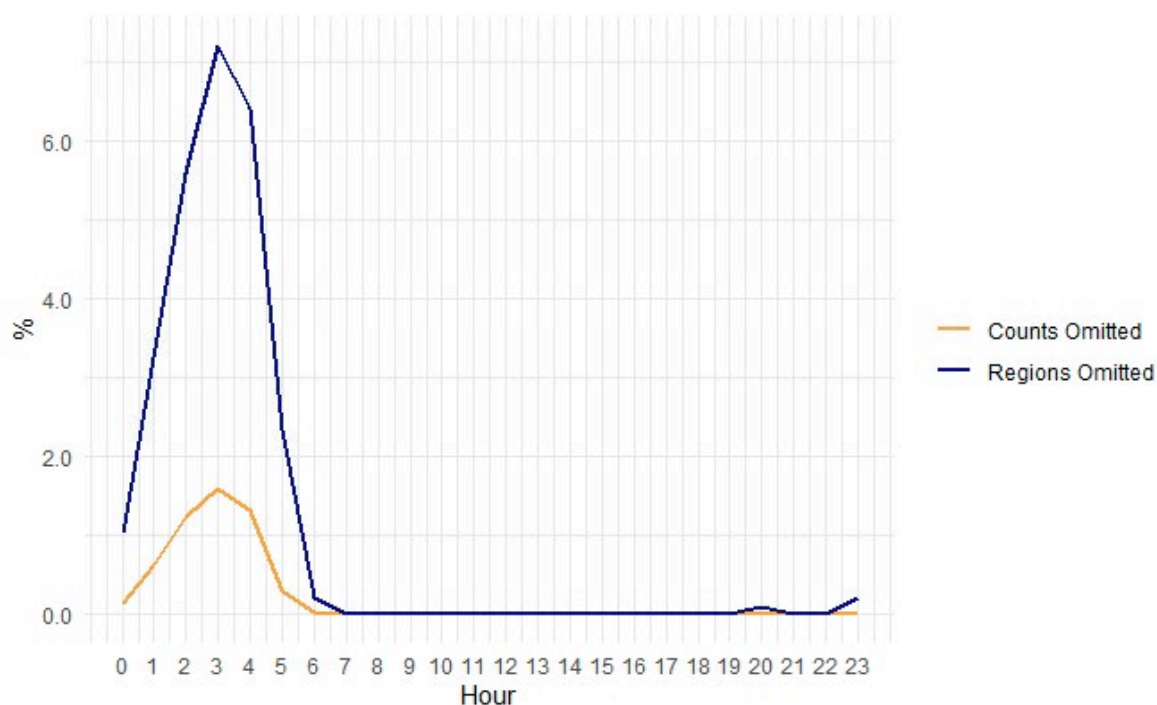


Figure 7. Counts and regions omitted per MSOAs per hour.

For 16 out of 24 hours, all MSOAs meet the threshold, and none are omitted (Figure 7). However, at 3 am, 71 MSOAs are removed due to IDE (activity < 10). This represents 7% of Greater London MSOAs. Despite these omissions, the average activity per MSOA at 3 am is of 35 devices. This means that, on average, each MSOA could be divided into 3 smaller regions at 3 am, though 71 of them would remain undivided. Overall, the results are encouraging: the average activity per hour per MSOA is of 152 (implying 15 regions per MSOA on average at any given time), with some hours seeing very high counts throughout (256 devices per MSOA

on average at 5pm). In light of this test, we can conclude it is possible to create an average hourly H3BR (H3BR_H) nested within MSOAs, and perhaps even investigate what a bespoke 3am set of regions would look like compared to a 5pm set, for instance.

On top of confirming the plausible use of MSOAs as the merge boundary for H3BR_H, this test provides two key insights for the following steps. Firstly, the mean activity per MSOA is lower for H3BR_H than it was for H3BR_D confirming that the H3BR_H will be larger than H3BR_D (the mean activity per MSOA for a day of data was 192, see Chapter 5). Secondly, the large difference in activity (and omissions) between hours (Figure 7) indicates that the data is negatively skewed by the most active hours. Thus the “typical hourly regions” should be made from the median activity across all hours rather than the mean, as the median is less sensitive to outliers.

6.2.2.2. Comparing H3BR_D and H3BR_H

The same data sample as H3BR_D was selected for the making of H3BR_H: September 2019. For each hour, the mean activity per hour across the month was retrieved, and the median of all those hours was kept for making the regions. Thus, the regions are made with a “typical hour’s” worth of data across the month. As per with the H3BR_D, this typical hour’s activity is calculated per H310 cell, linking median hourly activity to the H310 lookup tables created in Chapter 5 with the terrain information and MSOA labels for each H310. This data, with the median hour activity and all contextual attributes, is then run through the PNMx algorithm to join all H310s of less than 10 devices. As it requires more iterations to reach the threshold due to the lower starting counts, the algorithm ran for around 20 minutes on a single threaded Intel 64 processor, 5 minutes longer than for the H3BR_D. The resulting regions are roughly twice as big as the H3BR_D on average, though some areas with high activity remain spatially granular. Figure 8 presents a side-by-side comparison of H3BR_D and H3BR_H, still with a focus on the borough of Camden. The statistics provided with Figure 8 are for the Greater London area, describing the number of regions and average unit area for both versions of the H3BR.

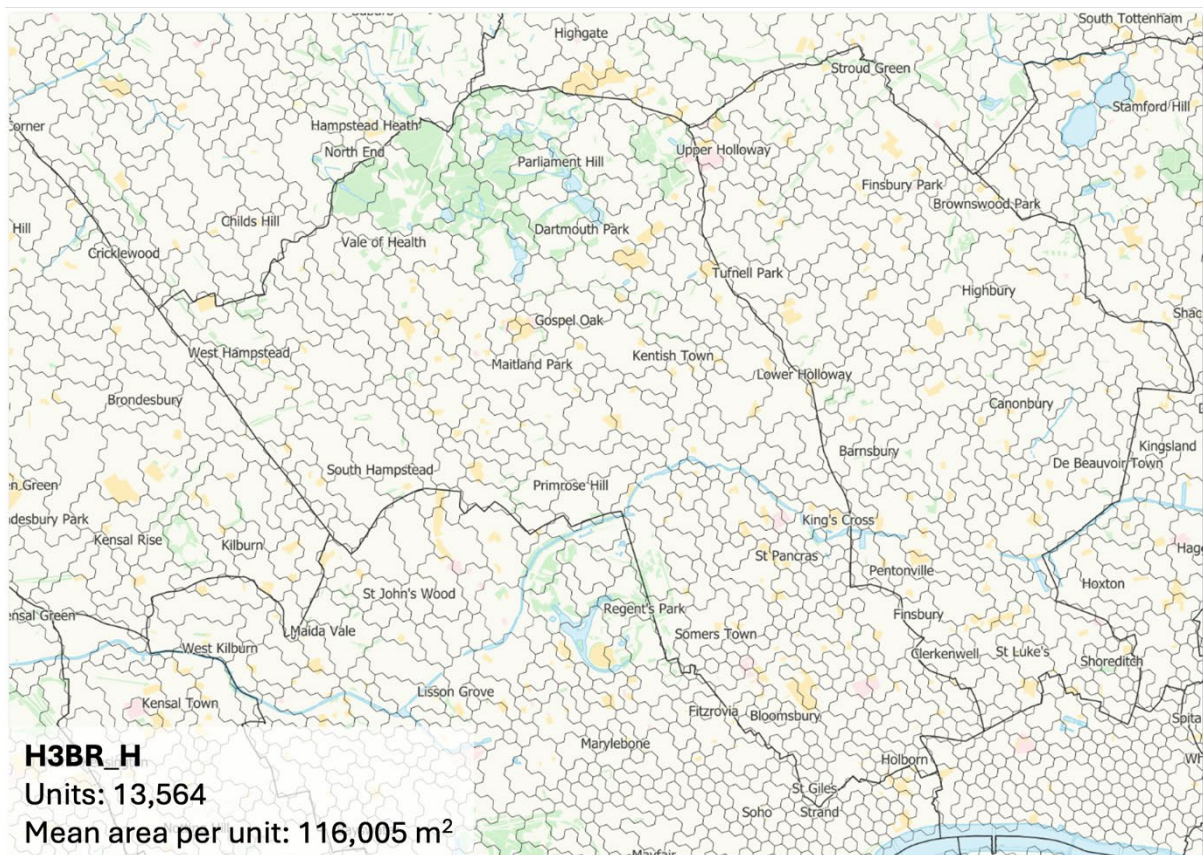
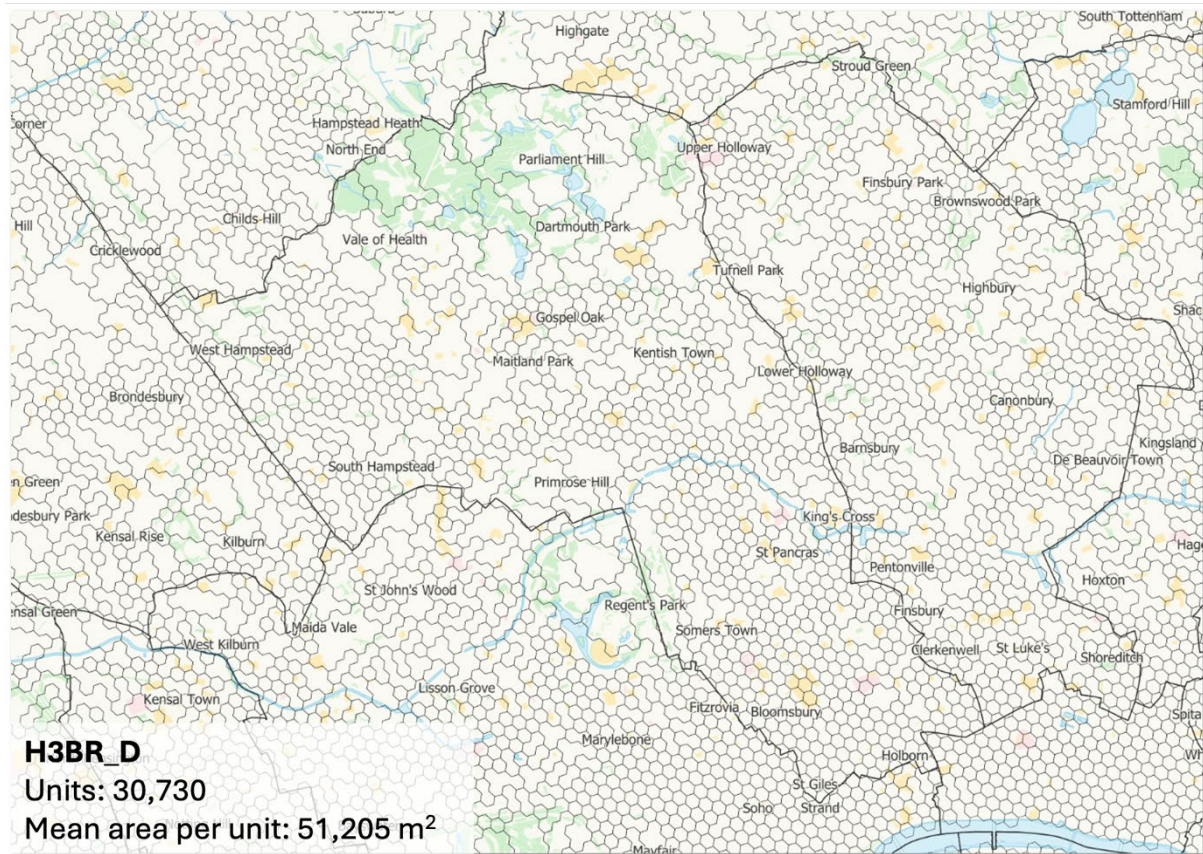


Figure 8. Comparison of H3BR_D and H3BR_H at the London scale, with a focus on Camden and surrounding areas for ease of comparison of the partitioning differences.

The H3BR_H are thus spatially coarser than H3BR_D, but hypothetically more appropriate for hourly analysis, which constitutes an interesting compromise between spatial and temporal granularities. Using the H3BR_H for daily analysis would not be optimal, as they are not as granular as the daily version and thus information is lost where the data counts would otherwise allow for a higher level of detail. The H3BR_H could also be arbitrary in light of daily data distribution, as can be assumed from the daily versions being unsuitable for hourly data.

The H3BR_H are then compared with OSGB 250 and H3BR_D, following the same process as Section 6.1.2.3. Hourly data was aggregated to H3BR_H, and the omitted counts and regions were returned and plotted alongside the previous results of Figure 4. This is displayed by Figure 9 below.

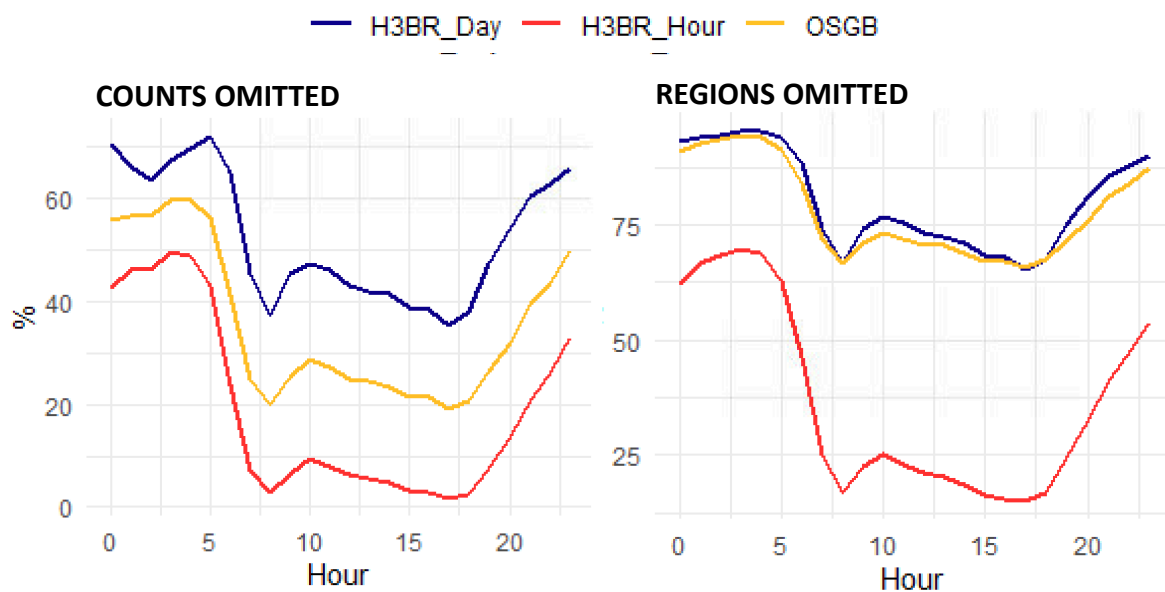


Figure 9. Comparison of percentage omitted counts and regions per hour - H3BR_H added

A clear improvement is seen from using H3BR_H compared to H3BR_D and OSGB250, especially in daytime hours, perhaps busier than the median activity used to delineate the regions. In fact, during busy daytime hours, the percentage of regions omitted is almost 50 points lower than the region omission incurred by OSGB250 and H3BR_D (20% omission against 68%). This gap is narrower for times of low data defined previously. At 3am, the time with the least activity throughout the dataset (See Figure 5), the H3BR_H still omit close to half of the counts and 69% of regions due to not meeting the threshold. These results highlight two main findings. Creating an H3BR_H indeed makes hourly data more exploitable, by significant margins across all hours, however, that margin is much less conclusive for times of low data at

this temporal scale. We remember that, for daily data, Sunday experienced a slight increase in omission compared to other weekdays when aggregated at the H3BR_D made with a mean day of data (Figure 1). This is due to Sunday being the day with least data, but this difference in omission proportions was not as drastic at the day-scale as it is at the hour-scale.

Nonetheless, an average or median-based region is helpful to conduct hourly comparisons throughout the day. More specific regions are tested later to assess whether more precise zoning is helpful to preserve data at lowest times, but first the H3BR_H are applied to explore daytime activity against nighttime activity overall.

6.2.2.3. Using H3BR_H to map daytime and nighttime activity

To expand on the previous exploratory analysis of daytime, nighttime, weekend and weekday regions, we use the H3BR_H to visualise activity levels during the night and the day. Here, nighttime corresponds to 11pm to 5am, the hours with the lowest activity, which we could identify from Figure 5 displayed very different levels to daytime hours. Here, distinguishing nighttime and daytime by stretches of 12 hours (6am to 6pm being daytime and 6pm to 6am nighttime, for example) would not capture the realistic activity during the times of low data of the night, with the activity potentially being driven up by the busiest evening hours.

Figure 10 overlaps the daytime activity and nighttime activity in a bivariate choropleth map, with the activity levels aggregated at H3BR_H. The activity levels correspond the average hourly activity over the respective time periods. Though the levels of activity are lower for nighttime, using H3BR_H allows us to visualise nighttime activity and plot it alongside daytime hours for comparison. At a glimpse, we can identify areas which are busy throughout the 24 hours of a day, and distinguish those that are notably active during the day or night only.

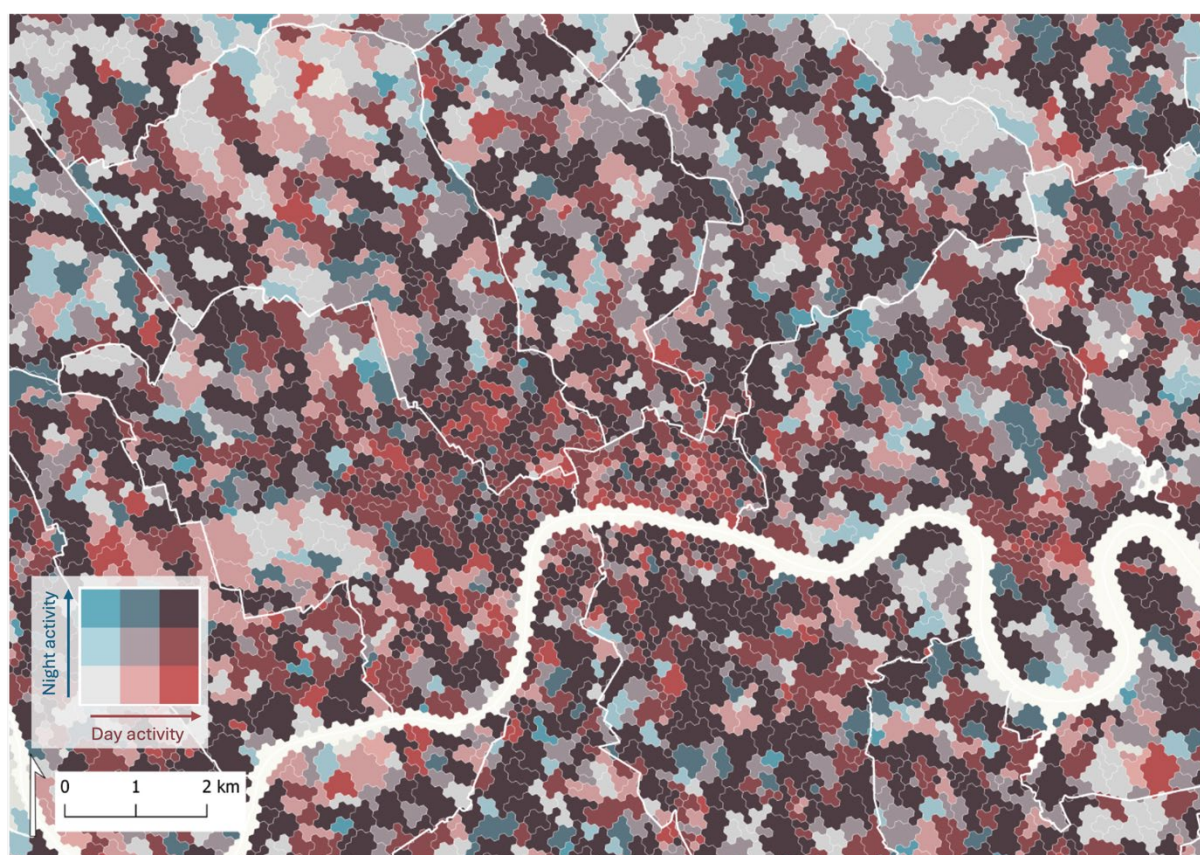
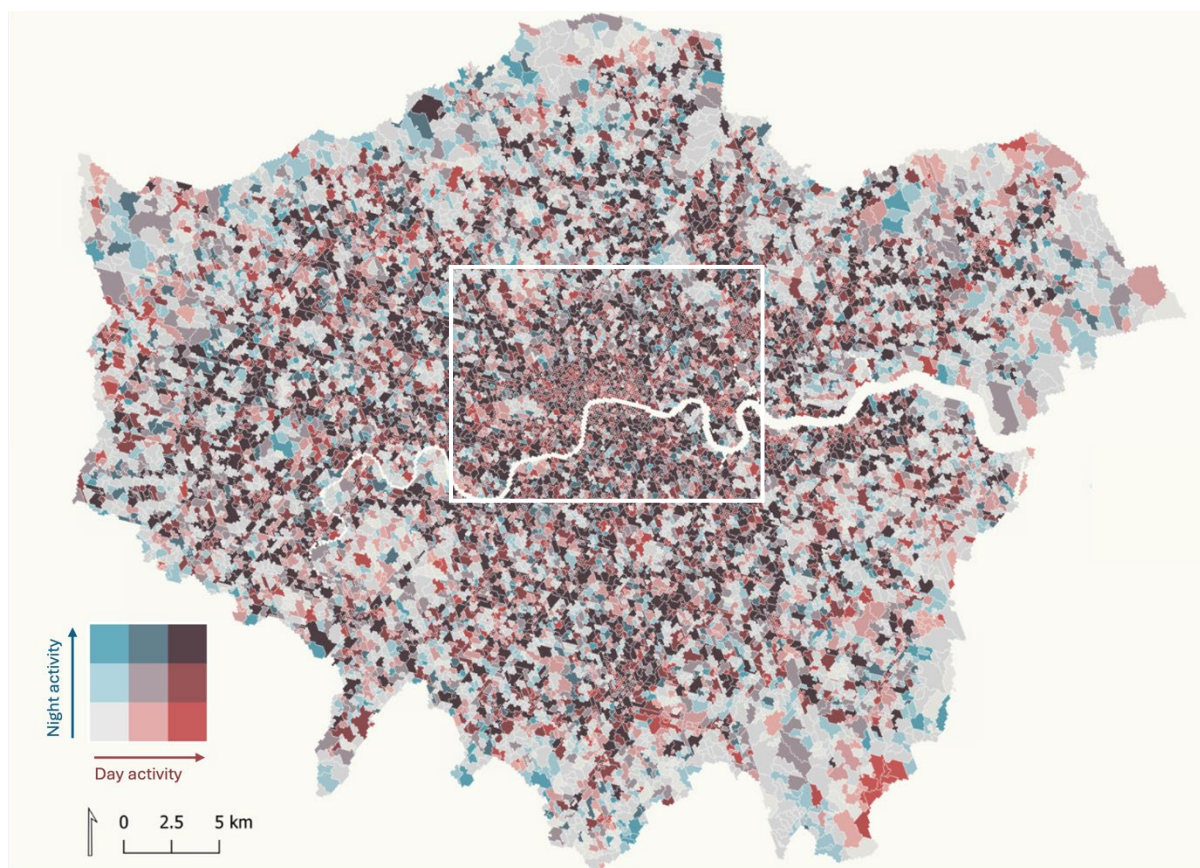


Figure 10. Bivariate choropleth maps showing night and day activity levels across Greater London, aggregated at H3BR_H. Night activities range from 10 to 100 and day activities 10-150. Greys are under 10. Boroughs are outlined in white, and the zoom focuses on central areas.

With nighttime activity being overall much lower than daytime, one question persists as to whether the H3BR_H, which are drawn based on an average hour across the day, capture nighttime activity closely enough. The hypothesis here is that the H3BR_H, built on the median, are closer to daytime activity and the areas driven by it than nighttime activity, which might remain prone to omissions. Additionally, areas of the map with significant activity for both may have different shapes for night activity and day activity, if those two dimensions were to be distinguished in the regionalisation process. This sparks the question of how the same space may be partitioned differently between different times outside of scale. We refer to this as *thematic regionalisation* through the following sections (i.e. choosing the time dimension based on the field question or theme of research rather than only data availability).

6.3. Times of low data and thematic regionalisation

6.3.1. Controlled temporal input: making regions for specific hours or use cases

This section seeks to assess the value of making very specific regions in cases of low data, or to research specific phenomena (such as night activity levels against day activity). It also seeks to assess, through controlled temporal input, the effects of MTUP on the regions created. Following previous findings, questions remain about the fit of ‘generalised’ regions at the hourly scale. Though the daily average regions (H3BR_D) did not discriminate significantly between days when aggregating different days using the regions, the hourly median set (H3BR_H) seems to be more skewed towards times of higher activity, as the activities per hour range more widely than activities per day.

Would more targeted regions help mitigate this, and would these have any relevant analytical purpose in their specificity? What would come out of making ‘typical nighttime hour’ and ‘typical daytime hour’ sets of regions, anticipating their differences? In this section, it is shown that specific regions improve upon the omissions, but that this comes at a trade-off for interpretability and comparison potential across times. After detailing the rationale behind the making of the specified units and comparing them statistically, recommendations are proposed regarding the approach to take when choosing a temporal scale for the making of H3BR in the context of future analyses using in-app data aggregates.

Four additional sets of regions are made to compare with H3BR_H. Their fit to specific times of low data is assessed:

- (1) **H3BR_3am**, built from the average 3 am activity.
- (2) **H3BR_5pm**, built from the average 5pm activity.
- (3) **H3BR_NT**, based on average hourly nighttime activity (11pm to 5am).
- (4) **H3BR_DT**, based on average hourly daytime activity (6am-10pm) (not to confuse with the earlier H3BR_D, based on a full day on data).

The H3BR_5pm are made to compare the busiest hour (5pm) with the quietest one (3am) and see how they compare to median hour regions and night or day ones.

Following previous region-making methodologies, activity counts for each temporal scale and zone above are computed at the H310-scale, joined with the H310 attributes and run through the PNMx to assign merge neighbours and reach the 10-device threshold. Each timeframe tested (3am, 5pm, night average, day average) produce very different mean activity per H310 atomic unit prior to aggregation (Table 1). This already forecasts notable differences between the regions resulting from each sample as the algorithm follows device counts as its first and most important criteria.

Table 1. Descriptive statistics of the specific hourly H3BRs. Mean per H310 prior to regionalisation, and number of resulting H3BR units post regionalisation

(H3BR_)	3am	5pm	NT	DT	H (median)
<i>Mean activity per H310</i>	0.77	3.34	0.91	1.86	3.01
<i>Number of H3BR units</i>	3666	14734	4296	9025	13564

From Table 1, we see that the mean nighttime activity per H310 is slightly below one device per atomic unit, more than twice as less as the mean daytime activity (1.86 devices per H310). The resulting regions are thus twice as big for H3BR_NT. Regions have to recombine more often to reach a threshold of 10 for a typical nighttime hour. Interestingly, the 5pm regions are closer to the median regions than they are to the daytime ones. This indicates a possibility that 5pm (the busiest hour of the day) drives much of the H3BR_H's shapes due to its high activity. The hypothesis at this stage would be that 3am data looks suited to the nighttime regions: both have a mean activity of less than one device, and their region count is much closer than when compared to the other sets. To assess these differences, we compare each set of regions against one another using the Jaccard index introduced in Chapter 5. This allows for a quantitative

measure of similarity, inspecting if certain areas remain stable throughout the temporal zone changes, and whether the more ‘generalisable’ ones (nighttime, daytime, median) are spatially comparable to more specific times such as 3am and 5pm and useful to aggregate these hours.

6.3.2. Jaccard comparisons of resulting regions

As discussed above, the time samples used in making the 5 regions compared return very different counts per H310. These differences in turn result in drastically different regions, especially when the counts hover around the threshold much more than they did for a day’s average of data. In Chapter 4, the Jaccard relationship between regions made with different days of data was between 0.25 and 0.40 depending on the day (meaning roughly a third or more of the regions remained the same if made with any day of data). We can expect that, for hourly data, the Jaccard indexes will be much lower. In fact, 5pm data records more than 4 times more activity than 3am data, which will inevitably result in drastically different regions. This is exacerbated by low overall hourly counts. If hourly counts were often of 10 or more activity per H310, as they are for daytime data, more regions would remain unmerged, and thus identical between hours. Instead, when most, if not all, H310 are below a count of 10, an additional count in any cell may have a significant impact on the number of iterations needed to meet a threshold of 10. This is illustrated by Figure 11, and constitutes a key weakness of applying H3BR methodology at times of low data, as it increases the method’s volatility. In fact, Figure 11 illustrates in parts the impact of MTUP on the H3BR making methodology, showing that the notion of threshold can disproportionately impact data surrounding said threshold. The H3BR are adaptable and easy to apply, but it is important to note their increased volatility in areas and times of low data.

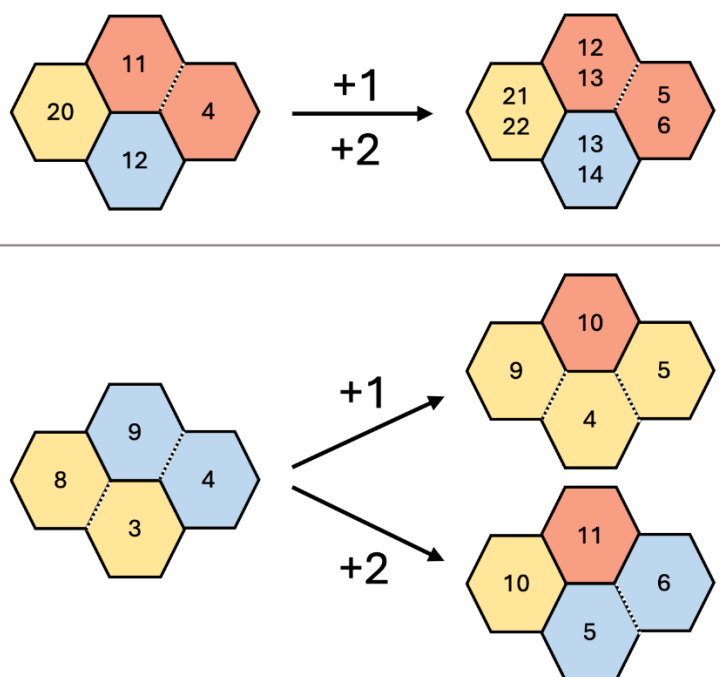


Figure 11. Diagram simplification of region reassignment when data is added. In the top case, regions with data counts largely above 10 will see little difference if an activity is or two is added. This is similar to comparing two days of data. However, starting H310s with low data, particularly counts hovering around 10, will be creating very different output regions if one count or two are added. This simulates what can happen when one hour sees slightly more counts than a previous hour when regionalising both separately, as the starting counts are lower, and the differences larger than between days.

As explained by Figure 11 samples with higher counts do not recombine as much, and are less prone to be impacted by small differences in data. As the hours sampled for the regions tested here are low in counts and vary by three folds or more between them, we can expect the Jaccard similarity indexes to be much lower than they were between days of data. Nonetheless, we can observe their relative relationships to assess the level of overlap between the regions. Table 2 displays the Jaccard similarity indexes between each of the four sets.

Table 2. Jaccard similarity indexes between hourly region sets. Colours correspond to relative similarity as indicated by the legend.

	5pm	3am	NT	DT	H	6pm
5pm	1					
3am	0.03	1				
NT	0.04	0.1	1			
DT	0.06	0.03	0.04	1		
H	0.2	0.05	0.04	0.08	1	
6pm	0.2	0.03	0.03	0.08	0.2	1

>= 10%

< 10%

< 5%

The 6pm results correspond to regions made with 6pm data, following the same methodology as the H3BR_5pm. These were made to get a sense of similarity between two adjacent hours of similar activity levels, demonstrating the above discussion regarding volatility at the hour scale (Figure 11). Despite their proximity in time and activity counts, H3BR_5pm and 6pm only share 20% of their units when regionalised, something to keep in mind when looking at the other relatively low results.

The focus of this analysis is the H3BR_3am, the time of lowest data. H3BR_3am share little similarities with other regions, but encouragingly, their closest counterpart are the NT (nighttime) regions. About 10% of their units are identical. This implies that aggregating 3am data to H3BR_NT could preserve more data than aggregating them to H3BR_H. 3am and 5pm regions are the ones which share the least overlap, which is to be expected given their drastically different activities and expected data distribution. Thus, in order, H3BR_3am are closest to NT first, then H, with DT and 5pm tied last.

The highest Jaccard indexes seen in Table 2 are between the H3BR_5pm and H3BR_H (median hour). This is interesting as one could have first hypothesised that 5pm would also share more similarities with DT (daytime) regions. However, looking at activity counts per H310 earlier (Table 1), it follows logic that the 5pm regions have more in common with H3BR_H.

It may not always be sustainable nor interesting to create new regions for every hour of data, particularly in cases where the spatial units must remain stable for comparison over time. Though the regions are delineated very differently between the NT and DT sets and share few identical regions with the specific hourly ones (3am, 5pm), can they be used to aggregate and potentially preserve counts compared to using the H3BR_H alone? Would aggregating nighttime hours to an average nighttime set of regions truly help preserve relevant information? The following analysis takes a sample of data from a single day (rather than hourly averages over the week of data) and aggregates specific hours using the thematic regions to compare against H3BR_H.

6.3.3. Sunday's 3am and 5pm omissions by regions

As the 3am and 5pm regions are made based on the 'typical' 3am and 5pm values (average of all 3am hours across the week, for example), all bespoke regions for these hours do not automatically meet the threshold when aggregating a random 3am or 5pm sample. Looking at

earlier findings regarding activity distribution across the day and week, we see that Sunday is the day with least activity. Data for 3am and 5pm on Sunday the 8th of September 2019 was thus extracted and aggregated at their respective regions to compare omissions. The samples were also aggregated to H3BR_H, and H3BR_NT (for 3am data) or H3BR_DT (for 5pm data). The idea here is to ensure that regions with higher similarities in the earlier Jaccard tests would also omit less data when used for aggregation. This can help inform the balance to strike between creating too specific a set of H3BR, which would be hard to compare or generalise, or use a version which might not be the best fit for the research question or data sample. Figure 12 displays the resulting omitted counts and regions for Sunday 3am (left) and Sunday 5pm (right) for each type of H3BR tested.

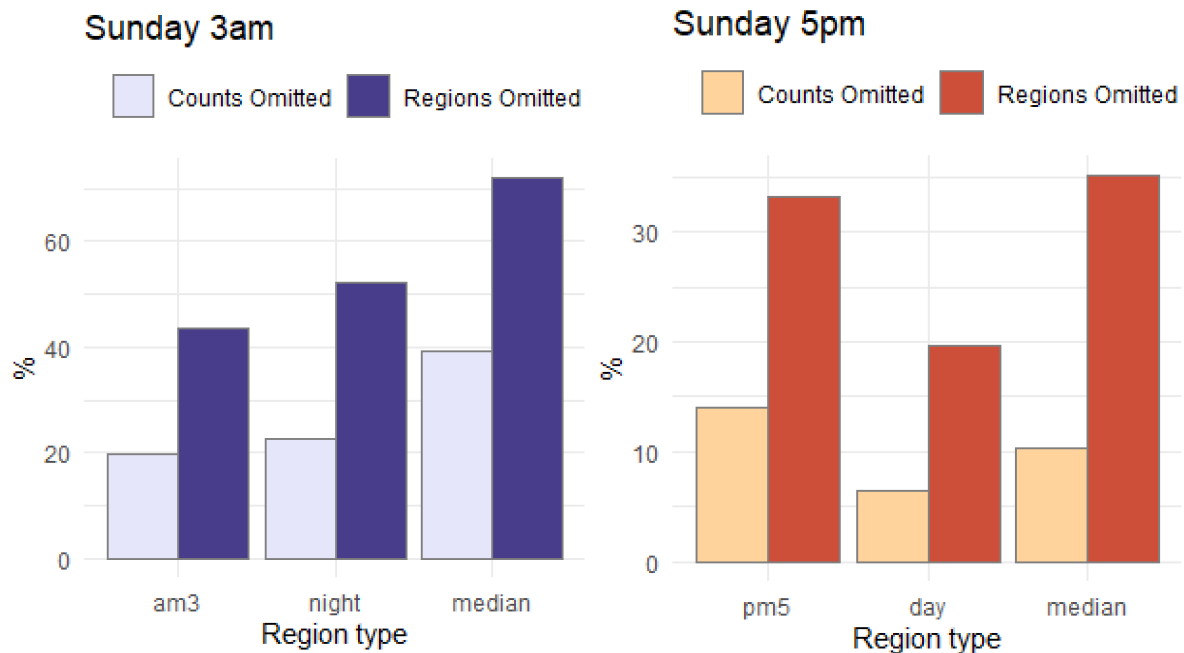


Figure 12. Bar plots showing the proportions of omitted counts and regions per region type when aggregating 3am (left) and 5pm (right) Sunday data.

For Sunday 3am data, the H3BR_3am are unsurprisingly the regions which omit the least data. The omission, however, remains relatively high, with 3am being a time of such low counts. when using the closest region, 20% of counts are omitted, with over 40% of regions omitted. However, this is a net improvement compared to H3BR_H, which omits 40% of counts and 70% of regions, as was illustrated in Section 6.2.2.2 (Figure 9). Using the H3BR_NT helps preserve close to 20% of counts and regions compared to the median hour regions, and they only perform slightly worse than the bespoke 3am regions. In this context, the H3BR_NT are a good choice. Significant context is gained from using regions which are not too specific (unlike single-hour

regions), as they allow for wider use, and the H3BR_NT provide an advantage over H3BR_H in terms of data and space preservation.

The results for Sunday 5pm are, at first glance, much more surprising. In fact, the bespoke 5pm regions perform less well for Sunday 5pm than the H3BR_DT. This can be explained by Sunday's 5 pm activity being atypical of the usual 5pm activity driving the H3BR_5pm regions. In this case, the H3BR_DT preserve more data, meaning Sunday 5pm activity is closer to the average daytime hour in terms of activity distribution. Similarly to Sunday 3am, the H3BR_H are the regions which see the highest omission; though we can note the proportions of omissions are two times lower for Sunday 5pm than 3am.

The unexpected result of the Sunday 5pm data not being best preserved by 5pm regions reflects the intricacies of MTUP. When making precise single-hour regions, one must also take into account weekly patterns, as they also constitute changes of zones. The recommendation is thus to privilege making regions based on stable medians or averages, preferably over longer periods of time rather than a single hour (such as 3am or 5pm). The making of the most specific regions (H3BR_3am, H3BR_5pm) is to be considered for times of low data where the count preservation is crucial and a core priority of the regionalisation. Otherwise, one is exposing themselves to MTUP-driven inconsistencies without notable advantages.

As demonstrated earlier, the H3BR_H do help preserve some data when compared to H3BR_D or OSGB250. This attests for the possibility to apply the H3BR methodology at changing scales. The question of zones then relies more heavily on field knowledge and regionalisation aims. The omissions comparison above highlighted the advantages of making thematic regions based on relevant temporal zones such as day or night, as these preserve data and remain more stable than specific hour-based ones. However, these must be considered with the considerable assumptions they carry. For instance, in aggregating 3am data to H3BR_NT, we assume that 3am data is most closely represented or captured by nighttime regions over daytime regions, something we felt safe to presume from initial exploratory analyses of hourly data distributions. This could be untrue depending on data samples and specific events occurring on specific days, and must be taken into account in the decision making.

Controlling the temporal input and testing its impact on data preservation helps build a more precise picture of realistic H3BR applications. An advantage of the region making methodology is its versatility and efficiency in producing the different sets of regions compared here. These

demonstrations help make a case for incorporating the H3BR into an aggregation querying system, where in-app data users outside of secure data environments could request data at specific temporal scales and retrieve ready-made aggregates corresponding to their needs.

However, the performance of each time-based H3BR was only assessed with regards to data omission as specific times. The impact of MTUP is crucially spatial as well as temporal, and the drastic difference in the regions sets impacts in the way space is interpreted by each of them. This is further explored below, in an attempt to illustrate the indissociable effects of MTUP and MAUP.

6.4. MTUP and dynamic reconfigurations of otherwise fixed space

The H3BR methodology inputs dynamic information (in-app data over time) to create a static snapshot (region outlines). However, as shown above, different hours of data produce different shapes of regions. This implies that, when changing the region's data times, there can be a shift in the way underlying static space is delineated and interpreted.

So far, this chapter has focused on omission of data counts and spatial units as a metric to assess the impact of temporal scale and zone changes on the H3BR methodology. In fact, throughout most of the thesis, the measure of success of regionalisation was largely quantified through omission reduction. MTUP has mostly been addressed as an issue of data preservation and method volatility throughout the chapter. However, changing the partitioning of space, whether due to temporal choices or spatial ones, can impact the interpretation of otherwise fixed space (Openshaw, 1979; Viegas, Martínez and Silva, 2009; Cheng and Adepeju, 2014). In Chapter 3, this was explored using LOAC classifications, a demographic descriptor, illustrating the ways in which the change of spatial zones and scales impacts the proportions of activities assigned to various populations. Here, the aim is to go one step further and demonstrate that static spaces (such as terrain, buildings), rather than population characteristics, are also subject to this effect. This is situated in the context of MTUP effects by assessing those changes based on changes of temporal units when regionalising.

This section takes points of interested (POI) data to assess whether the region compositions change between different versions of the hourly H3BRs. The aim is to assess whether the

changes in region numbers and shapes seen above drastically impact the way the city is interpreted, or whether, at the scales at which H3BR operate, these effects are negligible.

6.4.1. Methodology

For this assessment, we compare H3BR_H, H3BR_NT and H3BR_DT. These regions are chosen as they capture different levels of activity, and it can be assumed that their underlying trends are driven by different behaviours (Kim, 2020; Nemeškal *et al.*, 2020). Inspecting earlier results of using the H3BR_H to compare nighttime and daytime (Figure 10), it was noted that certain areas of London appeared to have significant levels of activity for both time frames, and questions persisted as to whether the same area would see different partitioning imposed at these varying times. Furthermore, using H3BR_NT and H3BR_DT is more informative than comparing the 3am and 5pm regions which were shown to be more volatile and appropriate only for very specific use cases. Thus, H3BR_H, NT and DT regions are linked to POI information, and the composition of the regions are compared to see how each capture the underlying space.

6.4.1.1. Data presentation: AddressBase Premium POIs

For this, we use the POI classifications from the Ordnance Survey's (OS) AddressBase Premium dataset. AddressBase datasets are derived from multiple authoritative sources including the Royal Mail's Postcode Address File (PAF), local authority data, Ordnance Survey mapping data, and valuation office data. They are designed to support various applications ranging from emergency planning and services, delivery logistics, infrastructure development, and government services. The AddressBase Premium (ABP) is the most comprehensive dataset offered by the OS, with information on objects without postal addresses (such as parks, open spaces etc.) and historical addresses.

We use the 2021 ABP classification scheme which provides a category description to each POI registered. The ABP shapefile contains coordinates for each POI along with different levels of classification. Table 3 lists the classification's primary and secondary codes selected for this analysis, along with their category descriptions. The full list of classifications is provided in Appendix 3.

Table 3. Primary and Secondary POI classification categories as described by the OS ABP dataset. Classes which have low presence across London, or low relevance to the analysis,

were not considered in the profiles. All 58 classes, including the ones not considered here, are listed in Appendix 3.

Primary category (Class 1)	Secondary category (Class2)	Description
C (Commercial)	CA	Agricultural
	CC	Community Services
	CE	Education
	CH	Hotel/Motel/Boarding house
	CI	Industrial
	CL	Leisure
	CM	Medical
	CO	Office
	CR	Retail
	CT	Transport
L (Land)	LD	Development
	LF	Forestry
	LL	Allotment
	LO	Open Space
	LP	Park
Z (Object of Interest)	ZW	Place of Worship

Given the level of granularity of the regions, and with the aim to differentiate them in a helpful way, the primary categories of Class 1 appear too coarse. The C (Commercial) category, for instance, does not provide a sufficient level of information, whereas Class 2 categories help differentiate if regions capture primarily retail spaces, offices, industrial complexes etc. It is thus interesting to investigate the secondary categories of POI. Which ones are most prevalent across the NT and DT regions and do the distribution differ only based on region type?

Classes not listed in Table 3 (See Appendix 3) were removed for their low relevance to the analysis. For example, CZ (Information) was largely composed of signalisation (road signs, monument descriptions). The primary category named “Other” is also difficult to conceptualise. Support POIs or ancillary buildings say little about the type of activity that characterises the area. “Monuments” (ZM) recorded many statues and other small features, skewing the data to indicate large proportion of “monuments” in regions, but not adding to an understanding of the regions’ usage. Finally, parent shells were also removed as simply knowing there is a property shell does not help distinguish its type. Often, property shells are desirable to detect visits to specific POI; if a visit is recorded near an airport, the parent shell helps determine whether the visit enters the airport or not. However, parent shells could describe malls, universities, stations, and other large buildings containing varieties of POIs, and being unable to distinguish them from the P label alone, they were filtered out to prevent misinterpretation. Most importantly, residential POIs were also removed. They constitute the overwhelming majority of London

POIS (over 60%), and thus would not help distinguish space compared to more ‘functional’ POIs (such as retail, leisure, transport kept above) in the following exercise.

6.4.1.2. Linkage to H3BR_NT and H3BR_DT

The ABP points are spatially joined to the H3BR_H, H3BR_NT and H3BR_DT separately, using point-in-polygon operations, to obtain datasets listing all the POIs and their corresponding Class 2 labels per region for each H3BR version. From these POI lists per region, frequency tables are then computed to obtain the most frequent POI class for each unit. This is called the unit’s dominant ABP. Thus, we obtain the distribution of dominant ABP categories per region type across Greater London. This helps assess what type of POI is captured by the various versions of the hourly H3BR, and investigate whether changing the hour zone results in different dominant ABP summaries across London. Figure 13 maps the dominant ABP per region for all three sets with a focus on Camden. The first map snippet corresponds to the raw POI points, visualising how the region delineation around static points creates different pictures of the dominant POIs per region.

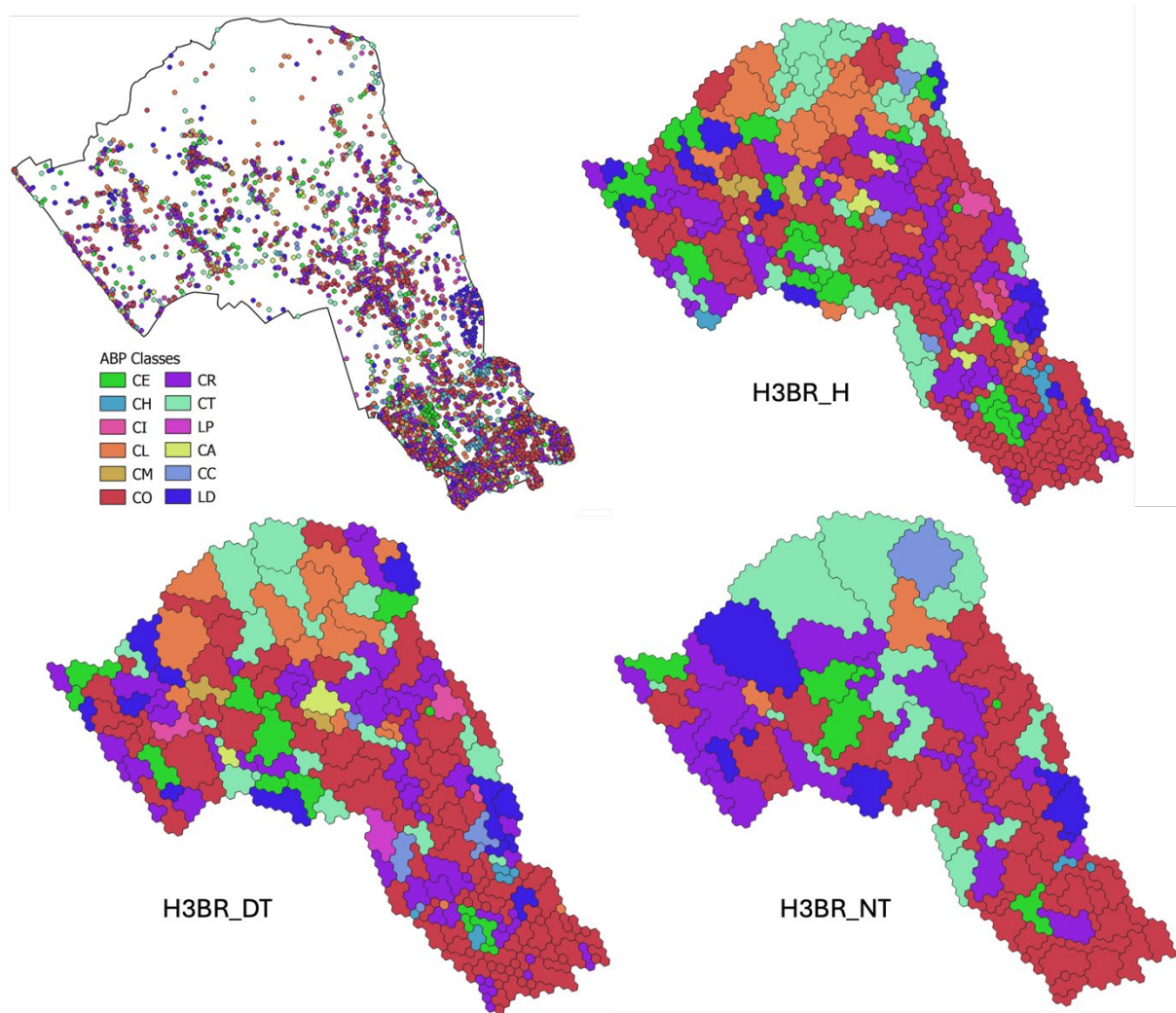


Figure 13. Maps of dominant ABP per region for each region types, focus on Camden. In order, from left to right and top to bottom: (1) POI original points coloured by ABP class (2) H3BR_H coloured by their majority (dominant) POI's class (3) same thing for H3BR_DT and (4) H3BRD_NT. Though the underlying POIs are the same as in the first map for all, the majority class varies spatially depending on the delineation chosen.

6.4.2. Results

The dominant POI per regions were then summed across Greater London and compared between sets. This is visualised in the bar chart of Figure 14. Firstly, we can note that the order of most prominent POI is different for the H3BR_NT than the two other sets. All three have retail (CR) and transport (CT) POIs as the top 2 most frequent dominant POIs. However, in third position, the nighttime regions have industry (CI) POIs instead of offices (CO) captured third by the other two region sets. This corroborates Greater London Administration (GLA) research reporting that nighttime workers in 2019 worked mostly in health sectors, professional

services and transport and storage (this last category often being located within CI POIs) (Greater London Authority, 2024).

Mainly, Figure 13 and Figure 14 show that the dominant POIs are captured in different proportions across regions. Though the POIs are static, the description of Greater London provided through the filtering of key POIs per region is dynamic and dependant on the regions used. This is an effect of MAUP, but in this specific study, it also demonstrates clearly the effects of MTUP: changing the temporal zone in turn changes the regions’ spatial scales and zones and the way the same city is summarised. Across London, the CR category is more present for nighttime regions by 4% when compared to daytime regions and by 8% when compared to the median regions (Figure 14).

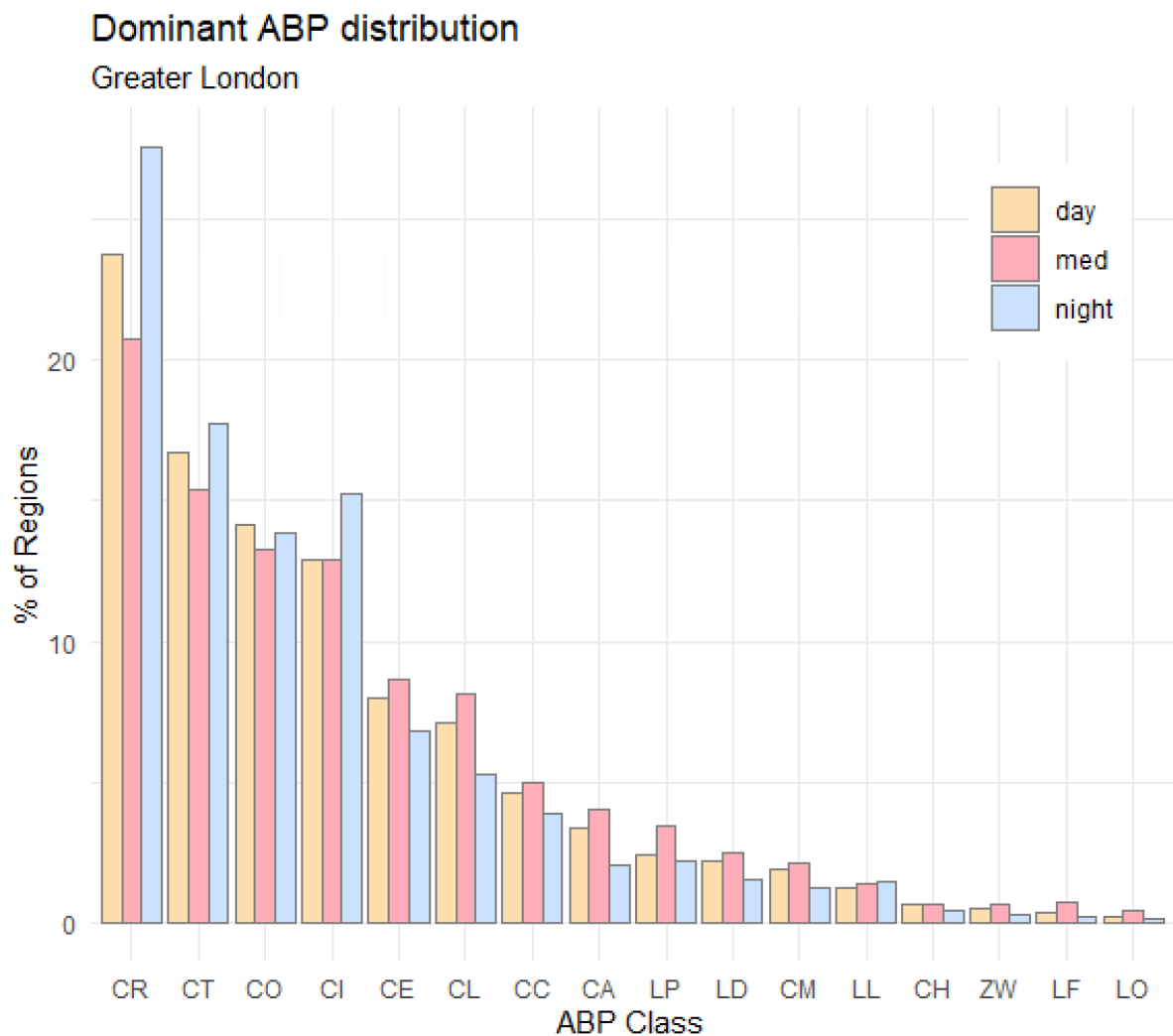


Figure 14. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H over Greater London.

Then, the same distributions were plotted at borough level, inspecting whether certain neighbourhoods of London may be disproportionately affected by this effect. The focus was put on boroughs with both day and night activity identified through Figure 10, to see whether the same relatively busy space is captured differently through temporal zone selection.

For Camden (**Error! Reference source not found.**), the proportions are relatively similar. However, POIs of industrial, agricultural, medical categories, as well as parks, are not present in the list of dominant POIs for H3BR_NT. This could be explained by the larger size of the H3BR_NT not allowing for a high level of specificity.

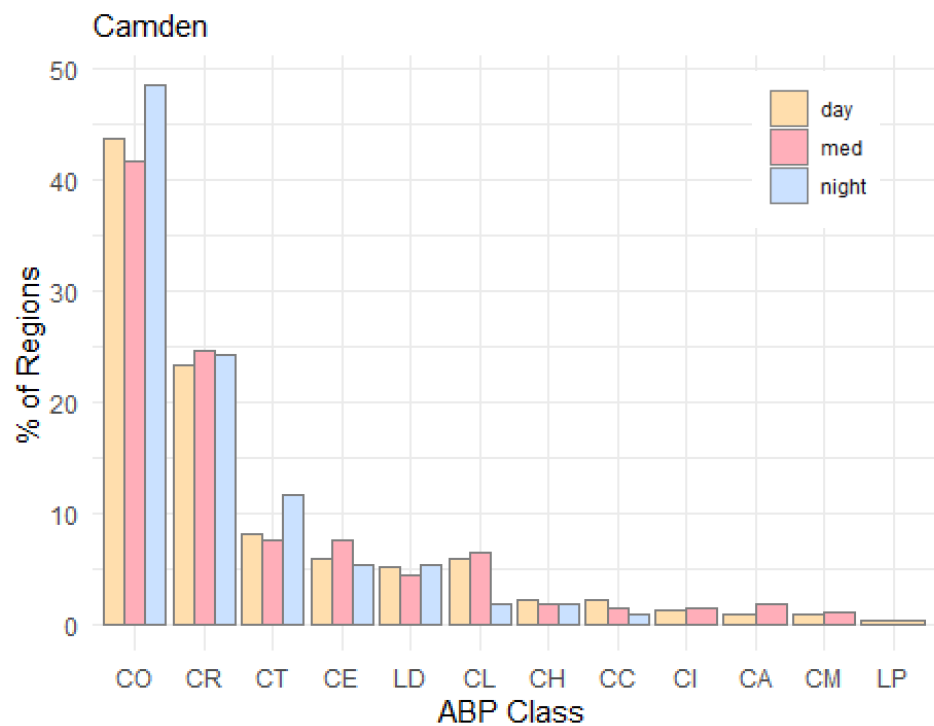


Figure 15. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H across Camden.

In Westminster (Figure 16), regions with a dominance of places of worship are only captured by the median regions. Again, the H3BR_NT sees high proportions of the dominant types, here offices and retail, which does not automatically indicate more activity in areas of retail and offices for nighttime region, but could be that the regions delineated may not be specific enough to capture detailed activity in other POI categories (see Figure 13). However, the GLA reports that Westminster experiences the most nighttime retail visits (between 6pm to midnight) when compared to any other borough (Greater London Authority, 2024). Thus, the H3BR_NT regions are expected to be comparable to the median and daytime versions for Westminster, which they appear to be with this assessment.

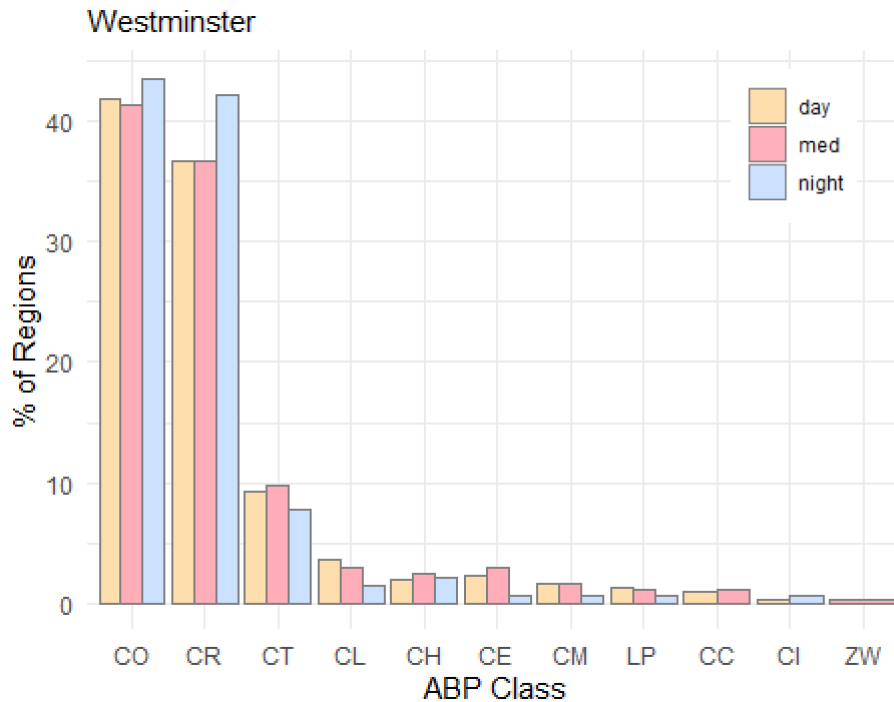


Figure 16. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H across Westminster.

Encouragingly, throughout those first examples, the daytime and median H3BRs propose very similar POI profiles of Camden and Westminster. It is encouraging that, despite a low Jaccard relationship (below 1% similarity) these two regions do not differ drastically in their POI summaries.

Differences across Tower Hamlet are more pronounced (Figure 17). The order of key dominant ABP is different between the regions. H3BR_NT are led by CI POIs, whereas this category only come at the 4th position for H3BR_DT and H3BR_H. There is a marked 10% difference in CI POI proportion between H3BR_NT and H3BR_DT. Nighttime regions display less regions with dominant CR classes, and CT dominated regions only come second after industry. Using one set of regions rather than another to summarise Tower Hamlets provides very different profiles of the borough. For the daytime and median regions, Tower Hamlets is a retail, office and transport led borough. For the nighttime region, it is dominated by industry and offices.

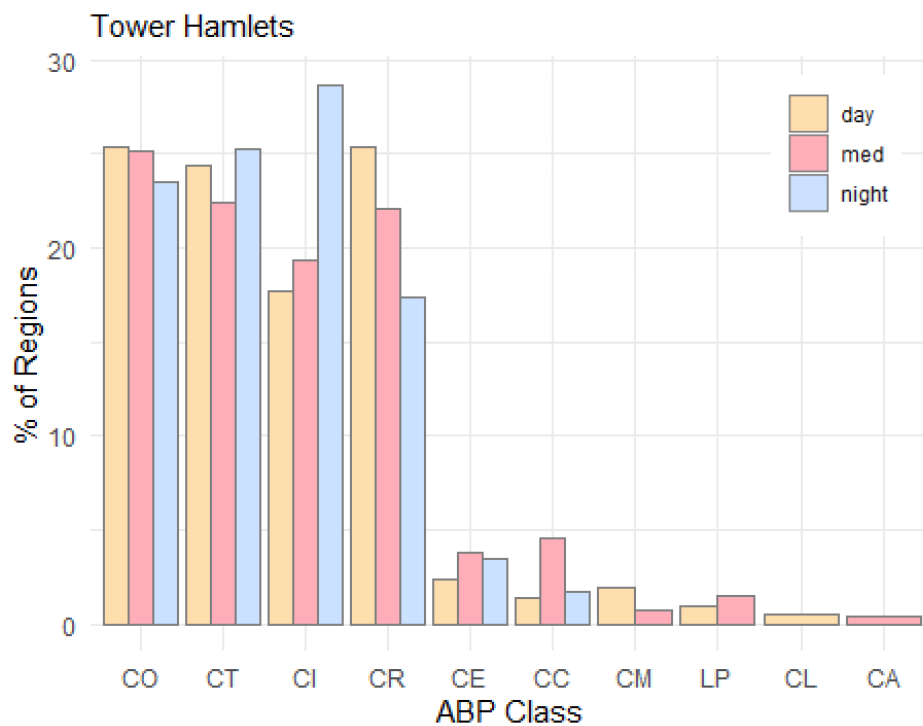


Figure 17. Distribution of dominant ABP classes per region. Comparison of H3BR_NT, H3BR_DT and H3BR_H across Tower Hamlets.

6.4.3. Implications

The results presented above show that there are differences in the way POIs are summarised across space depending on the time frames used for aggregation. The choice of temporal units impacts the profiles of London, Camden, Westminster and Tower Hamlets obtained. The changing mixes of POIs in these areas alert to scale and location issues. This finding echoes earlier findings from Chapters 3 and 4, which showed that, for different aggregates, LOAC and WPZ profiles were notably different. However, here, the POIs do not describe the data itself, but the underlying, fixed space. There is thus a dynamic reshaping of otherwise static elements through the choice of temporal units. Aggregation that follows mobile points captures space in varying ways, which had not been showed explicitly before in this work.

The results above show a closer relationship between the POI profiles of H3BR_DT and H3BR_H, pressing that the times of low data captured by H3BR_NT are disproportionately impacted by the MTUP and MAUP effects. In a sense, using a more general set of regions (H3BR_H) would have less impact for daytime analysis with higher counts, than it does for nighttime, low activity-based analysis. In many ways, the findings through this chapter repeatedly highlight the intertwined natures of MAUP and MTUP. Much like with MAUP,

there is also no “one size fit all” solution to prevent MTUP effects in the aggregation process. The key takeaways from these results are discussed further below.

6.5. Summary and discussion

This chapter highlighted the importance of considering time in regionalisation and aggregation processes. It started off by reminding us of the H3BR’s good performance at the day scale before demonstrating its advantages were progressively lost as we strayed further away from the original H3BR underlying temporal scale. Following this finding, it was showed that the H3BR making methodology was applicable to hourly temporal scales, and multiple new region sets, based on varying temporal scales and zones, were created and tested. The tests built on what was developed throughout the thesis (Jaccard index, omission counts and analytical comparisons) to try to quantify and formulate the MTUP’s effect on the overall H3BR making process. In this context, the H3BR were harnesses as a tool to help assess what incremental changes to temporal input can change in outcome regions. This was a great opportunity, both to inform on the realistic capabilities of the method, but also to insist on the MTUP’s importance. The key conclusion showed that the impacts of these incremental changes were significant, and many factors can and should inform decision making.

Some important limitations to the studies above can be noted. Firstly, there were no attempts to change the atomic unit of H310 to another H3 scale which could have been a better fit to hourly counts. This alteration would have increased the complexity of comparing H3BR_D with H3BR_H, and the impact of MTUP on the method would have been harder to assess, as the regions would have been different through other factors than temporal alone. MSOAs were assessed to ensure they could still be reasonably used, and from there onwards, the aim was to change the methodology as little as possible, so the outcome differences between versions of the H3BR would only be due to temporal scale and zone differences. Another limitation is that these analyses do not investigate the region’s appropriateness in picking up events rather than overall patterns. What mix of POIs would capture unique events rather than repeatable space? It would have been valuable to explore this more thoroughly through the final section of this chapter concerned with spatial compositions.

Regardless, the findings help draw some recommendations for H3BR in light of MTUP effects. First, using a summary region (average of a period of time, such as the H3BR_H and H3BR_D)

was found to provide advantages compared to traditional methods. These summary regions allow for comparison across different times, and promote general applications at their temporal scale. However, the closest fit to the data, especially for areas and times of low activity, might not be these summary regions. In many contexts, nothing can replace having a research direction or informed hypothesis. If research is conducted specifically on times of low data, such as nighttime, the thematic regions (H3BR_NT or DT) can be a better fit for the analysis. They, however, won't be as generalisable as the summary ones. This is especially true for times of low data, as it was shown through the analysis in this chapter that these are the times that gain the most from bespoke regionalisation and are the most affected by unfit regions.

Using data at specific hourly scales also relies on the data being good enough at said scales. 'Big data' does not equate to the data being appropriate and plenty at all temporal and spatial scales. Choosing the scale requires good field knowledge, a good notion of the applications, and an understanding of the dataset. In fact, this dataset has low hourly counts, especially at night. This could be driven by underlying factors such as app types, but also simply by the type of activity captured being lower at night. If a research aim was to conduct an analysis of night activity around specific, granular areas, perhaps this dataset might not be the most suitable, as it is biased towards daytime activity and is best aggregated to small areas by day rather than by hour. Reducing the impact of MTUP and MAUP also relies on a strong understanding of the dataset dynamics. We can confidently recommend its use for tracking daily activities. Hours-wise, it captures interesting dynamics across days, but for very precise hourly comparisons it would be helpful to gain more confidence that an hour's activity in a small area is due to specific events rather than apps-based dynamics.

The final analysis stressed the importance of seeing MAUP and MTUP as a combined set of problems rather than separate effects. The kernel of time geography remains this intrinsic relationship. Accounting for both in reasonable ways ensures conclusions reached through analyses are due to the data and behaviour studies rather than a tampering of spatial and temporal scales.

Regardless, this chapter helps close the current H3BR experiments on a relatively high note. The methodology proved to be rapidly and easily reusable to the creation of multiple sets of units, which enabled this exploration of MTUP. Thus, they can also become a research tool as well as a flexible aggregation unit. This is encouraging as the long journey towards temporal tessellation starts with more incremental and quantitative explorations of the MTUP effect.

7. DISCUSSION

7.1. Introduction

Traditionally, in-app mobile phone datasets and other sensitive consumer location datasets have been mostly aggregated to arbitrary grids or administrative boundaries. When the data is accessible in its raw form, researchers have been able to utilise it to conduct valuable analysis in epidemiology, transport planning and countless other fields. However, preliminary tests proposed in this work demonstrated that aggregating the datasets to protect privacy significantly alters the data, often impacting the results of analysis conducted with them. This proves to be particularly troublesome as the access to raw datasets is still largely limited, with the more frequent use cases relying on aggregates. Consequently, the aim of this thesis has been to spark conversation about the way these datasets are accessed and disseminated, and propose a new tractable method for regionalising and aggregating sensitive location dataset. This work benefitted from a privileged access to an in-app dataset, a novel source of location data with established potential for geodemographic and mobility analysis. The efforts were directed towards generating regions which not only could preserve the underlying data well and protect privacy, but would also provide new granular insights on the spaces defined based on the times at which they are accessed and the type of activity they may capture.

The presentation of the in-app data provided in Chapter 3 identified issues with the dataset related mostly to coverage, uncertainty and inconsistencies in sampling. It also brought forth the necessity to produce aggregated products, to even be able to provide and present the data's initial descriptions. As a result, Chapter 4 sought to investigate more deeply the outstanding challenges posed by the MAUP during data aggregation. The regionalisation principles, along with the preliminary tests conducted in Chapter 4, provided a firm foundation enabling the key findings of Chapter 5. Learning from other methodologies, and assessing their strengths and drawbacks helped justify and inform the development of the proposed H3BR method. The analyses throughout provided evidence that bespoke data-driven regionalisation facilitates aggregation and helps propose better quality aggregates that meet disclosure control requirements and protect individual and commercial information. Chapter 6 followed the thematic thread woven throughout this work: the value of applying these new forms of data at granular temporal and spatial scales for mobility analysis. It thus proposed an initial assessment

of the effect of MTUP on the H3BR methodology, aiming to highlight the key impacts of changing temporal scales and zones, to better inform future regionalisation efforts and applications of the method developed here. Together, these findings suggest that alternative geographic units derived from consumer data can be made which capture granular insights on data users and spaces. These kinds of outcomes are now possible due to the advent of new data sources and new GIS tools enabling the development of these data-driven regionalisation methodologies.

This final discussion chapter aims to consolidate this thesis' contributions by first reflecting on the methods used throughout. The key limitations of the analyses presented are also described in more depth. Then, the applications of this work are explored primarily in the contexts of academic research, with acknowledgments of the implications for data providers and users alike. Future prospects for this study are discussed, exploring the research avenues which could potentially acknowledge some of the limitations and shortcomings discussed. The chapter concludes on the key takeaways from this thesis.

7.2. Reflection on Methods

Research conducted on in-app data is often exploratory, due to the novelty and recent uptake of this specific data type. On the other hand, regionalisation methodologies are more established, but often remain proprietary or project specific. In developing this thesis, it thus became necessary to devise new methods to suit the specific research objectives of regionalising in-app data, but also produce more generalisable regionalisation methodologies. In the approach to assessing the impact of MAUP on this dataset, new criteria for what constitutes data loss needed to be defined. However, the most prominent example where new heuristics were required for this work was the development of the unique regionalisation methodology. This method relied on several context-based decisions and pragmatic steps which are discussed here in more detail. The final assessment of MTUP effects in Chapter 6 hinged on various definitions and tests, which should also be reflected upon. Given that other methodological approaches were more established, such as the point-in-polygon techniques and the app accuracy assessments, this section does not extensively discuss them. This discussion will thus focus on three main areas: the initial data assessment with a special emphasis on the MAUP tests, the regionalisation methodology, and finally, the MTUP analysis, outlining the rationale and impact of each. Each stage required a mix of subjective, field-informed and quantitative decision making, and where

possible, tests were implemented to justify the choices as objectively as possible in the context of the desired outputs.

7.2.1. Data assessment methods

The dataset assessment (Chapter 3), comprising of metadata and data description, was largely exploratory. Other studies have described with much greater detail what composes in-app data, and what sets them apart from other large human-generated location datasets (See Calabrese *et al.*, 2013; Almuhimedi *et al.*, 2015; Kishore *et al.*, 2020; Gibbs, Eggo and Cheshire, 2024 for some examples). However, the descriptions provided in Chapter 3, pertaining to spatial and temporal coverage, app counts, device and impression distributions, as well as accuracy levels throughout the dataset, amply served the purpose of contextualising the subsequent analysis. The coverage analysis helped decide which samples of the data were the most appropriate for future tests, settling largely on Greater London and the years 2018-2019. Low-accuracy locations were filtered out following these assessments, and the data description provided a strong basis to scope the type of insights one could expect from the in-app data, when contextualising its different attributes (such as the implication of certain apps and devices producing more data). Mostly, this initial data assessment served to build an understanding of what was at stake with the subsequent MAUP analyses. What was implied by “data alteration” in this context, would be loss of valuable and interesting information, potentially uniquely provided by these types of data.

Focusing on the MAUP assessment, quantifying the data loss through changes of scales and zones required the invention of a stable metric to compare across aggregates. This was done through the creation of the activity count: the count of unique devices per day per area tested. This is a new heuristic, as there is no general consensus on how this should be accounted for in the literature, each decision remaining project specific. It should be noted that other metrics were presented, such as total impressions or modal locations per area. These were not tested against one another beyond exploratory assessments, which would have been a more inductive approach. However, the choice of the activity count was justified in the context of the data’s nature and potential in providing mobility insights. As explained throughout the work, capturing individuals as they travel through various regions provides interesting information closer to capturing the original data than modal locations, and removes noise from otherwise counting all impressions. It is, however, important to acknowledge that the impact of double counting unique devices in that matter is an outstanding challenge. Claiming an output is more

granular, when it is perhaps simply counting the same device more times, could be inaccurate in multiple contexts. Differentiating granularity from multiplicity, and proposing a consistent and transferable way to quantify granularity requires further heuristics. Overall, the activity count produced the level of insight needed to assess the various regionalisation methods tested. It allowed for a quantitative comparison of population counts obtained through the standard aggregate against a control in the context of Chapter 3, and was applied in Chapter 4 to identify which aggregate scale best fit the original data.

The initial assessment of MAUP (Chapter 4) was conducted by varying both scale and zone. This helped describe both key MAUP effects and show the importance of creating bespoke regions on top of aggregating at larger scales. However, this assessment could have been conducted with more precision by introducing the scale and zone changes more incrementally and reporting exact measures for each, as well as the WPZ group comparison displayed. This was improved upon in Chapter 5, in two ways: (1) when H3 resolutions are assessed to settle on the H310 size for the building block selection and (2) when comparing the regions quantitatively to OAs and OSGBs through the tracking of omitted counts rather than population counts.

7.2.2. Reflections on regionalisation methodologies

Developing the regionalisation methodology was the core output of this thesis. The literature review's focus on aggregation provided the backbone for much of the initial decision making. The core principles proposed by Casado Díaz and Coombes (2011) for the making of functional regions were integrated to help provide a more formal understanding of what constitutes an acceptable output. Considering the 'building block' methodology, famously used in OA delineation, among others, helped conceptualise and develop the core function of the H3BR methodology (Cockings *et al.*, 2013). However, as the first in-app data specific regionalisation process, many considerations were not directly answered via traditional methods.

A key methodological consideration was to create static regions. This constitutes a drawback, especially in the context of mobility analysis. Not having a notion of origin-destination and flows significantly reduces the scope of analysis conducted with the aggregate. However, the aim was to create outputs more akin to WPZ, which may not be renewed past the 2011 census. Travel-to-work areas could also be devised, potentially by readapting the proposed method. This represents a commonly recognised issues when aggregating granular point datasets to

areas, which was not addressed as part of this work's methodologies. This key decision sought to simplify the method with the aim of making it more generalisable. It is hoped that, nevertheless, the static snapshots provided still constitute clear improvements over the industry standard aggregates to arbitrary grids. That being said, future research, with more specific and less global focus, could explore the generalisation of the regionalisation method to specific flows and origin-destination behaviours not captured here.

The criteria selected for the regionalisation methodology were in part informed by literature and in part justified by the aims outlined. Homogeneity is a key criterion of most rule-based functional regions (Martin, 2000; Cockings *et al.*, 2011). However, this is limited in our methodology, as the dataset did not contain geodemographic information to quantify that homogeneity. The choice of land use and terrain became a practical alternative, as it also promoted the creation of regions considerate of underlying land attributes. The decision to nest the regions within administrative boundaries was also justified by the intended use of these types of data aggregates. By allowing direct linkage with MSOAs, a nested geography has more transfer potential for analysis, and reduces the necessity for another spatial join and change of spatial units (which introduce more uncertainty and MAUP impacts) to compare with other dataset or attribute geodemographic characteristics. Consistency between geographies is valued in geographical analysis (Walford and Hayles, 2012). There is additional value in proposing nested geographies as they are more easily interpretable and understood by users outside specific research communities, and reflect the shared advantages of a recognisable region outside of solely being a statistical summary. It is important to note that nesting this methodology inside other geographies may have given the regions another dimension which may be harder to predict and quantify, as these overarching geographies carry their own meaning and criteria.

The ranking of criteria was also justified by the aims. The first criterion for merging was to meet the threshold number of devices, as the key motivation of the regionalisation was disclosure control. Terrain came second, and the whole process was nested within administrative boundaries. No assessment was conducted to investigate how removing the nesting would impact the region shapes, or if data-only regions would capture the same thing as those made with the inclusion of the land-use attribute. The method proposed, however, provides the same result every time, as long as the criteria and thresholds set are the same (no seed, the method does not require a specific starting point, etc.). Though the changes in outputs

resulting from the removal of certain criteria were not investigated, a quantitative sensitive analysis was conducted using the established Jaccard similarity index to assess the method's volatility.

Another key methodological decision in regionalising the data was the choice of hexagons for the initial tessellation. As explained in Chapter 2, they provide many advantages, particularly for large scale computations and uniform comparisons across neighbours. It should be noted that they are not ideal for all applications. In the case where the data is to be compared to physical geography databases, commonly provided in rasters or square grids, a square-based methodology would provide a more appropriate and direct relationship between the two datasets. However, for nesting and comparison with administrative boundaries, hexagons were the preferred choice. They fit regional outlines more closely, particularly at the high resolution of H310 (De Sousa *et al.*, 2006).

Many other choices were motivated by the desire to propose a clear, reusable and tractable methodology. The simplicity of the method, which could be one of its main drawbacks, is here privileged for its adaptability and versatility. It is hoped that this thesis helped consider regionalisation with a less project-specific approach than usual. Great regionalisation examples cited throughout this work sought to output specific datasets, such as the OA. However, the frequency at which new consumer datasets appear called for generalisable methodologies for sensitive point datasets. Often, they carry similar privacy risks, and could benefit from an approachable, widely applicable methodology to conduct initial regionalisation tests. The methodologies and considerations provided in this work, though currently specific to in-app data, can still be primordial in the development of such new tools.

The regionalisation validation was conducted mostly through the assessment of count preservation. This helped assess the performance of the regions and clearly rank them amongst other methods. However, to further validate the proposed regionalisation method, additional statistical approaches could be applied. These include testing spatial autocorrelation to measure internal consistency, and evaluating how much information retention occurs in the new regions, such as with entropy measures. Bootstrapping could be used to validate the robustness of regional boundaries under resampling, strengthening confidence in the methodology's reliability.

This focus on omissions (and remaining counts) to compare the performance of aggregation methods is a core element throughout the thesis. While focusing on omission as a performance indicator in the thesis is valuable for ensuring data retention and minimising privacy risks, it is important to recognize that omission alone may not fully capture the effectiveness or usability of the regionalisation method, especially in different application contexts. The significance of omission depends largely on the intended use of the dataset. For example, in mobility research, minimising data omission is critical for maintaining detailed movement patterns. However, in other contexts such as policy-making or urban planning, factors like the representativeness or predictive power of the regional units might be more important than simply reducing omitted data points. The choice of minimising data omission throughout the thesis was motivated by its focus on mobility analysis, but this means that findings must be recontextualised within this exercise and thematic emphasis.

To consider other contexts, a more comprehensive validation framework could be considered, incorporating additional performance criteria beyond omission. These might include: (1) data retention and information loss: assessing how well the regionalisation preserves key patterns or relationships within the data. Entropy measures listed above could quantify the extent to which the original data's variability is retained after aggregation. (2) Predictive accuracy: evaluating how well the regionalised units can predict relevant real-world outcomes (e.g. mobility patterns, socioeconomic trends, or transport flows) would provide a practical measure of their utility. (3) Clustering consistency: assessing the internal consistency of the regions, in terms of spatial homogeneity or spatial autocorrelation, would also help determine whether the regions reflect natural spatial patterns. By incorporating these additional criteria, the evaluation framework could offer a more balanced view of the performance. This would have helped ensure that the regionalisation process is better aligned with the varying needs of different applications, but was outside of the scope of this specific thesis, which focused on mobility analysis applications to drive necessary choices in refining the methodology.

It finally should be noted that the assessment of other data-driven methodologies (quadtree and Voronoi) relied heavily on the aforementioned qualitative assessments built on the Casado Díaz and Coombes (2011) principles. Voronoi, quadtree and H3BRs are differently generated, some of them requiring prior clustering or aggregation to building blocks. Their incomparable shapes and resulting outputs are complicated to assess quantitatively against one another. Thus, the final claim that H3BRs are more appropriate is heavily reliant on the initial criteria set out to

frame this project. H3BRs proved especially performant in creating regions which were recognisable as such and interpretable in the context of other geographies, underlying terrain and mobility studies. As there was no statistical assessment between the three proposed data-driven methods, this project does not necessarily claim that the final methodology (H3BR) outperforms the others quantitatively in every context.

7.2.3. Reflection on MTUP assessment

The final area of analysis which required methodological considerations was the assessment of MTUP effects on the regionalisation methodology. As touched upon in the Chapter 6 summary, this assessment could be improved upon in a number of ways. It was primarily investigational, and though it provided considerations and recommendations to address the issues identified, it did not provide technical solutions to the MTUP. It applied the heuristics developed throughout other chapters, and most of the methodological considerations listed above for the making of the regionalisation algorithm carried over into the MTUP analysis.

Nonetheless, one key area where new methods were implemented was to define and quantify what was understood by temporal scales and zones. This was built from MTUP literature and echoed MAUP terminology used throughout previous chapters, to homogenise the conversation, help in quantifying the effects, and relate findings to previous assessments of MAUP. This description of temporal scale changes on H3BR relevance focused mostly on reducing temporal scales rather than increasing them, as this work aimed to promote granularity. Zones could have been explored through dimensions other than hours, but multiple angles of the zoning question were explored throughout the analysis of ‘thematic regionalisation’, which was also defined for this analysis. Finally, the MTUP assessment attempted to go one step further than simply tallying omission proportions by assessing the distribution of POIs across different sets of regions. It should be noted that this was done through the comparison of ‘majority’ or dominant POI types. Perhaps a more complete and nuanced analysis could have explored mixes of POIs, and try to address whether the regions are helpful in capturing singular and notable events rather than simply generalising space. Nevertheless, it was necessary to attempt to summarise spatial interpretations for comparative analysis, to grasp the broader impacts of MTUP on the H3BR method and inform its realistic use.

Regretfully, the complexity of quantifying the relationship of MAUP and MTUP in a more explicitly paired analysis was proven to be outside the scope of this work. However, this

assessment shed light on their intertwined natures to motivate analysts to consider them in equal measure, and perhaps helped drive future spatial and temporal tessellation attempts.

7.3. Limitations

This thesis provides numerous insights and outputs, despite some methodological considerations evoked above. The work also has limitations, primarily stemming from the inherent uncertainty of the data and the ways sensitive datasets are restricted. These limitations are acknowledged and discussed below

7.3.1. Dataset limitations

Issues of data quality introduced in Chapter 3 brought about concerns of uncertainty at various stages of analysis, which could not always be quantified and accounted for. One example is the unequal spatial and temporal data distribution. Attempts were made to reduce this concern by focusing on Greater London, the area with the most consistent coverage throughout the data period. Trying to acknowledge the disparate coverage brought about a key limitation of the work, the exclusion of rural areas. Only conducting this assessment at the city scale means there is little understanding of how the regionalisation method would perform nationally, especially in areas of lowest data. Count preservation strategies may be most relevant in areas of low data, which could not be assessed here. However, the inconsistent coverage would have weakened the conclusions and reduced the likelihood of creating a stable, comparable output. Though the dataset provided was high in volume, not all points were retained, and the behaviours captured might represent very specific, mostly urban, populations.

This brings us to issues of representativeness being prominent with the in-app data provided. A concern with consumer datasets is that their fewest most active users generate a majority of the points (Lansley and Cheshire, 2018). Attempts were made to try and mitigate this, by removing large outliers from the data, or counting unique devices, thus filtering the noise overactive devices may produce. However, with these new forms of data, some uncertainty could remain undetected and unaccounted for. For instance, it was assumed that the large variations in device proportion throughout 2018 were explained by changes in operating software policies or apps, but this remains speculative insofar as we were not provided empirical evidence to make causation out of correlation.

The data bias analysis of Chapter 3 attempted to identify which populations are best captured by the dataset. Certain populations across Greater London were in fact under or overrepresented by the data when compared to census information. Though this analysis was imperfect due to the linkage necessary for the assessment, and the filtering of nighttime points to reflect census data, it pointed out uncertainties in how closely populations are captured by these datasets. Additional research could be conducted to provide a more thorough assessment of population representation. Recent studies using the same in-app dataset described here have developed home-location assignment heuristics to more closely describe the underlying populations driving these data impressions (Gibbs et al., 2024).

Finally, notable uncertainty stemmed from the proprietary aspect of consumer data. For example, there were no details provided on the apps, likely for commercial data protection reasons. This meant that, despite benefiting from very privileged access to raw point data, many questions pertaining to the dataset collection processes remained unanswered. Therefore, this work could self-dom account from uncertainties arising at these stages of the data making process, or filter out specific apps to capture specific behaviours. The full list of variables provided in the dataset were also not communicated here to prevent commercial disclosure, which could be seen as a limitation to this thesis's aims of transparency and reproducibility. Consequently, apart from the data representativeness assignments laid out in Chapters 3 and 4, this work had little insight as to what type of transaction or behaviour this in-app datasets may capture. This was less of an issue for developing the regionalisation strategy, especially as we aimed for a generalisable method. However, studies seeking to harness in-app mobile phone data to produce geodemographic or mobility insights would require a more in depth understanding of the underlying behaviours and collection methods.

7.3.2. Secure environments and software access

Due to the sensitive nature of the dataset, all data processing was constrained to the DSH, including the development of the regionalisation methodology. This carried additional limitations. A lot of time was required to reach a balance between the data requirements (e.g. size, format etc.) and the DSH permissions. It is acknowledged that access to remote secure environments greatly improved the feasibility of this work, in comparison to the traditional physical labs, by reducing the physical barriers required to access the dataset. However, it can

be reiterated that the skills necessary to navigate technical hurdles inside remote servers required substantial training and resources.

More importantly, some tools were unavailable due to DSH restrictions. An aim of this work was to test the AZTool automated zone design software (Martin, 2003). This was not feasible as it could not be onboarded onto the DSH within an appropriate time frame for this project. This further motivated the choice to develop a methodology in R, or other programming languages, as, in the DSH experience, these are more directly adoptable. As most sensitive dataset would be subjected to similar restrictions, providing R functions increases the likelihood that the method is authorised across platforms. In light of these limitations, a number of interesting automated zone design methodologies could not be properly compared using the sensitive in-app data.

In summary, this thesis underscores key considerations for using such data to capture specific patterns, noting that the analysis is contingent on the data available, which may restrict the range of insights that can be obtained. Despite this, it is hoped the data quality was appropriately assessed to meet the main aim of addressing issues pertaining to regionalisation and aggregation.

7.4. Applications and Implications

Despite the recognized limitations above, this thesis offered key insights for the aggregation, dissemination and integration of in-app mobile phone datasets in and beyond geographic research. The applications and implications of this work can be broadly categorised into five parts: (1) enabling real world applications of promising mobile phone data through their safe dissemination (2) contributions towards an understanding of the utilised in-app data, (3) provision of a formal framework for addressing MAUP and assess data regionalisation and aggregation, (4) implications derived from the MTUP assessment (5) contributions toward data protection, privacy and ethical discussions.

7.4.1. Harnessing promising datasets

As listed in the literature review (Chapter 2, Section 2.2.2.2), mobile phone data has a wide range of applications. The potential of these datasets grows further as they are disseminated with more stakeholders and applied in new areas to provide real-time insights in a safe manner. The regionalisation method proposed is essential for disseminating sensitive data in ways that

maintains individual privacy but still provides actionable insights. In doing so, the work directly addresses one of the most critical barriers to the use of mobile phone data in applied contexts: how to balance utility with confidentiality. This newfound balance opens up important opportunities for use in public health, disaster management, and humanitarian response, amongst many others. For example, during the COVID-19 pandemic, mobile phone data played a key role in many countries, tracking movement patterns and investigating the efficiency of lockdown strategies (Jeffrey *et al.*, 2020; Kang *et al.*, 2020). In conflict zones or areas experiencing natural disasters, humanitarian actors could use similarly regionalised and aggregated datasets to monitor displacement, aiding the protection of civilians, as activity counts at granular scales provide valuable insights when timely information is vital. The spatial sensitivity preserved through regionalisation enhances the ability to respond dynamically to unfolding crises using novel and promising datasets.

Beyond these example applications, the methodology developed is intentionally flexible and can be adapted to other forms of dynamic spatial data (such as transport flows, social media activity, or other point data) making the approach transferable. This broad applicability increases the relevance of the work for a wide range of actors, including governments, NGOs, urban planners, and public health agencies. Ultimately, the uptake and impact of this work will depend on institutional efforts to safely integrate data-driven information into research, policy and practice.

7.4.2. Understanding in-app data

Regarding the usability of the specific in-app data used throughout this thesis, it is evident that, when considered alone, this dataset would be insufficient to accurately represent the general public. However, the analyses provided illustrate ways in which this data could offer relevant insights into its *target* populations. Due to their recent emergence, few public studies provide in-app data descriptions inclusive of information such as app distributions and other key metadata. This work, through providing such descriptions, contributes to building a knowledge base of what is captured by these datasets. The high granularity of the data can be harnessed with a strong understanding of its representativeness, which could enrich representations of space and populations outside of traditional census-based measures. Providing this information in an accessible format could offer academics a reliable reference point, and allow further investigation into the biases and representation issues mentioned previously.

The specific applications of the dataset could take many forms, and should be considered independently, on a case-by-case basis. This decision making can be enabled by an access to the dataset provided by this study. The aggregated products proposed allow researchers outside of the CDRC to inspect the dataset and assess its usability in the context of their own research. The H3BRs' linkage potential helps couple these datasets to other sources of data, potentially facilitating these analyses. By making the dataset both easily accessible and generalisable, this thesis hopes to promote their use and inspection by the wider research community, which in turns will help build a rigorous, shared and transparent understanding of these novel sources of data.

7.4.3. The MAUP, formal frameworks for aggregation, and dissemination

Perhaps the most notable application of this thesis relates to MAUP and regionalisation. There has been a significant gap in the assessment of these issues, particularly with regards to new forms of data and spatial aggregation. This thesis provided ample illustrations of the impact of MAUP when aggregating in-app data. The basis provided encourages future assessments, as new forms of data continue to arise and permeate geographic and social science research.

For the regionalisation methodology proposed, a formal framework was designed, building on past literature and adding elements centred around reproducibility and transparency. A formal framework generally refers to a structured set of rules, principles, or guidelines that are explicitly defined and rigorously applied to a specific process, with explicit rules and definitions, including generalisability and scalability. The assessment principles proposed could thus be considered as a formal framework for benchmarking regional outputs for use in mobility and geographical analysis. It is hoped that this can inform future regionalisation methodologies for aggregation and consolidate decision-making.

The detailed methodological descriptions of the H3BR algorithm were also provided to promote future applications. Point datasets could now more readily be aggregated to bespoke regions rather than arbitrary grids where a project-specific regionalisation is unnecessary or inaccessible. Standard practice aggregates could thus be improved upon, without requiring in-depth knowledge of the datasets or investing human and financial resources into creating dataset-specific regions each time. If the regions remain general, as described in this work (static, capturing only a general time frame), they still improve upon arbitrary aggregates for dissemination.

7.4.4. Applications of the MTUP assessment

Insights from the MTUP assessment revealed several implications for optimised region profiles that are both spatially and temporally pertinent for analysis. This analysis showed how incrementally changing the temporal input can help pinpoint the scales and zones at which a certain method is most useful. This information both guides the potential applications of the dataset used to create the regions, and provides additional dimension for those seeking to use the output regions themselves to conduct analysis. The evidence regarding the differences between the compared temporal regions may also help grasp the temporality and behaviour of users within the data.

Three key applications can thus be retained from this temporal assessment. First, the regions are better understood as a result. Future users are aware of their limitations and the boundaries imposed by the underlying dataset, specifically in terms of temporal scale. Clear recommendations were provided for selecting temporal scales when using the H3BR in general, which provides overall guidance for their best practice use.

In turn, questioning the aggregate's quality for times of low data provides further insights into the dataset, confirming that it captures daily activities over the daytime the best, especially if aggregated. Big data does not equate to perfect data across all spatial and temporal scales, and these analyses reminded future users of the importance of considering data capacities, as well as spatial and temporal dimensions, when choosing aggregate zones.

Finally, the assessment demonstrated the versatility of the H3BR methodology. Automated zone designs beyond the H3BR can be used as exploration tools where the spatial and temporal dimensions can be tested and better quantified. Generalisable methodologies thus have this added benefit of being adaptable to multiple test scenarios for conducting tests.

7.4.5. Data protection, privacy and ethical implications

A more theoretical perspective on this thesis' implications includes a discussion of how the work proposed throughout contributes to debates around data ethics. The main emphasis was put on preserving data privacy through disclosure control, largely by removing counts below 10 devices. This was successfully achieved. However, the creation of spatial aggregates of these data types carries ethical implications beyond simple disclosure control. Producing ethical

outputs using in-app data does not equate solely to producing non-disclosive, law-abiding outputs. Considerations of context and thinking through potential scenarios of use is paramount. This is in part because data could still potentially be misused despite being non-disclosive *per se* (Ruppert *et al.*, 2017; Williams, 2020; Sieg *et al.*, 2024). Additionally, transparency and reproducibility are core ethical principles which are sought to be enabled through the proposed research.

‘Where’ data is collected plays into the ethical challenges that may arise when it is disseminated. Data travels but is always encountered within significant local settings. The invisible aspects of data (origin, locality of collectors, initial cultural context etc.), must be thought of critically and always in relation to other localities rather than as general entities (Loukissas, 2019). This geographical recontextualization is particularly difficult with datasets such as in-app data, as the sources are heterogeneous, with various and often unknown local attachments. In line with this, another ethical concern in treating these datasets is their dehumanisation, partly leading us to believe that the data cannot be dangerous because it is impersonal (which has been demonstrated to be untrue), and partly masking the often-complex social dynamics of the individuals who make up the data. As explained by Ruppert *et al.* (2017) “to speak about data as though it records subjects and objects independent from the social and political struggles that govern them is to mask such struggles” (p.1). Grasping the social and political context is not simply a matter of disclosure control, but also a matter of understanding the acceptable dynamics to represent, and how to visualise underlying struggles and power dynamics represented by the data and its use (Ruppert *et al.*, 2017). The work presented here sought to define the dataset as best as possible prior to aggregation in the hopes of recontextualising the data within its local setting, rather than only presenting it as disjointed point data. The literature review (Chapter 2) sought to place ethical concerns pertaining to the data’s nature at the centre of the conversation. It is hoped that improving aggregate data quality can help address many of the issues relating to misuse and misinterpretation of these inherently socio-political datasets.

Additional ethical challenges have also arisen from the rapid appearance of new actors (third party data holders) and technologies which are hard to foresee. The development of technical solutions to strengthen disclosure control proposed in this thesis offers promising solutions to this. These allow for analysis to be obtained without the researchers having access to any individual’s data, thus preventing issues stemming from data sharing or mistakes in aggregation processes (Jain *et al.*, 2016; Bampoulidis *et al.*, 2020). These developments promise a wider

access to data by researchers, reducing imbalances in power-dynamics. This shift from expert to non-expert driven data access and spatial data science helps democratise tools, data and information access, and empowers users. Non-experts can thus both advocate for better use of their data and utilise them themselves. Defining privacy helps shift the debate focus from what privacy *is* to the outcomes of societal struggles over privacy (Ash, 2016). Thus, this work promotes the safe dissemination of these new data types in these ever-evolving contexts and proposes technologies which could potentially increase individual's control and understanding over their data. Safe dissemination of location data and understanding of its place in the physical world can open up the ethical debates to help define and legislate "acceptable use" and "public benefit" of sensitive data, and enable wider discussions and explorations of other ethical debates centred around data ownership and usage.

7.5. Future Prospects and Conclusion

In the context of data dissemination, a notable area of future research would be to investigate how the proposed regionalisation methodology could be integrated as part of a database querying system (see de Montjoye *et al.*, 2018). This would mean users could request a data aggregate of the in-app data or other sensitive point datasets, specifying the criteria required, and receive an aggregated product meeting these requirements. This would be interesting as researchers would not automatically depend on a direct access to the raw data to obtain more specific aggregates. For instance, one could specify if they would like total impressions or unique device counts, which time scale and zone are preferred, which geography to merge the outputs in, and which thresholds to meet. This carries over this work and the dataset from a restricted access model to a querying system model and would provide valuable insights on what is feasible in terms of data-driven bespoke regionalisation for project-specific requests.

Avenues for future research should also include more exploration of time analysis in conjunction with space. The conclusions drawn from Chapter 6 emphasised the need to design regions for both space and time, and seeking ways to establish how to simultaneously account for the two together rather than as separate elements. Perhaps considering space and time variograms or space-time point patterns, exploring ways to more systematically measure the stability of spatial units over time, and ultimately designing the first *space time units* could be exciting future developments which would build upon the exploratory space-time research conducted in this thesis.

As was demonstrated throughout this discussion section, this research carried significant amounts of uncertainty. The aim here was to focus on addressing key issues of data aggregation, but much more could be done with regards to adapting in-app data for mobility research. Particularly, it is deemed necessary to explore avenues of aggregation which could produce non-static grids, and consider key concepts in mobility analysis, namely origin and destination flows, and a clearer notion of movement across different key spaces at different times. Furthermore, developing case studies that apply the in-app data to specific problems, and address a particular social or public issue could be beneficial. This approach would clearly demonstrate the scenarios in which these datasets can offer social value or mobility insights. It is hoped that this work lays the groundwork for understanding the types of use cases that might be applicable. Several such research avenues may rely on robust aggregated products, where access to raw data is not available. Thus, we hope that this work also provided the initial dataset and theoretical basis required to encourage and inform future analysis conducted on these interesting and novel data sources.

Overall, although additional progress is necessary, this study offers encouraging evidence that granular insights into the population can be drawn from innovative location datasets in an ethical and non-disclosive way, using bespoke data-driven regionalisation. In-app data carries significant potential, but also significant risks. Their ethical implications are often poorly understood, making it critical to gain further access into the datasets to define what constitutes appropriate use, in both technical and ethical terms. Though they may not be appropriate substitutes for other established data sources (such as the census) their granularity can be harnessed to propose new ways of approaching mobility analysis and conduct research at scales previously too disclosive.

Finally, the MAUP and MTUP are well-known phenomena in geographic analysis. This work demonstrated that they cannot be neglected when creating data product for use in research, especially at granular scales for which the tolerable margins of error are smaller. With the technology made available to us as geographers, there is no longer a strict necessity to drastically sacrifice data quality and analytical completeness in the name of privacy protection, as disclosure control principles can be respected using new bespoke methodologies.

References

- Achara, J.P., Acs, G. and Castelluccia, C. (2015) ‘On the Unicity of Smartphone Applications’, *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, pp. 27–36. Available at: <https://doi.org/10.1145/2808138.2808146>.
- Alan Turing Institute (2021) *Ecosystems of digital twins* | *The Alan Turing Institute*. Available at: <https://www.turing.ac.uk/research/research-projects/ecosystems-digital-twins> (Accessed: 15 September 2021).
- Alessandretti, L., Aslak, U. and Lehmann, S. (2020) ‘The scales of human mobility’, *Nature* 2020 587:7834, 587(7834), pp. 402–407. Available at: <https://doi.org/10.1038/s41586-020-2909-1>.
- Allshouse, W.B. *et al.* (2010) ‘Geomasking sensitive health data and privacy protection: an evaluation using an E911 database’, *Geocarto international*, 25(6), pp. 443–452. Available at: <https://doi.org/10.1080/10106049.2010.496496>.
- Almuhimedi, H. *et al.* (2015) ‘Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging’, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery (CHI ’15), pp. 787–796. Available at: <https://doi.org/10.1145/2702123.2702210>.
- Antoine, S. and Weill, G. (1968) ‘Les metropoles et leur region’, *L’Espace et les Poles de Croissance* [Preprint].
- Ash, A. (2016) *Whistleblowing and Ethics in Health and Social Care*. Jessica Kingsley Publishers.
- Atkinson, P.M. and Tate, N.J. (2000) ‘Spatial Scale Problems and Geostatistical Solutions: A Review’, *The Professional Geographer*, 52(4), pp. 607–623. Available at: <https://doi.org/10.1111/0033-0124.00250>.
- Aubin, M. (2020) *ndjson Newline Delimited JSON*, *ndjson.org*. Available at: <http://ndjson.org/> (Accessed: 14 August 2023).
- Aurenhammer, F., Klein, R. and Lee, D.-T. (2013) *Voronoi Diagrams and Delaunay Triangulations*. Available at: <https://www.worldscientific.com/worldscibooks/10.1142/8685#t=aboutBook> (Accessed: 20 September 2023).
- Axhausen, K.W. (1995) ‘Travel diaries: An annotated catalogue’, p. 133 p. Available at: <https://doi.org/10.3929/ETHZ-B-000024848>.
- Azaele, S. *et al.* (2009) ‘Predicting spatial similarity of freshwater fish biodiversity’, *Proceedings of the National Academy of Sciences*, 106(17), pp. 7058–7062. Available at: <https://doi.org/10.1073/pnas.0805845106>.

Bampoulidis, A. *et al.* (2020) ‘Privately Connecting Mobility to Infectious Diseases via Applied Cryptography’. Available at: <https://eprint.iacr.org/2020/522> (Accessed: 14 March 2023).

Basellini, U. *et al.* (2020) ‘Linking excess mortality to Google mobility data during the COVID-19 pandemic in England and Wales’. Available at: <https://doi.org/10.31235/OSF.IO/75D6M>.

Becker, R. *et al.* (2013) ‘Human Mobility Characterization from Cellular Network Data’, 56(1). Available at: <https://doi.org/10.1145/2398356.2398375>.

Bengtsson, L. *et al.* (2015) ‘Using Mobile Phone Data to Predict the Spatial Spread of Cholera’, *Scientific Reports*, 5. Available at: <https://doi.org/10.1038/srep08923>.

Berry, B.J. (1965) ‘Research frontiers in urban geography’, *The study of urbanization*, pp. 403–430.

Berry, B.J.L., Barnum, H.G. and Tennant, R.J. (1962) ‘Retail location and consumer behavior’, *Papers of the Regional Science Association*, 9(1), pp. 65–106. Available at: <https://doi.org/10.1007/BF01947623>.

Berry, T. *et al.* (2016) ‘Using workplace population statistics to understand retail store performance’, *The International Review of Retail, Distribution and Consumer Research*, 26(4), pp. 375–395. Available at: <https://doi.org/10.1080/09593969.2016.1170066>.

Bertini, R.L. and El-Geneidy, A. (2003) ‘Generating Transit Performance Measures with Archived Data’, *Transportation Research Record*, 1841(1), pp. 109–119. Available at: <https://doi.org/10.3141/1841-12>.

Biehl, A., Ermagun, A. and Stathopoulos, A. (2018) ‘Community mobility MAUP-ing: A socio-spatial investigation of bikeshare demand in Chicago’, *Journal of Transport Geography*, 66, pp. 80–90. Available at: <https://doi.org/10.1016/j.jtrangeo.2017.11.008>.

Birenboim, A. and Shoval, N. (2016) ‘Mobility Research in the Age of the Smartphone’, *Annals of the American Association of Geographers*, 106(2), pp. 1–9. Available at: <https://doi.org/10.1080/00045608.2015.1100058>.

Bock, M. *et al.* (2010) ‘Generalized Voronoi Tessellation as a Model of Two-dimensional Cell Tissue Dynamics’, *Bulletin of Mathematical Biology*, 72(7), pp. 1696–1731. Available at: <https://doi.org/10.1007/s11538-009-9498-3>.

Bondaruk, B., Roberts, S.A. and Robertson, C. (2019) ‘Discrete Global Grid Systems: Operational Capability of the Current State of the Art’, *Spatial Knowledge and Information Canada*, 7(6), p. 1.

Brennetot, A. and Ruffray, S. de (2015) *Une nouvelle carte des régions françaises, Géoconfluences*. École normale supérieure de Lyon. Available at: <http://geoconfluences.ens-lyon.fr/informations-scientifiques/dossiers-regionaux/la-france-des-territoires-en-mutation/articles-scientifiques/regions-francaises> (Accessed: 4 September 2023).

Brockmann, D., Hufnagel, L. and Geisel, T. (2006) ‘The scaling laws of human travel’, *Nature*, 439(7075), pp. 462–465. Available at: <https://doi.org/10.1038/nature04292>.

- Brodsky, I. (2018) *H3: Uber's Hexagonal Hierarchical Spatial Index*, *Uber Blog*. Available at: <https://www.uber.com/en-GB/blog/h3/> (Accessed: 28 November 2022).
- Brodsky, I. and Uber Technologies Inc. (2015) *H3: A hexagonal hierarchical geospatial indexing system*. Available at: <https://h3geo.org/#/> (Accessed: 28 November 2022).
- Brunet, R., Ferras, R. and Théry, H. (1993) *Les Mots de la Géographie, Dictionnaire Critique*. 3rd edn. Montpellier-Paris: Reclus - La Documentation Française.
- Brunsdon, C. and Comber, A. (2020) 'Opening practice: supporting reproducibility and critical spatial data science', *Journal of Geographical Systems*, 23, pp. 477–496. Available at: <https://doi.org/10.1007/s10109-020-00334-2>.
- Bustos, M.F.A. *et al.* (2020) 'A pixel level evaluation of five multitemporal global gridded population datasets: a case study in Sweden, 1990–2015', *Population and Environment* 2020 42:2, 42(2), pp. 255–277. Available at: <https://doi.org/10.1007/S11111-020-00360-8>.
- Bwambale, A. *et al.* (2020) 'Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling', *Transportation* 2020, pp. 1–28. Available at: <https://doi.org/10.1007/S11116-020-10129-5>.
- Cabinet Office (2020) *[Withdrawn] Staying at home and away from others (social distancing)*, *GOV.UK*. Available at: <https://www.gov.uk/government/publications/full-guidance-on-staying-at-home-and-away-from-others/full-guidance-on-staying-at-home-and-away-from-others> (Accessed: 19 July 2023).
- Calabrese, F. *et al.* (2013) 'Understanding individual mobility patterns from urban sensing data: A mobile phone trace example - ScienceDirect', *Transportation research. Part C, Emerging technologies*, 26. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0968090X12001192> (Accessed: 29 July 2023).
- Casado Díaz, J. and Coombes, M. (2011) 'The Delineation of 21st Century Local Labour Market Areas: A Critical Review and a Research Agenda', *Boletín de la Asociación de Geógrafos Españoles*, pp. 7–32.
- CDRC (2022) *Protecting Data | CDRC Data*. Available at: <https://data.cdrc.ac.uk/protecting-data> (Accessed: 15 June 2023).
- Chai, Y. (2013) 'Space—Time Behavior Research in China: Recent Development and Future Prospect', *Annals of the Association of American Geographers*, 103(5), pp. 1093–1099.
- Chang, S. *et al.* (2021) 'Mobility network models of COVID-19 explain inequities and inform reopening', *Nature*, 589(7840), pp. 82–87. Available at: <https://doi.org/10.1038/s41586-020-2923-3>.
- Cheng, T. and Adepeju, M. (2014) 'Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-Time Cluster Detection', *PLoS ONE*, 9(6), p. e100465. Available at: <https://doi.org/10.1371/journal.pone.0100465>.

Cheshire, J. (2020) 'Code/Space', in A. Kobayashi (ed.) *International Encyclopedia of Human Geography (Second Edition)*. Oxford: Elsevier, pp. 303–308. Available at: <https://doi.org/10.1016/B978-0-08-102295-5.10524-4>.

CMA (2015) *The commercial use of consumer data - Report on the CMA's call for information*. 38. Available at: www.nationalarchives.gov.uk/doc/open-government- (Accessed: 2 September 2021).

Cockings, S. *et al.* (2011b) 'Maintaining Existing Zoning Systems Using Automated Zone-Design Techniques: Methods for Creating the 2011 Census Output Geographies for England and Wales', *Environment and Planning A: Economy and Space*, 43(10), pp. 2399–2418. Available at: <https://doi.org/10.1068/a43601>.

Cockings, S. *et al.* (2013) 'Getting the Foundations Right: Spatial Building Blocks for Official Population Statistics', *Environment and Planning A: Economy and Space*, 45(6), pp. 1403–1420. Available at: <https://doi.org/10.1068/a45276>.

Cockings, S. and Martin, D. (2005) 'Zone design for environment and health studies using pre-aggregated data', *Social Science & Medicine*, 60(12), pp. 2729–2742. Available at: <https://doi.org/10.1016/j.socscimed.2004.11.005>.

Cöltekin, A. *et al.* (2011) 'Modifiable temporal unit problem'. Available at: <https://doi.org/10.5167/UZH-54263>.

Cook, S. (2018) 'Geographies of mobility: a brief introduction', *Geography*, 103(3), pp. 137–145.

Coombes, M. *et al.* (1982) 'Functional regions for the population census of Great Britain', in. Available at: <https://www.semanticscholar.org/paper/Functional-regions-for-the-population-census-of-Coombes-Dixon/eba727f1d168fe0aebc421d1a93909335be55cbd> (Accessed: 20 June 2023).

Costa, L. da F. (2021) 'Further Generalizations of the Jaccard Index', *Computer Science and Machine Learning* [Preprint]. Available at: <https://doi.org/10.48550/arXiv.2110.09619>.

Cowpertwait, P.S.P. (2011) 'A regionalization method based on a cluster probability model', *Water Resources Research*, 47(11). Available at: <https://doi.org/10.1029/2011WR011084>.

Crampton, J.W. *et al.* (2013) 'Beyond the geotag: Situating "big data" and leveraging the potential of the geoweb', *Cartography and Geographic Information Science*, 40(2), pp. 130–139. Available at: <https://doi.org/10.1080/15230406.2013.777137>.

Crang, M. (2001) 'Rhythms of the City: temporalised space and motion.', in *Timespace*. London: Routledge. Available at: https://www.academia.edu/27783141/Crang_M_2001_Rhythms_of_the_City_temporalised_space_and_motion_TimeSpace_J_May_and_N_Thrift_London_Routledge_187_207 (Accessed: 4 September 2023).

Crawford, K. (2013) 'The Hidden Biases in Big Data', *Harvard Business Review*, 1 April. Available at: <https://hbr.org/2013/04/the-hidden-biases-in-big-data> (Accessed: 4 September 2023).

- Dalton, C.M. and Thatcher, J. (2015) 'Inflated granularity: Spatial "Big Data" and geodemographics', *Big Data & Society*, 2(2), p. 2053951715601144. Available at: <https://doi.org/10.1177/2053951715601144>.
- D'Angelo, A. (2016) 'A Brief Introduction to Quadrees and Their Applications', in. Available at: <https://www.semanticscholar.org/paper/A-Brief-Introduction-to-Quadrees-and-Their-D%27Angelo/beffc40cd086948bbfcd28b3a3ce2b974498aadb> (Accessed: 19 September 2023).
- De Meersman, F. *et al.* (2016) *Assessing the Quality of Mobile Phone Data as a Source of Statistics*. Available at: <http://economie.fgov.be/nl/consument/Internet/telecommunicatie/teledistributie/> (Accessed: 21 April 2021).
- De Sousa, L. *et al.* (2006) 'Assessing the accuracy of hexagonal versus square tiled grids in preserving DEM surface flow directions'.
- Degirmenci, K. (2020) 'Mobile users' information privacy concerns and the role of app permission requests', *International Journal of Information Management*, 50, pp. 261–272. Available at: <https://doi.org/10.1016/J.IJINFOMGT.2019.05.010>.
- Dewille, P. *et al.* (2014) 'Dynamic population mapping using mobile phone data', 111(45), pp. 15888–15893. Available at: <https://doi.org/10.1073/pnas.1408439111>.
- Diestel, R. (2000) *Graph theory*. 2. ed. New York, NY: Springer (Graduate texts in mathematics, 173).
- Diggelen, F.S.T.V. (2009) *A-GPS: Assisted GPS, GNSS, and SBAS*. Artech House.
- D'Ignazio, C. and Klein, L. (2020) *Data Feminism*. Available at: <https://data-feminism.mitpress.mit.edu/pub/visobxh7/release/4>; (Accessed: 16 March 2023).
- van Dijck, J. (2014) 'Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology', *Surveillance and Society*, 12(2), pp. 197–208. Available at: <https://doi.org/10.24908/ss.v12i2.4776>.
- Diker, A. and Nasibov, E. (2012) 'Estimation of traffic congestion level via FN-DBSCAN algorithm by using GPS data', in, pp. 1–4. Available at: <https://doi.org/10.1109/ICPCI.2012.6486279>.
- Domingo-Ferrer, J. and Mateo-Sanz, J.M. (2002) 'Practical data-oriented microaggregation for statistical disclosure control', *IEEE Transactions on Knowledge and Data Engineering*, 14(1), pp. 189–201. Available at: <https://doi.org/10.1109/69.979982>.
- Dong, P. (2008) 'Generating and updating multiplicatively weighted Voronoi diagrams for point, line and polygon features in GIS', *Computers & Geosciences*, 34(4), pp. 411–421. Available at: <https://doi.org/10.1016/j.cageo.2007.04.005>.
- Duke-Williams, O. and Rees, P. (1998) 'Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure', *International Journal of Geographical Information Science*, 12(6), pp. 579–605. Available at: <https://doi.org/10.1080/136588198241680>.

Dusek, T. (2004) ‘Spatially aggregated data and variables in empirical analysis and model building for economics’, *Cybergeo: European Journal of Geography* [Preprint]. Available at: <https://doi.org/10.4000/cybergeo.2654>.

Ebdon, D. (1992) ‘SPANS—A quadtree-based GIS’, *Computers & Geosciences*, 18(4), pp. 471–475. Available at: [https://doi.org/10.1016/0098-3004\(92\)90077-5](https://doi.org/10.1016/0098-3004(92)90077-5).

Edensor, T. (2016) ‘Rhythm’, in *Urban Theory*. Routledge.

ESRI (2020) *Boundary Effect Definition, GIS Dictionary*. Available at: <https://support.esri.com/en-us/gis-dictionary/boundary-effect> (Accessed: 20 September 2023).

EthicalGeo (2021) *Locus Charter*.

Eurostat and Coombes, M.G. (1992) *Étude sur les zones d’emploi*. 20. Luxembourg: Office for Official Publications of the European Communities.

Farmer, C.J.Q. and Fotheringham, S. (2011) ‘Network-based functional regions’, *Environment and Planning A*, 43(11), pp. 2723–2741. Available at: <https://doi.org/10.1068/a44136>.

Fletcher, S. and Islam, M.Z. (2018) ‘Comparing sets of patterns with the Jaccard index’, *Australasian Journal of Information Systems*, 22. Available at: <https://doi.org/10.3127/ajis.v22i0.1538>.

Forgó, N. *et al.* (2021) ‘An ethico-legal framework for social data science’, *International Journal of Data Science and Analytics*, 11(4), pp. 377–390. Available at: <https://doi.org/10.1007/s41060-020-00211-7>.

Fortunato, S. (2010) ‘Community detection in graphs’, *Physics Reports*, 486(3), pp. 75–174. Available at: <https://doi.org/10.1016/j.physrep.2009.11.002>.

Fotheringham, A.S., Densham, P.J. and Curtis, A. (1995) ‘The Zone Definition Problem in Location-Allocation Modeling’, *Geographical Analysis*, 27(1), pp. 60–77. Available at: <https://doi.org/10.1111/j.1538-4632.1995.tb00336.x>.

Fotheringham, A.S. and Rogerson, P.A. (1993) ‘GIS and spatial analytical problems’, *International journal of geographical information systems*, 7(1), pp. 3–19. Available at: <https://doi.org/10.1080/02693799308901936>.

Gahegan, M.N. (1989) ‘An efficient use of quadtrees in a geographical information system’, *International journal of geographical information systems*, 3(3), pp. 201–214. Available at: <https://doi.org/10.1080/02693798908941508>.

Garfinkel, S. (2022) ‘Differential Privacy and the 2020 US Census’, *MIT Case Studies in Social and Ethical Responsibilities of Computing* [Preprint], (Winter 2022). Available at: <https://doi.org/10.21428/2c646de5.7ec6ab93>.

Garrison, W.L. *et al.* (1959) ‘Studies of highway development and geographic change’. Available at: <https://trid.trb.org/view/89512> (Accessed: 4 September 2023).

GDPR (2018a) *Art. 5 GDPR – Principles relating to processing of personal data, General Data Protection Regulation (GDPR)*. Available at: <https://gdpr-info.eu/art-5-gdpr/> (Accessed: 26 August 2024).

GDPR (2018b) *Art. 7 GDPR – Conditions for consent | General Data Protection Regulation (GDPR)*. Available at: <https://gdpr-info.eu/art-7-gdpr/> (Accessed: 17 August 2021).

Geofabrik (2022) *Geofabrik, Maps and Data*. Available at: <https://www.geofabrik.de/data/> (Accessed: 28 November 2022).

Georgiadou, Y., By, R.A. de and Kounadi, O. (2019) ‘Location Privacy in the Wake of the GDPR’, *ISPRS International Journal of Geo-Information* 2019, Vol. 8, Page 157, 8(3), p. 157. Available at: <https://doi.org/10.3390/IJGI8030157>.

Gibbons, A. (1985) *Algorithmic Graph Theory*. Cambridge University Press.

Gibbs, H. *et al.* (2021) ‘Population disruption: estimating changes in population distribution in the UK during the COVID-19 pandemic’, *medRxiv*, p. 2021.06.22.21259336. Available at: <https://doi.org/10.1101/2021.06.22.21259336>.

Gibbs, H., Eggo, R.M. and Cheshire, J. (2024) ‘Detecting behavioural bias in GPS location data collected by mobile applications’. *medRxiv*, p. 2023.11.06.23298140. Available at: <https://doi.org/10.1101/2023.11.06.23298140>.

Giraudel, J.L. and Lek, S. (2001) ‘A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination’, *Ecological Modelling*, 146(1), pp. 329–339. Available at: [https://doi.org/10.1016/S0304-3800\(01\)00324-6](https://doi.org/10.1016/S0304-3800(01)00324-6).

González, M.C., Hidalgo, C.A. and Barabási, A.L. (2008) ‘Understanding individual human mobility patterns’, *Nature*, 453(7196), pp. 779–782. Available at: <https://doi.org/10.1038/nature06958>.

González-Bailón, S. (2013) ‘Big data and the fabric of human geography’, <https://doi.org/10.1177/2043820613515379>, 3(3), pp. 292–296. Available at: <https://doi.org/10.1177/2043820613515379>.

Goodchild, M.F. (2007) ‘Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0’, *International Journal of Spatial Data Infrastructures Research*, 2(2), pp. 24–32.

Goodchild, M.F. (2013) ‘The quality of big (geo)data’, *Sage*, 3(3). Available at: <https://journals.sagepub.com/doi/10.1177/2043820613513392> (Accessed: 4 September 2023).

Goodchild, M.F. (2018) ‘Reimagining the history of GIS’, *Annals of GIS*, 24(1), pp. 1–8. Available at: <https://doi.org/10.1080/19475683.2018.1424737>.

Goodchild, M.F. and Proctor, J. (1997) ‘Scale in a digital geographic world’, *Geographical and Environmental Modelling*, 1(1), pp. 5–23.

Google (2020) *COVID-19 Community Mobility Reports*. Available at: <https://www.google.com/covid19/mobility/> (Accessed: 15 September 2021).

- Greater London Authority (2024) *London at Night Update*. Available at: <https://data.london.gov.uk/blog/london-at-night-an-updated-evidence-base-for-a-24-hour-city/> (Accessed: 25 August 2024).
- Grieco, M. and Urry, J. (2012) 'Introduction: Introducing the Mobilities Turn', in *Mobilities: New Perspectives on Transport and Society*. Routledge.
- Griffiths, E. *et al.* (2019) 'Handbook on Statistical Disclosure Control for Outputs'.
- Gu, X. and Shen, Q. (2022) 'Self-organizing Divisive Hierarchical Voronoi Tessellation-based classifier', *Information Sciences*, 603, pp. 106–129. Available at: <https://doi.org/10.1016/j.ins.2022.04.049>.
- GVR (2022) *Location Intelligence Market Size, Share & Trends Analysis Report*. GVR-2-68038-401-7. Grand View Research. Available at: <https://www.grandviewresearch.com/industry-analysis/location-intelligence-market> (Accessed: 28 February 2023).
- Hagen, T., Hamann, J. and Saki, S. (2022) 'Discretization of Urban Areas using POI-based Tessellation'. Available at: <https://doi.org/10.48718/7JJR-1C66>.
- Hägerstrand, T. (1970) 'What About People in Regional Science?', *Papers in Regional Science*, 24(1), pp. 7–24. Available at: <https://doi.org/10.1111/j.1435-5597.1970.tb01464.x>.
- Hägerstrand, T. (1974) 'On socio-technical ecology and the study of innovations', *Ethnologia Europaea*, 7(1). Available at: <https://doi.org/10.16995/ee.3295>.
- Hägerstrand, T. (1975a) 'On the definition of migration', in *Readings in Social Geography*. E.Jones. Oxford University Press, pp. 200–209.
- Hägerstrand, T. (1975b) 'Space, time and human conditions', in *Dynamic allocations of urban space*. Farnborough: Saxon House/Lexington, D.C Heath, pp. 3–12.
- Hahsler, M. and Piekenbrock, M. (2022) 'dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms'.
- Halás, M. *et al.* (2015) 'An alternative definition and use for the constraint function for rule-based methods of functional regionalisation', *Environment and Planning A: Economy and Space*, 47(5), pp. 1175–1191. Available at: <https://doi.org/10.1177/0308518X15592306>.
- Hales, T.C. (2001) 'The Honeycomb Conjecture', *Discrete & Computational Geometry*, 25(1), pp. 1–22. Available at: <https://doi.org/10.1007/s004540010071>.
- Haley, A. (2017) 'Defining geographical mobility: Perspectives from higher education', *Geoforum*, 83, pp. 50–59. Available at: <https://doi.org/10.1016/j.geoforum.2017.04.013>.
- Hallisey, E. *et al.* (2017) 'Transforming geographic scale: a comparison of combined population and areal weighting to other interpolation methods', *International Journal of Health Geographics* 2017 16:1, 16(1), pp. 1–16. Available at: <https://doi.org/10.1186/S12942-017-0102-Z>.

Han, K. (2021) ‘Broken promises: How Singapore lost trust on contact tracing privacy | MIT Technology Review’, *MIT Technology Review*. Available at: <https://www.technologyreview.com/2021/01/11/1016004/singapore-tracetogether-contact-tracing-police/> (Accessed: 16 February 2021).

Harrower, M., MacEachren, A. and Griffin, A.L. (2000) ‘Developing a Geographic Visualization Tool to Support Earth Science Learning’, *Cartography and Geographic Information Science*, 27(4), pp. 279–293. Available at: <https://doi.org/10.1559/152304000783547759>.

Helbich, M., Mute Browning, M.H.E. and Kwan, M.-P. (2021) ‘Time to address the spatiotemporal uncertainties in COVID-19 research: Concerns and challenges’, *The Science of the Total Environment*, 764, p. 142866. Available at: <https://doi.org/10.1016/j.scitotenv.2020.142866>.

Helmpecht, B. and Schackis, D. (1996) *Manual on disclosure control methods*. Luxembourg : Lanham, MD: Office for Official Publications of the European Communities ; UNIPUB, [distributor] (Theme 9--Research and development).

Holmes, J.H. and Haggett, P. (1977) ‘Graph Theory Interpretation of Flow Matrices: A Note on Maximization Procedures for Identifying Significant Links’, *Geographical Analysis*, 9(4), pp. 388–399. Available at: <https://doi.org/10.1111/j.1538-4632.1977.tb00591.x>.

Hormann, K. and Agathos, A. (2001) ‘The point in polygon problem for arbitrary polygons’, *Computational Geometry*, 20(3), pp. 131–144. Available at: [https://doi.org/10.1016/S0925-7721\(01\)00012-8](https://doi.org/10.1016/S0925-7721(01)00012-8).

Hornsby, K. and Egenhofer, M.J. (2002) ‘Modeling Moving Objects over Multiple Granularities’, *Annals of Mathematics and Artificial Intelligence*, 36(1), pp. 177–194. Available at: <https://doi.org/10.1023/A:1015812206586>.

Huang, C.-W. and Shih, T.-Y. (1997) ‘On the complexity of point-in-polygon algorithms’, *Computers & Geosciences*, 23(1), pp. 109–118. Available at: [https://doi.org/10.1016/S0098-3004\(96\)00071-4](https://doi.org/10.1016/S0098-3004(96)00071-4).

Hubrich, S. and Wittwer, R. (2017) ‘Effects of improvements to survey methods on data quality and precision – Methodological insights into the 10th wave of the cross-sectional household survey ”Mobility in Cities – SrV”’, *Transportation Research Procedia*, 25, pp. 2276–2286. Available at: <https://doi.org/10.1016/j.trpro.2017.05.436>.

Hut, D. *et al.* (2020) ‘Statistical Disclosure Control When Publishing on Thematic Maps’, in J. Domingo-Ferrer and K. Muralidhar (eds) *Privacy in Statistical Databases*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 195–205. Available at: https://doi.org/10.1007/978-3-030-57521-2_14.

Iacus, S.M. *et al.* (2021) ‘Anomaly detection of mobile positioning data with applications to COVID-19 situational awareness’, *Japanese Journal of Statistics and Data Science* [Preprint]. Available at: <https://doi.org/10.1007/s42081-021-00109-z>.

ICO (2022) *What is personal data?* ICO. Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/> (Accessed: 9 January 2023).

INSEE (2020) *Méthode de constitution des zones d'emploi 2020*, pp. 1–4.

Jaccard, P. (1912) 'The Distribution of the Flora in the Alpine Zone.', *New Phytologist*, 11(2), pp. 37–50. Available at: <https://doi.org/10.1111/J.1469-8137.1912.TB05611.X>.

Jain, P., Gyanchandani, M. and Khare, N. (2016) 'Big data privacy: a technological perspective and review', *Journal of Big Data* 2016 3:1, 3(1), pp. 1–25. Available at: <https://doi.org/10.1186/S40537-016-0059-Y>.

Järv, O., Tenkanen, H. and Toivonen, T. (2017) 'Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation', *International Journal of Geographical Information Science*, 31(8), pp. 1630–1651. Available at: <https://doi.org/10.1080/13658816.2017.1287369>.

Jeffrey, B. *et al.* (2020) 'Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with covid-19 social distancing interventions was high and geographically consistent across the UK', *Wellcome Open Research*, 5, pp. 1–10. Available at: <https://doi.org/10.12688/WELLCOMEOPENRES.15997.1>.

Jiang, B. *et al.* (2019) 'Higher order mobile coverage control with applications to clustering of discrete sets', *Automatica*, 102, pp. 27–33. Available at: <https://doi.org/10.1016/j.automatica.2018.12.028>.

Jiron, P. and Carrasco, J.A. (2020) 'Understanding Daily Mobility Strategies through Ethnographic, Time Use, and Social Network Lenses', *Sustainability*, 12(1), p. 312. Available at: <https://doi.org/10.3390/su12010312>.

Kang, Y. *et al.* (2020) 'Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic', *Scientific Data* 2020 7:1, 7(1), pp. 1–13. Available at: <https://doi.org/10.1038/s41597-020-00734-5>.

Keegan, J. and Ng, A. (2021) *There's a Multibillion-Dollar Market for Your Phone's Location Data – The Markup, The Markup*. Available at: <https://themarkup.org/privacy/2021/09/30/theres-a-multibillion-dollar-market-for-your-phones-location-data> (Accessed: 1 February 2023).

Kim, Y.-L. (2020) 'Data-driven approach to characterize urban vitality: how spatiotemporal context dynamically defines Seoul's nighttime', *International Journal of Geographical Information Science*, 34(6), pp. 1235–1256. Available at: <https://doi.org/10.1080/13658816.2019.1694680>.

Kishore, N. *et al.* (2020) 'Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis', *The Lancet Digital Health*, 2(11), pp. e622–e628. Available at: [https://doi.org/10.1016/S2589-7500\(20\)30193-X](https://doi.org/10.1016/S2589-7500(20)30193-X).

Kitchin, R. (2013) 'Big data and human geography: Opportunities, challenges and risks', *Dialogues in Human Geography*, 3(3), pp. 262–267. Available at: <https://doi.org/10.1177/2043820613513388>.

Kitchin, R. (2014a) 'Big Data, new epistemologies and paradigm shifts', *Big Data & Society*, 1(1), p. 205395171452848. Available at: <https://doi.org/10.1177/2053951714528481>.

- Kitchin, R. (2014b) *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. SAGE Publications Ltd. Available at: <https://doi.org/10.4135/9781473909472>.
- Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. Available at: <https://doi.org/10.7551/mitpress/9780262042482.001.0001>.
- Klapka, P., Ellegård, K. and Frantál, B. (2020) ‘What about Time-Geography in the post-Covid-19 era?’, *Moravian Geographical Reports*, 28(4), pp. 238–247. Available at: <https://doi.org/10.2478/mgr-2020-0017>.
- Kraak, M.J. (2000) ‘Visualisation of the time dimension’, in *Time in GIS: Issues in spatio-temporal modelling*. Netherlands Geodetic Commission (NCG), pp. 27–35. Available at: <https://research.utwente.nl/en/publications/visualisation-of-the-time-dimension> (Accessed: 16 August 2024).
- Kung, K.S. *et al.* (2014) ‘Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data’, *PLoS ONE*, 9(6), p. e96180. Available at: <https://doi.org/10.1371/journal.pone.0096180>.
- Kwan, M.-P. (1999) ‘Gender, the Home-Work Link, and Space-Time Patterns of Nonemployment Activities’, *Economic Geography*, 75(4), pp. 370–394. Available at: <https://doi.org/10.2307/144477>.
- Lagonigro, R., Oller, R. and Carles Martori, J. (2020a) ‘AQuadtree: an R Package for Quadtree Anonymization of Point Data.’, *R journal* [Preprint]. Available at: <https://doi.org/10.32614/RJ-2021-013> (Accessed: 14 September 2021).
- Lagonigro, R., Oller, R. and Carles Martori, J. (2020b) ‘Title Confidentiality of Spatial Point Data’. Available at: <https://doi.org/10.2436/20.8080.02.55>.
- Lange, O. and Perez, L. (2020) ‘Traffic prediction with advanced Graph Neural Networks’, *Deepmind Applied*, 3 September. Available at: <https://www.deepmind.com/blog/traffic-prediction-with-advanced-graph-neural-networks> (Accessed: 1 March 2023).
- Lansley, G. and Cheshire, J. (2018) ‘Challenges to representing the population from new forms of consumer data’, *Geography Compass*, 12(7), p. e12374. Available at: <https://doi.org/10.1111/gec3.12374>.
- Lapin, M. and Barnes, B.V. (1995) ‘Using the Landscape Ecosystem Approach to Assess Species and Ecosystem Diversity’, *Conservation Biology*, 9(5), pp. 1148–1158. Available at: <https://doi.org/10.1046/j.1523-1739.1995.9051134.x-11>.
- Lassave, P. and Haumont, A. (2001) *Mobilités Spatiales, Une question de société*. L’Harmattan. Available at: https://www.editions-harmattan.fr/index_harmattan.asp?navig=catalogue&obj=livre&razSqlClone=1&no=5607 (Accessed: 29 May 2023).
- Lazer, D. *et al.* (2009) ‘Life in the network: the coming of age of computational social science’, *Science*, 323(5915), pp. 721–723. Available at: <https://doi.org/10.1126/science.1167742>.

- Lefebvre, H. (2004) *Rhythmanalysis: Space, Time and Everyday Life*. Bloomsbury Publishing.
- Lenormand, M. *et al.* (2014) ‘Cross-Checking Different Sources of Mobility Information’, *PLOS ONE*, 9(8), p. e105184. Available at: <https://doi.org/10.1371/journal.pone.0105184>.
- Lévy, J. (2000) ‘Les nouveaux espaces de la mobilité’, in *Les Territoires de la mobilité*. Paris cedex 14: Presses Universitaires de France (Sciences sociales et sociétés), pp. 155–170. Available at: <https://doi.org/10.3917/puf.bonne.2000.01.0155>.
- Liang, Y. *et al.* (2022) ‘Region2Vec: Community Detection on Spatial Networks Using Graph Embedding with Node Attributes and Spatial Interactions’, in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pp. 1–4. Available at: <https://doi.org/10.1145/3557915.3560974>.
- Lloyd, A.S. (2018) *The Applications of Loyalty Card Data for Social Science*. Doctor of Philosophy (Ph.D.). University College London.
- Longley, P. (2005) *Geographic Information Systems and Science*. John Wiley & Sons.
- Longley, P. and Singleton, A. (2018) *London Output Area Classification (LOAC)*. London: Censur Information Scheme.
- Longley, P.A. *et al.* (2015) *Geographic Information Science and Systems*. John Wiley & Sons.
- Lovelace, R., Ballas, D. and Watson, M. (2014) ‘A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels’, *Journal of Transport Geography*, 34, pp. 282–296. Available at: <https://doi.org/10.1016/J.JTRANGE.2013.07.008>.
- Maas, P. *et al.* (2019) ‘Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery’, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Anchorage AK USA: ACM, pp. 3173–3173. Available at: <https://doi.org/10.1145/3292500.3340412>.
- Mai, W., Xie, S. and Chen, X. (2019) ‘Travel Time Estimation and Urban Key Routes Analysis Based on Call Detail Records Data: A Case Study of Guangzhou City’, in X.B. Zhai, B. Chen, and K. Zhu (eds) *Machine Learning and Intelligent Communications*. Cham: Springer International Publishing (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), pp. 659–676. Available at: https://doi.org/10.1007/978-3-030-32388-2_55.
- Martin, D. (2000) ‘Towards the Geographies of the 2001 UK Census of Population’, *Transactions of the Institute of British Geographers*, 25(3), pp. 321–332.
- Martin, D. (2003) ‘Extending the automated zoning procedure to reconcile incompatible zoning systems’, *International Journal of Geographical Information Science*, 17(2), pp. 181–196. Available at: <https://doi.org/10.1080/713811750>.

- Martin, D. (2010) 'Optimizing census geography: the separation of collection and output geographies', <http://dx.doi.org/10.1080/136588198241590>, 12(7), pp. 673–685. Available at: <https://doi.org/10.1080/136588198241590>.
- Massey, D. (2005) *For Space*. SAGE.
- McNutt, M. (2014) 'Journals unite for reproducibility', *Science*, 346(6210), pp. 679–679. Available at: <https://doi.org/10.1126/science.aaa1724>.
- Mearns, B. (2015) *QGIS Blueprints*. Packt Publishing Ltd.
- Meentemeyer, V. (1989) 'Geographical perspectives of space, time, and scale', *Landscape Ecology*, 3(3), pp. 163–173. Available at: <https://doi.org/10.1007/BF00131535>.
- Megerian, S. *et al.* (2005) 'Worst and best-case coverage in sensor networks', *Mobile Computing, IEEE Transactions on*, 4, pp. 84–92. Available at: [https://doi.org/10.1109/TMC.2005.15\(410\)](https://doi.org/10.1109/TMC.2005.15(410)).
- Mennis, J. (2019) 'Problems of Scale and Zoning', *Geographic Information Science & Technology Body of Knowledge*, 2019(Q1). Available at: <https://doi.org/10.22224/gistbok/2019.1.2>.
- Miller, H.J. (2010) 'The Data Avalanche Is Here. Shouldn't We Be Digging?', *Journal of Regional Science*, 50(1), pp. 181–201. Available at: <https://doi.org/10.1111/j.1467-9787.2009.00641.x>.
- Miller, H.J. (2017) 'Time Geography and Space-Time Prism', in D. Richardson *et al.* (eds) *International Encyclopedia of Geography*. Oxford, UK: John Wiley & Sons, Ltd, pp. 1–19. Available at: <https://doi.org/10.1002/9781118786352.wbieg0431>.
- Miller, H.J. and Goodchild, M.F. (2015) 'Data-driven geography', *GeoJournal*, 80(4), pp. 449–461.
- Minot, N. and Baulch, B. (2005) 'Poverty Mapping with Aggregate Census Data: What is the Loss in Precision?', *Review of Development Economics*, 9(1), pp. 5–24. Available at: <https://doi.org/10.1111/j.1467-9361.2005.00261.x>.
- Molloy, J. and Moeckel, R. (2017) 'Automated design of gradual zone systems', *Open Geospatial Data, Software and Standards* 2:1, 2(1), pp. 1–10. Available at: <https://doi.org/10.1186/S40965-017-0032-5>.
- de Montjoye, Y.A. *et al.* (2018) 'On the privacy-conscious use of mobile phone data', *Scientific Data*, 5(1), pp. 1–6. Available at: <https://doi.org/10.1038/sdata.2018.286>.
- Morphet, C.S. (1993) 'The Mapping of Small-Area Census Data—A Consideration of the Role of Enumeration District Boundaries', *Environment and Planning A: Economy and Space*, 25(9), pp. 1267–1277. Available at: <https://doi.org/10.1068/a251267>.
- Nemeškal, J., Ouředníček, M. and Pospíšilová, L. (2020) 'Temporality of urban space: daily rhythms of a typical week day in the Prague metropolitan area', *Journal of Maps*, 16(1), pp. 30–39. Available at: <https://doi.org/10.1080/17445647.2019.1709577>.

Neutens, T. *et al.* (2011) 'The relationship between opening hours and accessibility of public service delivery', *Journal of Transport Geography*, 25. Available at: <https://doi.org/10.1016/j.jtrangeo.2011.03.004>.

Office for National Statistics (2016a) *Early census-taking in England and Wales - Office for National Statistics*. Available at: <https://www.ons.gov.uk/census/2011census/howourcensusworks/aboutcensuses/censushistory/earlycensustakinginenglandandwales> (Accessed: 22 June 2023).

Office for National Statistics (2016b) *Output areas*. Available at: <https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas> (Accessed: 9 January 2023).

Office for National Statistics (2017) *Policy for social survey microdata*. Available at: <https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata> (Accessed: 9 January 2023).

Office for National Statistics (2020) *ONS methodology working paper series no. 8- Statistical uses for mobile phone data: literature review - Office for National Statistics*. Available at: <https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno8statisticalusesformobilephonedataliteraturereview> (Accessed: 21 April 2021).

Oliver, N. *et al.* (2020) 'Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle', *Science Advances*, 6(23), p. eabc0764. Available at: <https://doi.org/10.1126/sciadv.abc0764>.

Ollivro, J. (2005) 'Les classes mobiles', *L'Information Géographique*, 69(3), pp. 28–44. Available at: <https://doi.org/10.3406/ingeo.2005.3008>.

Openshaw, S. (1977) 'A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling', *Transactions of the Institute of British Geographers*, 2(4), pp. 459–472. Available at: <https://doi.org/10.2307/622300>.

Openshaw, S. (1979) 'A million or so correlation coefficients, three experiments on the modifiable areal unit problem', *Statistical Applications in the Spatial Science*, pp. 127–144.

Openshaw, S. (1981) 'The modifiable areal unit problem', *Quantitative geography: a British view*, 68 A(2), pp. 60–69. Available at: https://doi.org/10.4157/GRJ1984A.68.2_71.

Openshaw, S. (1984) 'Ecological Fallacies and the Analysis of Areal Census Data', *Environment and Planning A: Economy and Space*, 16(1), pp. 17–31. Available at: <https://doi.org/10.1068/a160017>.

Ordnance Survey (2020) *A Guide to Coordinate Systems in Great Britain*. Southampton: Ordnance Survey.

Oxford Reference (2011) *time-space geography*, *Oxford Reference*. Available at: <https://doi.org/10.1093/oi/authority.20110803104651770>.

Oyabu, Y. *et al.* (2013) 'Evaluating Reliability of Mobile Spatial Statistics', *NTT DOCOMO Technical Journal*, 14(3), pp. 16–23.

- Peuquet, D.J. (1994) 'It's About Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems', *Annals of the Association of American Geographers*, 84(3), pp. 441–461. Available at: <https://doi.org/10.1111/j.1467-8306.1994.tb01869.x>.
- Phillips, M. and Knoppers, B.M. (2019) 'Whose Commons? Data Protection as a Legal Limit of Open Science', *Journal of Law, Medicine & Ethics*, 47(1), pp. 106–111. Available at: <https://doi.org/10.1177/1073110519840489>.
- Phithakkitnukoon, S., Smoreda, Z. and Olivier, P. (2012) 'Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data', *PLoS ONE*, 7(6), p. e39253. Available at: <https://doi.org/10.1371/journal.pone.0039253>.
- Piastra, R. (2015) 'Brève histoire des régions françaises de Serge Antoine à François Hollande', la Revue des Collectivités Locales.
- Purdam, K. and Elliot, M. (2007) 'A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records', *Environment & Planning A*, 39(5), pp. 1101–1118. Available at: <https://doi.org/10.1068/a38335>.
- Qi, Y. and Wu, J. (1996) 'Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices', *Landscape Ecology*, 11(1), pp. 39–49. Available at: <https://doi.org/10.1007/BF02087112>.
- Qin, J. and Liao, F. (2021) 'Space–time prism in multimodal supernetwork - Part 1: Methodology', *Communications in Transportation Research*, 1, p. 100016. Available at: <https://doi.org/10.1016/j.commtr.2021.100016>.
- R Core Team (2024) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.
- Raento, M.U., Antti Oulasvirta, L. and Eagle, N. (2009) 'Smartphones An Emerging Tool for Social Scientists', *Sociological Methods & Research*, 37, pp. 426–454. Available at: <https://doi.org/10.1177/0049124108330005>.
- Reades, J. *et al.* (2007) 'Cellular Census: Explorations in Urban Data Collection'. Available at: www.computer.org/pervasive (Accessed: 15 August 2021).
- Roper, C. (2015) *GitHub - charlesroper/OSGB_Grids: A collection of Ordnance Survey National Grids in Shapefile [OSGB 1936] and GeoJSON [WGS84] formats*. Available at: https://github.com/charlesroper/OSGB_Grids (Accessed: 18 September 2023).
- Rosales, A., Fernández-Ardèvol, M. and Svensson, J. (2023) *Digital Ageism: How it Operates and Approaches to Tackling it*. 1st edn. London: Routledge. Available at: <https://doi.org/10.4324/9781003323686>.
- Roux, V. (2020) 'The French rolling census: A census that allows a progressive modernization', *Statistical Journal of the IAOS*, 36(1), pp. 125–134. Available at: <https://doi.org/10.3233/SJI-190572>.

- Ruggles, S. and Van Riper, D. (2022) ‘The Role of Chance in the Census Bureau Database Reconstruction Experiment’, *Population Research and Policy Review*, 41(3), pp. 781–788. Available at: <https://doi.org/10.1007/s11113-021-09674-3>.
- Ruppert, E., Isin, E. and Bigo, D. (2017) ‘Data politics’, *Big Data & Society* [Preprint]. Available at: <https://doi.org/10.1177/2053951717717749>.
- Sahebgharani, A., Mohammadi, M. and Haghshenas, H. (2019) ‘Computing Spatiotemporal Accessibility to Urban Opportunities: A Reliable Space-Time Prism Approach in Uncertain Urban Networks’, *Computation*, 7(3), p. 51. Available at: <https://doi.org/10.3390/computation7030051>.
- Sahr, K. (2011) ‘Hexagonal discrete global grid systems for geospatial computing’, *Archiwum Fotogrametrii, Kartografii i Teledetekcji*, 22, pp. 363–376.
- Sahr, K., White, D. and Kimerling, A.J. (2003) *Geodesic Discrete Global Grid Systems, Cartography and Geographic Information Science*, pp. 121–134.
- Salat, H., Smoreda, Z. and Schläpfer, M. (2020) ‘A method to estimate population densities and electricity consumption from mobile phone data in developing countries’, *PLOS ONE*. Edited by S. Fu, 15(6), p. e0235224. Available at: <https://doi.org/10.1371/journal.pone.0235224>.
- Schubert, E. *et al.* (2017) ‘DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN’, *ACM Transactions on Database Systems*, 42(3), pp. 1–21. Available at: <https://doi.org/10.1145/3068335>.
- Schwanen, T. (2009) ‘Time-Space Diaries’, in R. Kitchin and N. Thrift (eds) *International Encyclopedia of Human Geography*. Oxford: Elsevier, pp. 294–300. Available at: <https://doi.org/10.1016/B978-008044910-4.00547-2>.
- Segaud, M., Brun, J. and Driant, J.-C. (2003) *Dictionnaire critique de l’habitat et du logement*. Armand Colin. France. Available at: <https://www.eyrolles.com/BTP/Livre/dictionnaire-critique-de-l-habitat-et-du-logement-9782200261733/> (Accessed: 31 May 2023).
- Sekara, V. *et al.* (2021) ‘Temporal and cultural limits of privacy in smartphone app usage’, *Scientific Reports*, 11(1), p. 3861. Available at: <https://doi.org/10.1038/s41598-021-82294-1>.
- Sevtsuk, A. and Ratti, C. (2010) ‘Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks’, *Journal of Urban Technology*, 17(1), pp. 41–60. Available at: <https://doi.org/10.1080/10630731003597322>.
- Sieg, L. and Cheshire, J. (2023) ‘Regionalising mobile phone data: attributing time components to optimised regions to create region classifications’, in. Available at: <https://zenodo.org/records/7835017> (Accessed: 13 May 2024).
- Sieg, L. and Cheshire, J. (2024) ‘The Regionalization and Aggregation of In-App Location Data to Maximize Information and Minimize Data Disclosure’, *Geographical Analysis* [Preprint]. Available at: <https://doi.org/10.1111/gean.12406>.

- Sieg, L. and Chesire, J. (2021) 'Spatially aggregating mobility data – implications and challenges'. Available at: <https://doi.org/10.5281/ZENODO.4669903>.
- Silva, K.D.' *et al.* (2017) 'If I build it, will they come? Predicting new venue visitation patterns through mobility data'. Available at: <https://doi.org/10.1145/3139958.3140035>.
- Singleton, A., Longley, P. and Duckenfield, T. (2017) *London Workplace Zones Classification Technical Report*.
- Skinner, C. *et al.* (1994) 'Disclosure Control for Census Microdata', *Journal of Official Statistics*, 10(1), pp. 31–51.
- Slivinskas, G., Jensen, C.S. and Snodgrass, R.T. (2001) 'A foundation for conventional and temporal query optimization addressing duplicates and ordering', *IEEE Transactions on Knowledge and Data Engineering*, 13(1), pp. 21–49. Available at: <https://doi.org/10.1109/69.908979>.
- Smart, M.W. (1974) 'Labour market areas: Uses and definition', *Progress in Planning*, 2, pp. 239–353. Available at: [https://doi.org/10.1016/0305-9006\(74\)90008-7](https://doi.org/10.1016/0305-9006(74)90008-7).
- Snyder, J.P. (1992) 'An Equal-Area Map Projection For Polyhedral Globes', *Cartographica: The International Journal for Geographic Information and Geovisualization*, 29(1), pp. 10–21. Available at: <https://doi.org/10.3138/27H7-8K88-4882-1752>.
- Song, Y., Dahlmeier, D. and Bressan, S. (2014) 'Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data'.
- Stevens, N. (2006) *Changing Postal ZIP Code Boundaries*. Congressional Research Service.
- Sutherland, W.J. *et al.* (2012) 'A Collaboratively-Derived Science-Policy Research Agenda', *PLoS ONE*. Edited by E. von Elm, 7(3), p. e31824. Available at: <https://doi.org/10.1371/journal.pone.0031824>.
- Taylor, L. (2015) 'No place to hide? The ethics and analytics of tracking mobility using mobile phone data', <http://dx.doi.org/10.1177/0263775815608851>, 34(2), pp. 319–336. Available at: <https://doi.org/10.1177/0263775815608851>.
- Thrift, N.J. (1977) *An Introduction to Time-geography*. Geo Abstracts, University of East Anglia.
- Tiede, D. and Strobl, J. (2006) 'Polygon-based Regionalisation in a GIS Environment Modelling and forecasting land changes using cellular automata and Markov chain View project Sen2Cube.at (Sentinel-2 Semantic Data Cube Polygon-based Regionalisation in a GIS Environment'. Available at: <https://www.researchgate.net/publication/252121443> (Accessed: 7 July 2021).
- Tizzoni, M. *et al.* (2014) 'On the Use of Human Mobility Proxies for Modeling Epidemics', *PLoS Computational Biology*, 10(7). Available at: <https://doi.org/10.1371/journal.pcbi.1003716>.

- Tooru, O. and Kawakami, H. (2013) 'Using Mobile Spatial Statistics in Field of Urban Planning 2. The Need for Population Statistics in Urban Development', *NTT DOCOMO Technical Journal*, 14(3), pp. 31–36.
- Transport For London (2020) *Quieter times to travel*, *Transport for London*. Available at: <https://www.tfl.gov.uk/status-updates/busiest-times-to-travel> (Accessed: 18 September 2023).
- Trasberg, T. and Cheshire, J. (2020) 'Towards data-driven human mobility analysis', in. Available at: <https://huq.io/> (Accessed: 9 February 2021).
- UCL and GLA (2015) *London Output Area Classification - London Datastore*, *London Datastore*. Available at: <https://data.london.gov.uk/dataset/london-area-classification> (Accessed: 14 August 2023).
- UCL and GLA (2017) *London Workplace Zone Classification - London Datastore*. Available at: <https://data.london.gov.uk/dataset/london-workplace-zone-classification> (Accessed: 9 February 2021).
- Urry, J. (2007) *Mobilities*. Polity.
- Vallina-Rodriguez, N. *et al.* (2013) 'When assistance becomes dependence: characterizing the costs and inefficiencies of A-GPS', *ACM SIGMOBILE Mobile Computing and Communications Review*, 17(4), pp. 3–14. Available at: <https://doi.org/10.1145/2557968.2557970>.
- Vanhoof, M. *et al.* (2018) 'Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics', *Journal of Official Statistics*, 34(4), pp. 935–960. Available at: <https://doi.org/10.2478/jos-2018-0046>.
- Viegas, J.M., Martínez, L.M. and Silva, E.A. (2009) 'Effects of the modifiable areal unit problem on the delineation of traffic analysis zones', *Environment and Planning B: Planning and Design*, 36(4), pp. 625–643. Available at: <https://doi.org/10.1068/B34033>.
- Voronoi, G. (1908) *Nouvelle application des paramètres continus à la théorie des formes quadratiques*. Universitäts Göttingen.
- Walford, N.S. and Hayles, K.N. (2012) 'Thirty Years of Geographical (In)consistency in the British Population Census: Steps towards the Harmonisation of Small-Area Census Geography', *Population, Space and Place*, 18(3), pp. 295–313. Available at: <https://doi.org/10.1002/PSP.658>.
- Wang, F. *et al.* (2019) 'Extracting Trips from Multi-Sourced Data for Mobility Pattern Analysis: An App-Based Data Example', *Transportation research. Part C, Emerging technologies*, 105, pp. 183–202. Available at: <https://doi.org/10.1016/j.trc.2019.05.028>.
- Wang, F. and Chen, C. (2018) 'On data processing required to derive mobility patterns from passively-generated mobile phone data', *Transportation Research Part C: Emerging Technologies*, 87, pp. 58–74. Available at: <https://doi.org/10.1016/j.trc.2017.12.003>.
- Wang, J., Kim, J. and Kwan, M.P. (2022) 'An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data', *Cartography and Geographic Information Science*, 49. Available at:

https://doi.org/10.1080/15230406.2022.2056510/SUPPL_FILE/TCAG_A_2056510_SM4398.DOCX.

Watts, D. (2013) ‘Computational Social Science: Exciting Progress and Future Directions’, in *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2013 Symposium*. Available at: <https://doi.org/10.17226/18558>.

Wesolowski, A. *et al.* (2015) ‘Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data’, *POPULATION BIOLOGY* Downloaded by guest on August, 16, p. 2021. Available at: <https://doi.org/10.1073/pnas.1423542112>.

Wickham, H. (2014) ‘Tidy data’, *The Journal of Statistical Software*, 59(10). Available at: <http://www.jstatsoft.org/v59/i10/> (Accessed: 15 September 2021).

Willberg, E. *et al.* (2021) ‘Escaping from Cities during the COVID-19 Crisis: Using Mobile Phone Data to Trace Mobility in Finland’, *ISPRS International Journal of Geo-Information*, 10(2), p. 103. Available at: <https://doi.org/10.3390/ijgi10020103>.

Williams, S. (2020) *Data Action: Using Data for Public Good*. MIT Press.

Young, C., Martin, D. and Skinner, C. (2009) ‘Geographically intelligent disclosure control for flexible aggregation of census data’, *International Journal of Geographical Information Science*, 23(4), pp. 457–482. Available at: <https://doi.org/10.1080/13658810801949835>.

Zandbergen, P.A. (2014) ‘Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data’, *Advances in Medicine*, 2014, pp. 1–14. Available at: <https://doi.org/10.1155/2014/567049>.


Zang, H. and Bolot, J. (2011) ‘Anonymization of location data does not work: a large-scale measurement study’, *Proceedings of the 17th annual international conference on Mobile computing and networking - MobiCom '11*, p. 145. Available at: <https://doi.org/10.1145/2030613.2030630>.

Zhang, H. and McKenzie, G. (2022) ‘Rehumanize geoprivacy: from disclosure control to human perception’, *GeoJournal*, 88, pp. 189–208. Available at: <https://doi.org/10.1007/s10708-022-10598-4>.

Zhou, Y. *et al.* (2020) ‘Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data’, *The Lancet Digital Health*, 2(8), pp. e417–e424. Available at: [https://doi.org/10.1016/S2589-7500\(20\)30165-5](https://doi.org/10.1016/S2589-7500(20)30165-5).

Appendix

Appendix 1. CDRC User Guide, including Statistical Disclosure control guidance



Consumer
Data
Research
Centre

An ESRC Data
Investment

Introduction

CONSUMER DATA RESEARCH CENTRE DATA SERVICE USER GUIDE

Version: 7.0

The Consumer Data Research Centre (CDRC or Centre) is an academic led, multi-institution laboratory which discovers, mines, analyses and synthesises consumer-related datasets from around the UK. The CDRC forms part of the ESRC-funded Big Data network and offers a data service aimed at providing researchers with access to a wide range of consumer data to address many societal challenges. CDRC's key areas of interest include retail, transport, health, crime, housing, energy, mobility and sustainable consumption. We support the acquisition and analysis of data in these areas and others to achieve benefits for the 'public good'.

The purpose of this guide is to describe the Centre's data services and how researchers can access them. It identifies the different types of data the Centre holds and the service tiers through which these data sets are available. For data that are not publicly available, the guide details how researchers can register or apply for access and the kinds of support that is available to them.

CDRC Data Services

The CDRC provides data with three different levels of access. These correspond to the data levels described in the UK Data Service's three tier access policy:

- *Open data*: data which are freely available to all for any purpose. Data includes open datasets where CDRC have added value and non-sensitive and aggregated data and derivative products produced by the CDRC. Examples might include geodemographic data derived from the Census. Open data are accessed through the CDRC service via basic registration and download.
- *Safeguarded data*: data to which access is restricted due to licence conditions, but where data are not considered 'personally-identifiable' or otherwise sensitive – an example might include data from retail companies on store turnover. Access to safeguarded CDRC data is via a remote service that requires users to submit a project proposal. This proposal must receive approval from the Centre's Research Approvals Group (RAG) (see below) before

access to the data will be authorised. Users are able to retrieve data after authentication and authorisation by the service.

- **Controlled data:** data which need to be held under the most secure conditions with more stringent access restrictions, including data which are 'personally-identifiable' and therefore subject to Data Protection legislation or are considered commercially sensitive. Examples might include data on individual consumer purchases. Access to CDRC controlled data is provided through the CDRC-secure service. This service requires that individuals gain project

approval through the RAG and visit one of our secure facilities at either the University College London, University of Leeds or University of Liverpool.

Finding Data

All data available through the CDRC are accompanied by metadata that enable both attributes and geography to be searched.

Research Approvals Process

Access to both safeguarded and controlled data requires a process by which individuals submit project proposals for assessment and approval. The approval process is overseen by an independent Research Approvals Group (RAG) which comprises representation from the Data Partner(s) and the social science academic community. The Group may also draw upon the expertise from a social science ethics practitioner. The CDRC Senior Management Team provides comment on resource implications of a proposal. The composition ensures that the RAG has expertise in research design, analysis and impact, while also considering any commercial sensitivities a project may have. The RAG review process is overseen by the Chair of RAG.

For full details of the Research Approvals Process please see the Research Approvals Guidelines at <https://data.cdrc.ac.uk/using-our-data-services>.

Criteria for Approval

These criteria align with CDRC objectives and cover the following:

- **Scientific advancement** – how the project has the potential to advance scientific knowledge, understanding and/or methods using consumer data;
- **Public good** – how the project has the potential to provide insight and/or solutions that could benefit society;
- **Privacy and ethics** – the potential privacy impacts or risks, and wider ethical considerations relating to the project
- **Project Design and Methods** – how the project will be conducted and who will be involved with a focus on demonstrating project feasibility.
- **Cost and resources issues** – what impact the project is likely to have on CDRC resources, including CDRC staff time and use of infrastructure, as well as any data acquisition costs. Resource requirements should be justified.

The RAG typically considers applications remotely and is designed to be lightweight but robust, enabling timely decisions on user applications.

Approval will not be granted without evidence that the user has acquired ethical approval for the research through their institution, or supplied evidence that it

is not applicable. For non-academic projects, where there is no approval process in place the CDRC will assist the user with acquiring this.

Safe Researcher Training and Training and Development

Safe Research Training

Users, both academic and non-academic stakeholders, wishing to access controlled data and on occasion safeguarded data are required to have completed a safe researcher course, as offered by the Administrative Data Research Network (ADRN), HM Revenue and Customs (HMRC), Office for National Statistics (ONS) or the UK Data Service (UKDS). Evidence of valid accreditation for the duration of access to the data will be required. If the user has not previously completed such training the CDRC will offer access to training courses.

Training and Development

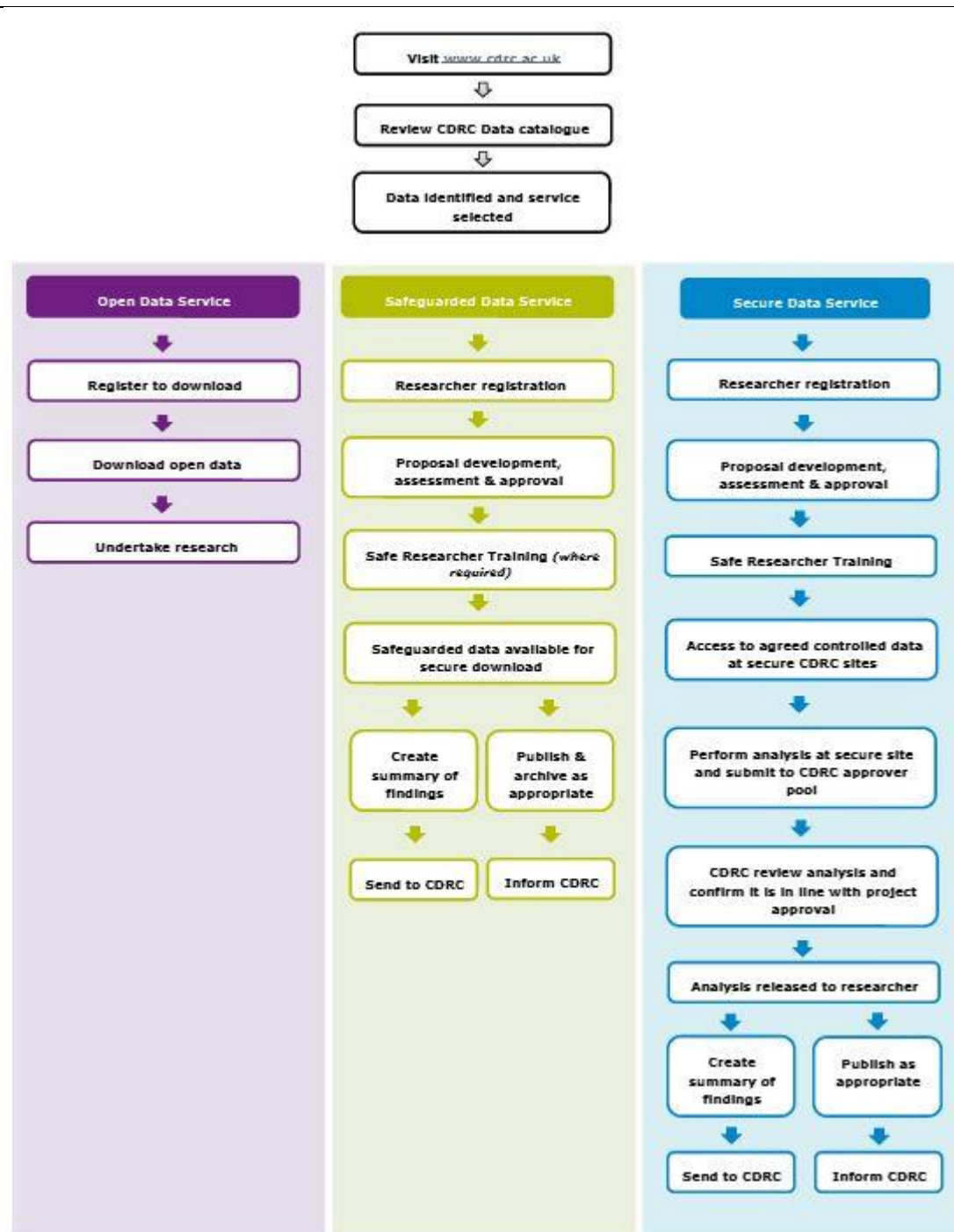
In addition to providing data services, the CDRC has a range of training courses and materials available. Many of these will be of benefit to those who wish to use our facilities, as they are aimed at enhancing capacity in data analytics and data visualisation methods. Full details of the training available can be found at <https://www.cdrc.ac.uk/education-and-training/> and online training tutorials at https://data.cdrc.ac.uk/search/field_tags/tutorial-1679. Our programme includes training in the following areas:

- Working on Big Data: introductory courses that explain the growing importance of Big Data; the importance of analytics and protocols; and standards for data management.
- Introductory and advanced courses in data analysis and visualisation, including courses in R.
- Introductory and advanced courses in Geographical Information Systems, including ArcGIS and QGIS.
- Advanced courses in microsimulation and geo-temporal demographics.
- Courses on how insights from Big Data analytics can enhance business.
- Visualisation.

Charges for CDRC Services

While a service will be provided to the academic community and stakeholders free of charge, researchers may need to apply for funding to cover the costs of additional data acquisitions, or be charged for access to certain, licensed software.

CDRC Services Overview and User Journey



CDRC Website: A Single Point of Entry into the CDRC Data Services

The CDRC website, www.cdrc.ac.uk, is designed to provide a single point of entry into our services and these are clearly linked from the homepage.

CDRC Data

Our data portal, CDRC Data, provides a complete listing of data available through the three tiers of the service and enables the dissemination of open data and application for access to safeguarded and controlled data.

Accessing data from CDRC Data data.cdrc.ac.uk Open Service:

Access to the Open Service requires:

1) Registration

Users will be required to provide contact details including a valid email address prior to download. This is to enable the CDRC to monitor the use of the resource. Data will then be available to the user to download for unrestricted use.

Safeguarded Service:

Access to the Safeguarded Service requires that users to obtain formal approval.

1) Initial Proposal

An approach is made to the CDRC by the user through completion of an online form, linked from the data pages. This initial proposal is processed and assessed by the Senior Management Team to see if it fits within the remit of the Centre. If not, the proposal may be referred to another Centre in the Big Data Network. Proposals that do not fit into either of these categories will be turned down at this stage.

2) Proposal Development

If the initial proposal fits within the Centre's remit, the user is supplied with the 'Safeguarded Data Project Proposal Form', and assigned to a CDRC data scientist who can advise on the technical aspects of the formal application. The aim is to co-produce an acceptable project proposal. Proposals will comprise:

1. a) Research motivation and purpose
2. b) Research impact
3. c) Planned outputs
4. d) Research team
5. e) Data requested
6. f) Data linkage
7. g) Duration of access

h) Ethical approval from user's institution¹

3) RAG Assessment and Approval

Once an application has been completed it is considered by the RAG against agreed criteria that are published on our website, <https://data.cdrc.ac.uk/using-our-data-services>. The number of rejected approvals will be minimised through initial interaction with the data scientists. Where approval is withheld, applications are referred back to the user for revision, and clear guidance will be given regarding those areas requiring clarity. If such amendments are agreeable by RAG, approval will be given. If minor,

the user may be asked to make further revisions, however, if issues are still considered to be major the RAG may decide to make a final decision to reject the proposal. Following approval, the user and their institution are required to agree to the CDRC User Agreement, including stipulations made by the Data Partner(s) and RAG.

4) Data Access

Access to a secure download of the agreed data is made available. This process requires that users telephone the CDRC to obtain a further password to unlock the encrypted download files. Once the user has downloaded the encrypted file, they are solely responsible for the data and its analysis.

5) Outputs

Users can use results of their analyses in publications, reports and presentations provided they abide by the terms and conditions with particular reference to the data partner publication terms. There is no screening of outputs by CDRC staff.

6) Completion, Reporting and Acknowledgement

Users are required to deposit copies of working papers, peer-reviewed journal articles, logs of impact and other publications for access with the CDRC site wherever copyright permits. Where this is not possible, full references to research outputs are required for CDRC audit purposes. Please email publications@cdrc.ac.uk when publications are ready for deposit or logging. The commitment to produce specified outputs is normally a condition of the data approval process. The terms of service require that published outputs include an acknowledgement stating: "The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC xxx, ES/L011840/1; ES/L011891/1". The acknowledgement will make further reference to the use of specific datasets according to the wishes and needs of individual data partners. After the project end date is reached, the CDRC will contact the user to confirm the destruction of the data and to document any outputs to date. The CDRC will contact users normally at 6 and 12 months after the project end date to request a log of any further publications or impact logs.

¹ If the user's institution does not have a system for data protection and ethics approval then the CDRC will assist with gaining ethical review if required.

7) Undergraduate and Postgraduate Student Applications

Undergraduate and Masters Students requesting access to data will be required to submit a proposal in the normal way including their academic supervisor as a named applicant.

CDRC Secure Service

Access to CDRC controlled data is via our Secure Service at one of three secure facilities located at University College London, the University of Liverpool and the University of Leeds. Independent analysis of secure data can be undertaken at all of our secure facilities. If users require bespoke guidance and support with analytics, this service is provided at the University of Leeds only.

Use of the CDRC-Secure service requires registration and project approval, with an additional step of booking into one of the secure facilities and meeting any site specific secure facility requirements. The user will be informed of these once the site to be visited has been selected.

Accessing data from CDRC secure sites

Access to this service requires that users obtain formal approval.

1) Initial Proposal

An approach is made to the CDRC by the user through completion of an online form. This initial proposal is processed and assessed by the Senior Management Team to see if it fits within the remit of the Centre. If not, the proposal may be referred to another Centre in the Big Data Network. Proposals that do not fit into either of these categories will be turned down at this stage.

2) Proposal Development

If the initial proposal fits within the Centre's remit, the user is supplied with the 'Controlled Data Project Proposal Form', and assigned to a CDRC data scientist who can advise on the technical aspects of the formal application. The aim is to co-produce an acceptable project proposal. Proposals will comprise:

1. a) Research motivation and purpose
2. b) Research impact
3. c) Planned outputs
4. d) Research team
5. e) Data requested
6. f) Data linkage
7. g) Access requirements
8. h) Ethical approval from user's institution²

3) RAG Assessment and Approval

² If the user's institution does not have a system for data protection and ethics approval then the CDRC will assist with gaining ethical review if required.

Once an application has been co-produced it is considered by the RAG against agreed criteria that are published on our website <https://data.cdrc.ac.uk/using-our-data-services>. The number of rejected approvals will be minimised through initial interaction with the data scientists. Where approval is withheld, applications are referred back to the user for revision, and clear guidance will be given regarding those areas requiring clarity. If such amendments are agreeable by RAG, approval will be given. If minor the user may be asked to make further revisions, however if issues are still considered to be major the RAG may decide to make a final decision to reject the proposal. Following approval, the user and their institution are required to agree to the CDRC User Agreement, including stipulations made by the Data Partner(s) and RAG.

4) Data Access

Following approval, the allocated CDRC data scientist arranges access for the registered user. Dates are booked to use the secure facility at either UCL, University of Liverpool or University of Leeds. Users will receive a document informing them of site specific secure facility requirements and instructions of use.

5) Data Analysis

The user works on the data only within the secure environment. If users wish to combine controlled data with other less sensitive data (open or safeguarded), then it will be necessary to have obtained consent for this from RAG as part of the project proposal. This supporting data will then be made available to the user in the secure facility. The same applies to software required for analysis. CDRC staff provide limited support through the advanced analytics service. At the University of Leeds, a supported analytics service is available which provides the user with bespoke guidance and support in both accessing and analysing data.

6) Outputs

All outputs that the user wants to take out of the secure environment must be vetted and cleared by the CDRC before they can be released. Source data do not leave the secure facility. Users can take results of their analyses for use in publications, reports and presentations provided they abide by the terms of the User Agreement and with particular reference to the data partner publication terms.

After completion of analysis the user informs the data scientist that the analysis is complete and that their files are now ready for vetting. For full details of the output process please see the CDRC site specific 'Secure Lab Data Import/Export Procedures'.

- a) Outputs will be checked by two CDRC data scientists to ensure that they conform to CDRC control criteria.
- i. Outputs requested should be 'finished outputs' i.e. the finished statistical analyses that you intend to present to the public, must be easy to read and interpret and how they are to be used explained and must be non-disclosive.
 - ii. The CDRC team will ensure that the outputs are the same specification as those agreed in the approved project proposal.
 - iii. The user is informed of the outputs vetting outcome within 5 working days and if successful with details about how the data extracts or analysis will be returned to them.
 - iv. Extracts that match approval are transferred by the CDRC team to a secure server from where outputs can be downloaded under the same arrangements as safeguarded data or transferred to the user on an encrypted USB/hard drive.
 - v. Where extracts are deemed not to match the required criteria, the user is informed.
 - i. Where there are issues with a part of the output, if feasible the user will be allowed to revisit the secure facility to rectify the problem.
 - ii. Major transgressions may be permanently deleted and the remaining output

is returned to the CDRC approver pool.

- vi. Once the user has completed all their analysis or their agreed lab access time has

been reached all passes or electronic fobs are returned and access to the secure facility is immediately revoked.

7) Completion, Reporting and Acknowledgement

Users are required to deposit copies of working papers, peer-reviewed journal articles, logs of impact and other publications for access with the CDRC site wherever copyright permits. Where this is not possible, full references to research outputs are required for CDRC audit purposes. Please email publications@cdrc.ac.uk when publications are ready for deposit or logging. The commitment to produce specified outputs is normally a condition of the data approval process. The terms of service require that published outputs include an acknowledgement stating "The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC xxx, ES/L011840/1; ES/L011891/1". The acknowledgement will make further reference to the use of specific datasets according to the wishes and needs of individual data partners. After the project end date is reached, the CDRC will contact the user to confirm the destruction of the data and to document any outputs to date. The CDRC will contact users normally at 6 and 12 months after the project end date to request a log of any further publications or impact logs.

8) Undergraduate and Postgraduate Student Applications

Undergraduate and Masters Students requesting access to data will be required to submit a proposal in the normal way including their academic supervisor as a named applicant.

9) Request for data not currently available through CDRC

It is possible to request access to data variables or datasets not currently available through the CDRC. To submit a request please send us an email to info@cdrc.ac.uk and we will contact you to discuss further.



Appendix 2. List of R packages used, including versions and documentation.

Note: this thesis relied on multiple versions of R software throughout the years, starting at version 3.6.2. At the date of submission, all analysis was running on R version 4.4.0 (v4.4.0, R Core Team, 2024).

Package	Version	Date	Source	Documentation and Citation
AQuadtrees	1.0.4	2023	CRAN	Lagonigro, R., Oller, R. and Carles Martori, J. (2020) ‘AQuadtrees: an R Package for Quadtree Anonymization of Point Data.’, <i>R journal</i> . https://CRAN.R-project.org/package=AQuadtrees
data.table	1.14.8	2023	CRAN	Dowle M, Srinivasan A (2023). <i>data.table: Extension of `data.frame`</i> . R package version 1.14.8. https://CRAN.R-project.org/package=data.table .
dbscan	1.2.0	2024	CRAN	Hahsler, M., Piekenbrock, M. and Doran, D. (2019) ‘dbscan: Fast Density-Based Clustering with R’, <i>Journal of Statistical Software</i> , 91, pp. 1–30. https://doi.org/10.18637/jss.v091.i01 .
dplyr	1.1.2	2023	CRAN	Wickham H, François R, Henry L, Müller K, Vaughan D (2023). <i>dplyr: A Grammar of Data Manipulation</i> . R package version 1.1.2, https://CRAN.R-project.org/package=dplyr
h3	3.7.2	2022	GitHub	Kueth S (2022). <i>h3: R Bindings for H3</i> . Package version 3.7.2, https://github.com/crazycapivara/h3-r
igraph	2.0.3	2006	CRAN	Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, Müller K (2024). <i>igraph: Network Analysis and Visualization in R</i> . doi:10.5281/zenodo.7682609 , R package version 2.0.3, https://CRAN.R-project.org/package=igraph .
ggplot2	3.4.1	2016	CRAN	Wickham H (2016). <i>ggplot2: Elegant Graphics for Data Analysis</i> . Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org .
lubridate	1.9.3	2011	CRAN	Grolemund G, Wickham H (2011). “Dates and Times Made Easy with lubridate.” <i>Journal of Statistical Software</i> , 40(3), 1-25. https://www.jstatsoft.org/v40/i03/ .

ndjson	0.9.0	2022	CRAN	Rudis B, Lohmann N, Bandyopadhyay D, Kettner L, Fultz N, Demeyer M (2022). “Ndjson: Wicked-Fast Streaming ‘JSON’ (‘ndjson’) Reader”. https://CRAN.R-project.org/package=ndjson
parallel	4.2.2	2022	CRAN	R Core Team (2022). <i>R: A Language and Environment for Statistical Computing</i> . R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ .
sf	1.0.12	2023	CRAN	Pebesma E, Bivand R (2023). <i>Spatial Data Science: With applications in R</i> . Chapman and Hall/CRC. https://r-spatial.org/book/ .
sp	1.6.0	2013	CRAN	Bivand RS, Pebesma E, Gomez-Rubio V (2013). <i>Applied spatial data analysis with R, Second edition</i> . Springer, NY. https://asdar-book.org/
terra	1.7.18	2023	CRAN	Hijmans R (2023). <i>terra: Spatial Data Analysis</i> . R package version 1.7-18, https://CRAN.R-project.org/package=terra .
tidyr	1.3.0	2023	CRAN	Wickham H, Vaughan D, Girlich M (2023). <i>tidyr: Tidy Messy Data</i> . R package version 1.3.0, https://CRAN.R-project.org/package=tidyr .

Appendix 3. Primary and Secondary POI classification categories as described by the Ordnance Survey's' AddressBase Premium dataset.

Primary category (Class 1)	Secondary category (Class2)	Description
<i>C</i> (Commercial)	CA	Agricultural
	CB	Ancillary Building
	CC	Community Services
	CE	Education
	CH	Hotel/Motel/Boarding/Guest House
	CI	Industrial
	CL	Leisure
	CM	Medical
	CN	Animal Centre
	CO	Office
	CR	Retail
	CS	Storage Land
	CT	Transport
	CU	Utility
	CX	Emergency/ Rescue service
	CZ	Information
<i>L</i> (Land)	LA	Agricultural (not business enterprise)
	LB	Ancillary Building
	LC	Burial Ground
	LD	Development
	LF	Forestry
	LL	Allotment
	LM	Amenity – Open area not attracting visitors
	LO	Open Space
	LP	Park
	LU	Unused Land
	LW	Water
<i>M</i> (Military)	MA	Army
	MB	Ancillary Building
	MF	Air Force
	MG	Defense Estates
	MN	Navy
<i>O</i> (Other)	OE	Emergency Support
	OG	Agricultural Support Objects
	OH	Historical Site/Object
	OI	Industrial Support
	OO	Ornamental/Cultural Object
	OP	Sport / Leisure Support
	OR	Royal Mail Infrastructure
	OS	Scientific/Observation Support
	OT	Transport Support
<i>P</i> (Parent Shell)	PP	Property Shell
	PS	Street Record
<i>R</i> (Residential)	RB	Ancillary Building
	RC	Car Park Space
	RD	Dwelling
	RG	Garage
	RH	House
	RI	Residential Institution

<i>U</i> (Unclassified)	UC	Awaiting Classification
	UP	Pending Internal Classification
<i>X</i> (Dual use)	X	Dual use
<i>Z</i> (Object of Interest)	ZA	Archaeological Dig Site
	ZM	Monument
	ZS	Stately Home
	ZU	Underground Feature (Natural)
	ZV	Other Underground Feature (Made)
	ZW	Place of Worship