University of Southampton

Faculty of Electronics and Computer Science

Web and Internet Science

Investigating the Properties of OpenStreetMap Provenance Graphs

by

Bernard Roper

ORCID ID: https://orcid.org/0000-0001-5011-846X

Thesis for the degree of Doctor of Philosophy

March 2023

Abstract

The production of geographic data has traditionally been the purview of institutions such as the Ordnance Survey and US Geological Survey. The past three decades have seen a technological revolution brought about by mobile computing resources and the World Wide Web. Ordinary citizens now have the tools to produce geographic information en-masse. OpenStreetMap, one of the world's most important and extensive geographic datasets has arisen out of this Volunteered Geographic Information phenomenon. Freely available to be both used and produced by ordinary people, it represents a paradigm shift which has changed our relationship with geographic information.

The free and open nature of OpenStreetMap has given rise to novel and often mission-critical uses, often among people with little or no interest in traditional quality assurance frameworks. The shift away from authoritative data sources and traditional quality assurance paradigms raises problems for geospatial data consumers who still need to make informed trust judgements. In a milieu characterised by a diverse and dynamic range of use cases, large volumes of data and no established quality assurance paradigms, we need new ways of understanding Volunteered Geographic Information. One of the most difficult components to document all the contributors and their contribution practices, which operate at a scale and diversity not found in traditional science.

Provenance data encodes much of this information and is useful for providing localised data documentation. Provenance Network Analytics is a methodology which has the potential to provide a principled automated means of analysing provenance data at scale. However, it has only been implemented in relatively simple, smaller scale use cases. OpenStreetMap does not explicitly record provenance data. There is also no framework for understanding of the necessary measurement and interpretation strategies

In this thesis we address these issues by providing a novel method of provenance reconstruction which produces a provenance dataset in an interoperable standard. We provide a framework for provenance measurement using metrics which allow the analysis of large volumes of data. Using OpenStreetMap provenance extracted from the Southampton area in the UK, we conduct a descriptive analysis of OpenStreetMap provenance data. The results provide an understanding of the drivers of variation in OpenStreetMap contribution practices. This work repositions VGI provenance as a new and novel form of geographic data which can provide insights into the nature of Volunteered Geographic Information and the human and physical environment it describes.

Keywords: VGI, Provenance, Graph Theory, OpenStreetMap

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Bernard Roper

Title of thesis: Investigating the Properties of OpenStreetMap Provenance Graphs

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signature: ..... Date:

Acknowledgements

First, apologies to the many people I have inevitably omitted to mention. You know who you are and I'm sure I will wake up tomorrow and remember somebody who should be included.

To Age, Dave, and Stefano for their patient mentoring, tireless support and guidance throughout this journey. For believing in me even when at times, I didn't. Thank you for being the best team of supervisors anyone could wish for.

To Nick and Heather for their encouragement, support, and feedback during the progression reviews, and to Heather and Giles for their inciteful comments and feedback during the final examination process.

To the denizens of the WAIS lab on the third floor of Building 32 for their companionship along the way: Callum, Amber, Mark, Sarah, Tyler, Tom, Jacqui, Rani, Belfrit, Miya, Nora, Iman, Zheng, Sami, and Elena. To Sophia for her wise perspectives, sympathetic ear, and excellent baked goods. To all at the Web Science Institute and CDT for their support and community, and to Alison for her sympathetic ear and wise words.

To the student support services at the University of Southampton who helped me cope in difficult times. Particularly Susannah and Kathryn who have helped me navigate many challenges, both personal and professional.

To all at the Ordnance Survey for their support and guidance, especially Jeremy, Izzy, and Nick for their support, perspective and enthusiasm.

To Rose and Rufus, my now grown-up children, who spent many lockdown hours nodding patiently while I told them about my research.

To Crooked Hayes Copse for being a place of tranquillity, and to its denizens for going about their daily business and reminding me that there is a world outside of academia and the affairs of humanity.

I would also like to acknowledge the use of the IRIDIS High-Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

This project was funded by the Ordnance Survey.

Finally a dedication to some good people who didn't emerge from the last three years: Kaz, Keith, Graeme, Steve, and Smut. You are missed, and I wish you could have seen this.

"You are done! Be done! Go eat ice cream!"

Definitions and Abbreviations

ANOVA ..... **A**nalysis **O**f **VA**riance, a statistical procedure to identify differences in the mean of a single variable by an independent grouping variable.

MANOVA ..... **M**ultivariate **A**nalysis **O**f **VA**riance, a statistical procedure to identify differences in the combined means of two or more dependent variables between groups of an independent grouping variable

OA ..... Output Area, a zone used to publish UK census data. Standardised by the number of households and population.

OAC ..... Output Area Classification, specifically referring to the 2011 demographic classification of census output areas

OSM..... OpenStreetMap

OWL..... Web Ontology Language, an RDF-based syntax for defining a vocabulary and set of logical rules used by machines to model and reason over a specific domain of discourse.

SWfMS..... Scientific Workflow Management System, a computer application or suite of applications designed to manage and execute the steps in a computer-based in silico scientific experiment

VGI..... Volunteered Geographic Information is geo-referenced data produced using tools and frameworks available to the general public, without any requirement for formal training or expertise. VGI can sometimes be produced by expert individuals, but these creators are operating as members of the general public and their work uses the same tools and has the same status.

URI..... Uniform Resource Identifier, a unique sequence of characters identifying a physical or logical resource on the World Wide Web. URIs can identify anything including and should provide a means of either retrieving a digital object or a representation of a physical object at its location.

W3C..... The W3C is an international standards body that maintains and develops standards and protocols for the World Wide Web.

XSLT ..... Extensible Stylesheet Language Transformations, an XML-based language for transforming documents in XML or other formats which support XPATH into other formats

Investigating the Properties of OpenStreetMap Provenance Graphs

Chapter 1 Introduction

"The foremost cartographers of the land have prepared this for you; it's a map of the area you will be traversing." <hands Blackadder a blank sheet of paper> ***"…they'll be very grateful if you could just fill it in as you go along."*** – Blackadder II, BBC comedy, 1986.

The creators of this iconic British comedy could not have known that they were describing the way masses of ordinary citizens would eventually produce the world's largest and most extensive GIS dataset. At that time there was no means of connecting that blank sheet of paper to millions of people. The foremost cartographers of the land have also been taken aback by the burgeoning growth of OpenStreetMap (OSM) from a blank sheet of paper in the hands of a group of cycling enthusiasts in 2004. Today, OSM data frequently sees use in mission-critical applications such as disaster relief planning, satellite navigation systems, public transport planning, policing, national mapping agencies. Despite this, it experiences the same question marks over its credibility as other important user generated content such as Wikipedia, and this is a problem academia is still grappling with.

Traditionally, cartographic and geographic data quality assurance works within a traditional academic paradigm in which the data is professionally produced and documented in accordance with institutionally agreed standards. These are designed to serve a limited set of well-understood use cases, and trust is derived from scientific authority and accreditation. User Generated Content such as Wikipedia and OSM, can be produced anonymously by anyone willing to spend 5 minutes creating an account. This data is also freely available to anyone with access to an Internet connection anywhere in the world, with few restrictions as to reuse. This unprecedented level of global access to geographic data has given rise to a fluid and unpredictable set of use cases, which renders traditional scientific paradigms of data quality and trust inadequate. Many users of today's OSM have little knowledge of or use for ISO data quality parameters. Traditional geographic data quality standards arose to cater for the needs of geographers in Western Europe and North America. Today's OSM is potentially available to users in sub-Saharan Africa who have their own notions of trust and data quality.

Despite this, researchers have studied data quality in OSM extensively and found that although variable, its quality is comparable to offerings from national mapping agencies. A major drawback though, is that much of this work depends upon comparison with authoritative reference data from national mapping agencies such as the Ordnance Survey and Institut Géographique

National. In many parts of the world such data is either not available or prohibitively expensive. Even more seriously, this limitation disproportionately affects the developing world, where people arguably have the most to gain from access to freely available GIS data. It is a fact which has led researchers to look for other ways of understanding and documenting data quality in OSM.

1.1 Background

Maps have been important content from the earliest days of The Web. At the inaugural International World Wide Web conference in May 1994, the Xerox Parc Map Viewer won awards for best use of interaction and technical merit [1]. This application produced and served static maps rendered on a powerful Sun Workstation at Xerox Parc [2], using data from the CIA World Databank II [3] and the 1:2,000,000-scale digital line graph from the US Geological Survey [4]. The emergence of MapQuest from the cartographic division of US publishing company RR Donnelly, based on their popular series of roadmaps, cemented the importance of online mapping. Its success turned MapQuest into a verb in much the same way as Google more recently [5]. These early maps benefited from professionally produced and accredited datasets and a track record of trusted publishing. Public trust judgements therefore relied on institutional quality assurance and the massive technical investment and oversight available to a large corporation.

When web designer Darcy DiNucci first coined the term “Web 2.0” in 1999 [6], the dominant paradigm of the World Wide Web was one of publication. Those with the resources published their content on a “read-only “web. DiNucci’s focus was on publication tools: web standards, interfaces, and new media. A few years later the term entered popular usage after the series of Web 2.0 conferences [7] shifting the focus to user participation, facilitated by the emergence of frameworks such as AJAX [8] which allowed web users to create and update content in real-time using a web browser. This was a dramatic paradigm shift from a web of publication to one of participation.

1.1.1 Produsers and Produsage

By 2005 the World Wide Web had entered a state of flux. Wikipedia had become the most popular reference website on the Internet with nearly a million user generated articles [9]. A combination of citizen journalism, the “blogosphere” and social media began to dominate the news and political discourse. By January 2021, there were an estimated 600 million blogs [10], with 6-8 million posts created each day [11]. In the UK, in 2019, YouTube alone accounted for 12% of all video viewed by adults [12], comparable to mainstream TV channels. Web-based Citizen science projects made it possible to classify galaxies, develop open-source protocols for insulin production, map air pollution, and create significant GIS data in a reconfiguration of science from a closed to an open

activity [13], and the newly founded OSM project would eventually become one of the world's most important GIS datasets.

This rapid rise of user generated content transformed the web from a medium of publication to one of participation, blurring the boundaries between producer and consumer. Our relationship with the web was profoundly changed as a result. In 2006, Bruns [14] described this shift as the *produsage paradigm*. He identified these defining factors:

Production by a community of "produsers": paradoxically, one of the defining features of produsers is that they are an ill-defined group, potentially consisting of anybody with an Internet connection. Bruns identifies this as an essential characteristic and strength [15]. Their numbers and heterogeneity give them an advantage over institutional producers.

Creation outside of professional realms [16]: producers are not part of any academic or professional discipline or institution and do not use any associated frameworks. Naab and Sehl [17] illustrate this by contrasting participatory journalism with citizen journalism. In the former, institutions such as the BBC seek contributions and participation in BBC editorial processes and production, whereas in citizen journalism all activity is entirely carried out by amateurs with no reference to any institution.

Iterative, evolutionary and palimpsestic development [14]: most produsage involves engaging with existing content, either revising. overwriting or copying. Wikipedia and OSM typify this and often have their own archives and complex edit histories resulting in a *palimpsest*, i.e. multi-layered content which is repeatedly overwritten and iteratively optimised, resulting in a complex editing history.

Collaborative engagement: produsage is rarely the sole work of individual users. Most platforms are designed for collaboration, e.g. Wikipedia discussion pages and OSM notes. Reuse of other content is one of the defining features of the produser such that intellectual property rights have also been transformed by this phenomenon [10]. This collaborative engagement, in which "many eyes make bugs shallow" [18] with a continuously improving iterative editing process, is a major strength of produsage as a mode for content creation.

1.1.2 Neogeography

The produsage paradigm has made its effect felt in geography and GIS and has given rise to what many see as a new subdiscipline. Neogeography is a term which has been in use since the early twentieth century to describe new approaches to the study of geography. The contemporary usage

dates to around 2006 [19], and is now widely understood to describe a blurring of the distinction between the expert "academic geography" and the geographic data and information produced by amateurs, i.e. produsaged geography. Its rise has been precipitated by a decline in the entry costs of geographic information production. Web 2.0, affordable GPS in smartphones, open-source GIS software and the publishing of open standards by the Open Geospatial Consortium (OGC) mean that anybody with a passing interest can publish geo-referenced information either by annotating content such as photographs and social media posts, uploading GPS traces and contributing to OSM.

Neogeography and the Democratisation of GIS. Neogeography, like produsage and Web 2.0, has been hailed as a democratising force, with potential to empower the lives of ordinary people. For example, neogeography and participatory mapping processes were crucial tools for the Arab spring democracy activists who used tools such as Google Maps and OSM to share "geospatial news" such as the locations of police, shelter, first-aid and other protesters [20]. Turner defines neogeography as being about people "*using and creating their own maps, on their own terms*", and sharing this information to provide context and understanding of place [19]. Goodchild, in [21], also states that neogeography democratises GIS in the same way that the PC democratised computing: "*It's like the effect of the personal computer in the 1970s, where previously there was quite an élite population of computer users. Just as the PC democratised computing, so systems like Google Earth will democratise GIS*." There was even a series of "Where 2.0" conferences starting in 2007, hot on the heels of the Web 2.0 series, with themes such as "web-based mapping is just the start - - what other geospatial barriers need to come down?" [22]. Literature dealing with user generated content is full of examples of its democratising potential but there are relatively few attempts to define precisely what is meant by democratisation in GIS.

Haklay defines it as: ***a process which makes geographic information technologies more accessible to excluded or marginalised groups in a way which helps them change their lives and the world around them*** [23]. This type of democratisation is about participation at all levels, including decision-making, and equality in terms of human rights and access to resources and opportunities, and Haklay prefers this as a route to evaluating the practice of neogeography to understand its limitations and enhance the potential for democratisation .

1.1.3 Critical Cartography

On the surface it seems that OSM, which provides anybody with web access with the tools to map their surroundings for inclusion in one of the world's most important GIS datasets, has largely achieved the aims of democratisation. Everyone can now benefit from maps and geospatial data made by and for them, which is the popular view of neogeography. Unfortunately, this hinges

on the idea that a map is simply a 2D representation of the earth surface, a scaled-down mirror of the world. This way of thinking has been challenged by a critical treatment of cartography which has its origins in the seminal essay written by Harley in 1989 [24] .

Drawing on the philosophy of Foucault and Derrida, Harley deconstructs the map and treats it as a textual object with an *implicit discourse* and meanings through which power is expressed and enacted. Foucault characterises a discourse as a set of rules governing the identification of truth and falsehood. This governs how textual objects are constructed, how we talk about them and what happens when we change them. These rules all exist for maps and GIS, which, Harley describes as *discursive objects* or *texts,* an idea which goes beyond literary media to encompass informational objects which are consciously constructed and systematically employ signs and signifiers bound by conventions. When we think of maps in this way, as culturally constructed texts, we can challenge their neutrality and employ frameworks for deconstructing them. We can for example, identify literary devices such as metaphor, which helps us to spot otherwise hidden meanings and statements of political authority and control, which also begs the question of who creates OSM and why.

OSM. OSM is by far the largest VGI dataset and is often seen as the "poster child" of neogeography. It is often lauded as a democratising success story of the open-source software movement. "*Free as in 'free speech' not as in 'free beer'*" [25]. Although the OSM map is created in true Web 2.0/produsage fashion, following a collaborative and participatory mode of production, it also follows the pattern of many crowdsourced projects with most data being created by a minority of users. 38% having done at least one edit and only around 5% making substantial changes to the map [26]. These are the people in OSM who exert the most influence and power, and they are not representative sample of those who stand to benefit from free geographic data. Most are European, male [27], [28], urban dwelling [29], wealthy and university educated [26], [29], [30] and many also have some level of cartographic technical skills [31]. Mechanisms of exclusion have been identified based on demographic and socio-political factors which are responsible for biases in the data [27], [28], [32]–[34].

1.1.4 OSM, Trust and Credibility

OSM is as culturally constructed as any other map, and the extent to which the interests of a given individual are represented in produsaged map content depends on who and where that individual is. Whilst a national mapping agency also has its own set of interests, the mode of map production is more transparent, documented, and standardised, as are the mechanisms and use-cases by and for which it is trusted. A major difficulty with OSM and other neogeography is that to

be truly democratising, it needs to incorporate notions of quality and trust which do not just reflect the interests of educated wealthy white European men.

The produsage phenomenon and breakdown of institutional paradigms of trust and quality assurance have resulted in scepticism and concerns about the veracity of produsaged content. These concerns have led commentators such as Keen [35] to decry the produsage phenomenon as a destructive "cult of the amateur". His polemic accuses produsers and Web 2.0 technology of destroying Western culture and democracy. His view is simplistic and anecdote based, and many of his criticisms have turned out to be unfounded, particularly those of Wikipedia which has since proved to be a reliable knowledge source comparable to institutions such as the Encyclopaedia Britannica [36]. However, some concerns are legitimate and raise questions as to how we live with and trust user generated content in the absence of institutional gatekeepers and traditional modes of quality assurance. Concerns over the quality of OSM content are well documented and discussed in Chapter 2, Section 2.3. OSM coverage, like Wikipedia, has been shown to be of a generally high standard and comparable to many commercial offerings, although various assessments of OSM quality have shown a high degree of variability.

Traditional Quality Assurance. Evidence of cartographic quality assurance can be found dating back to the nineteenth century and beyond e.g. in the Ordnance Survey Instructions to Field Examiners [37] originally dating from 1884. Early modern maps and geographic data tended to have a highly specific use cases and generally focused on single quality parameters, usually spatial accuracy. For example, compliance with the US national map accuracy standard (NMAS) from the 1940s required 90% of all the points tested to fall within an 8 mm threshold [38]. For the small number of highly specific, well understood use cases these maps were produced for, this standard was sufficient. As technology evolved, and use cases became more complex, so did quality standards. The current ISO standard for geographic data quality [39] uses six canonical quality dimensions originally described by Veregin in 1999 [40]. These are discussed in more detail in Chapter 2, Section 2.3.1

There are numerous studies investigating VGI and particularly OSM data quality in terms of these parameters [41]–[47], most of which depend on comparisons using authoritative reference data. In many parts of the world, no such reference data exists or is prohibitively expensive. Even where it is available, the parameters are designed with specific use cases in mind, to be understood by professional, academic geographers. The produsers of OSM have a much wider range use cases, many of which are unknown, and often have little or no interest in ISO data quality parameters.

Democratisation and Trust. Trust is a nebulous and problematic term which has a dizzying array of definitions, mostly in computer science, economics, sociology, and psychology. McKnight et al [48]in their interdisciplinary review of the study of trust note that it has been defined as both a noun and verb, as a personality trait, a belief, social structure and a behavioural intention. Its nebulous nature and variety of highly specific meanings is a good reason to avoid in-depth discussion and use of the concept in this work. Where we use the term, we mean: *"the extent to which something can be relied upon to behave in an expected manner"*

Credibility. Credibility theory is a potentially useful tool for understanding the critical assessment of produsaged web content [49], and as a potential evaluation tool for OSM and VGI, has been discussed by Flanagin and Metzger [50]. Credibility was originally described by Hovland in the 1950s and is broadly described as *"the believability of a source or message*". This seems vague, but it is important to understand that credibility is not a trust or quality metric. It is best seen as an epistemological framework to understand how people evaluate and trust information. There are thought to be two dimensions: trustworthiness, a subjective judgement about the source of the data, and expertise, primarily an objective judgement about the characteristics of the data or source [51]. It has been viewed in terms of accepted standards as an objective measure of quality: credibility-as-accuracy; or particularly in the social sciences, as a subjective *perceptual variable*: credibility-as-perception, which depends on the viewpoint of the user.

Credibility-as-perception considers the methods and heuristics which people use to evaluate information for trust judgements. Studies of credibility often focus on dimensions of information literacy and seek to understand the heuristics used by content consumers as they evaluate it. Metzger[52] identifies 5 dimensions: *accuracy*, the degree to which content is free from visible errors; *authority*, authorship credentials and affiliations; *objectivity*, assessing author intent; *currency*, whether the information is up-to-date; and *coverage*, the completeness or depth of information. These ideas provide a framework for a bottom-up assessment of information quality, which can be achieved by the collective effort of a large and disaggregated community such as the OSM contributor base, and is more likely to reflect their interests [50]. Traditional parameter-based assessments of quality (credibility-as-accuracy) still have a role to play, in assessing some of the credibility dimensions [50], [52] but assessments of authority and currency require access to authorship details. They are key components of credibility assessment [51]and in the case of palimpsestic produsaged content, where authorship and reputation are derived from multiple authors and tools in a complex editing history, analysis provenance becomes crucial [53].

Provenance is an active research area discussed in Chapter 2, Section 2.4, and one of the central problems faced by researchers is that provenance data can be highly complex and is often larger than the original dataset.

In OSM there is no explicit provenance assessment functionality beyond a browser view of the version history of individual features. There exists an XML dump containing a detailed edit history but both views of provenance are too cumbersome to provide a useful heuristic credibility assessment, so an automated provenance analytics functionality would be a valuable resource for the credibility assessment of OSM data. Provenance network analytics techniques have already been explored for crowdsourced geographic information [54] where they have been found useful for predicting user trust judgements obtained via the crowdsourcing platform. However, in this setting provenance had been explicitly recorded and the geographic information creation pipeline was in a more limited and much simpler setting. For OSM we do not have crowdsourced trust judgement data, and the scale and diversity of data and large, anonymous, and disaggregated contributor base combined with the lack of a principled provenance graph measurement framework presents a barrier to provenance network analytics in OSM

1.1.5 Summary

OSM is part of a paradigm shift in content creation on the web, as well as in geography and GIS. As a result, geographic data and the means of its creation are now available to a wider and more diverse range of people than ever before. GIS has also moved from a limited range of well understood use cases to a wider range of often poorly understood uses, with new ones emerging as more people become able to benefit from free geographic data. As a result, notions of trust and fitness for use have been disrupted as has our understanding of the modes of creation and interests of the users and contributors.

Credibility theory has emerged as a useful epistemology for thinking about trust and data quality in in produsage and VGI, and for enabling the bottom-up assessment of OSM data by the contributor community. The subjective judgements which build credibility require a principled understanding of the provenance and origins of the data, but its scale and complexity require automated analytic methods of provenance measurement which can display and document meaningful information derived from provenance graphs.

1.2 Motivation

The burgeoning range of new use cases for geographic data brought about by the move away from traditional academic paradigms is shifting the focus of data quality documentation from

western academic institutions and empirical analysis to a much more subjective, global, user-centred view. The shift towards notions of fitness for use and the mechanisms by which people trust map coverage, together with the sheer volume of map data available requires new approaches to analytics.

Huynh et al [55] and Ebden et al [56] proposed Provenance Network Analytics as a principled automated means of predicting user trust ratings using machine learning and graph theoretic metrics from provenance graphs. Their original work used a crowdsourced mapping application, CollabMap, a local scale application much less extensive than OSM. It was designed for a single use: escape route planning in the event of industrial accident from a nearby oil refinery. The producers were paid volunteers from an online crowdsourcing application which used a *find fix verify* pattern to provide quality assurance in which some users mapped escape routes and others rated their work. Huynh et al successfully used provenance network metrics to train a decision tree classifier to predict this user rating. Even more interestingly, they were able to gain insights about the data production using the decision tree output. To replicate this type of analysis in a global big data application such as OSM, we would need a meaningful and useful data quality/trust feature as a target for predictive analytics. Trying to identify such a feature using traditional data quality parameters raises methodological issues.

- It would restrict such work to areas with easy access to authoritative reference data.
- It would restrict the value of any findings to those communities for whom Western European and North American data quality standards were devised.
- It may work well in areas where the value of the feature is already known but would not necessarily generalise to other parts of the world.

Credibility theory offers a more useful global trust dimension which is more suited to global user generated content on the web, providing an understanding of trust as a *perceptual variable* incorporating the viewpoint of the user/consumer. A crucial parameter of credibility is the origins of content and the circumstances of its production, i.e. its provenance. Many OSM users are anonymous, and although some of the more prolific users who do much of the editing can often be contacted via forums and profile pages, the sheer volume of the OSM project and effort involved make the processing of provenance data for even a modest area of OSM coverage impractical. Provenance is an active area of research and frameworks exist for encoding provenance data, but unfortunately such data can often be many times the size of the original data, such that scalability is an important and often intractable problem for researchers. The value and nature of provenance data therefore suggests a need for principled automated way of understanding and quantifying it, including methods of spatial representation so that provenance can be displayed on a map.

In the light of this understanding, the focus of this research has shifted into a study of network analytics for provenance data using properties of provenance as a network graph. We lay the groundwork for further predictive provenance analytics by identifying ways of capturing provenance from OSM edit history data and building on this to develop an epistemological framework for understanding the measurements of provenance in both the domain specific and domain agnostic way. We then use this framework to identify crucial insights which can be gained from automated provenance analytics, examining network properties and spatial distributions using visual inspection and statistical methods to identify latent variables which cannot be measured directly as well as themes and patterns which drive variation in provenance graphs.

1.3 Research Questions

1.3.1 RQ1: Which approaches to the measurement of a provenance graph produce useful insights into the nature of VGI/UGC/OpenStreetMap ?

Contributions. Provenance has been identified as a valuable source of information for data analytics but often does not scale well. Analytics using the network properties of provenance data has been proposed as a vehicle for principled automated analysis and the use of graph data for analytics has been gaining traction in scientific investigations. The network properties of graphs have been of interest in education, neuroscience, other fields, for some time and such investigation of specific provenance graphs has been pioneered by Ebden et al[56] and Huynh et al [55]. We take a wider view of provenance information in User Generated Content, drawing on insights from quality assurance in both Wikipedia and OSM to resolve the issues of scaling this kind of analysis from a small crowd sourced application to global scale datasets. We have conceived a framework for describing graph analytics in terms of abstract and concrete graph measurement, providing methodological approaches which advance the field of graph data analytics.

1.3.2 RQ2: What insights can be demonstrated about map contribution behaviour and the mapped environment using provenance from OpenStreetMap?

Contributions. OSM is the world's largest and most extensive geospatial dataset, with an ever-increasing range of important and mission-critical use cases. Despite this, concerns about its quality and trustworthiness remain. User generated content from global projects such as Wikipedia and OSM represent a fundamental shift in data quality/trust paradigms, and one for which traditional quality/authority frameworks are no longer sufficient. These developments require new ways of

thinking about trust judgements which require an understanding of the way in which citizens interact with their surroundings in order to produce geographic data .

These editing patterns are themselves a type of geographic data, and we can show spatial variations which behave in a similar manner to other forms of geographic data. Although other studies investigate the provenance of geospatial data, we investigate the spatial properties of that provenance, treating it as geospatial data in its own right. Our approach is a descriptive analysis offering a framework that can enable the design of a principled, automated method of deriving insights into the way VGI is created at a spatial level, also providing insights into physical and geodemographic attributes of the area being mapped. This raises the possibility of using VGI provenance as a form of "remote sensing".

1.3.3 Other Contributions

This work successfully implements a data pipeline allowing provenance information to be practically extracted from VGI/UGC/OpenStreetMap, creating structured provenance data in an interoperable standard. In 2014, the Open Geospatial Consortium (OGC) produced the OGC Testbed-10 report [57], which recognised a need for provenance information that respects an internationally agreed interoperable standard, and recommended that the W3C's PROV-DM should be that standard. Provenance as a means of investigating trust, quality and other issues in OSM has interested several researchers e.g. [58]–[62] . Some of the studies have attempted to extract provenance information from edit history but none have fully addressed interoperability issues. Some have considered provenance standards such as the Open Provenance model and W3C which they have based their own provenance data models on but in each case the resulting provenance data is specific to that study and would not be suitable for the other studies without extensive modification. The dataset produced in our work respects the PROV-O ontology, and the W3C's PROV-DM and RDF standards. As such it could be used in all the studies mentioned with minimal effort. The data pipeline we have developed captures data for the Southampton area. It could also be used to produce a similar dataset for OSM anywhere in the world.

1.3.4 Structure of This Document

Chapter 2. In Chapter 2 we conduct a literature review, examining research into the nature of OpenStreetMap data and the circumstances of its production. We examine ways in which the data has been evaluated in order to understand and document its fitness for use. We identify the role of data provenance and look at the practical issues in encoding using provenance data for research.

Chapter 3. In Chapter 3 we define approaches to provenance measurement and analysis, defining measurement strategies as a framework for evaluating OpenStreetMap provenance. We outline three investigations to provide a descriptive analysis of OpenStreetMap provenance. We describe investigations of visualised provenance graphs, thematic maps visualising provenance graph measurements, and statistical analysis to investigate relationships with data quality proxies, discover latent variables using factor analysis, and physical and demographic variations using a MANOVA procedure to differentiate demographic output area classification groupings.

Chapter 4. Chapter 4 outlines the implementation of a novel data pipeline which extracts W3C PROV-DM RDF data from OpenStreetMap edit histories and generates a provenance graph data using geographic extracts based on UK census output area geometry.

Chapter 5. In Chapter 5, we evaluate individual provenance graphs, captured using census output area geometry alongside their OpenStreetMap coverage and associated metadata. Using detailed inspection and visualisation we interpret the graphs to explore drivers of their variation.

Chapter 6. In Chapter 6 we use thematic maps to explore the spatial properties of provenance data and their relationship to the physical environment. We also use Ordnance Survey MasterMap data to calculate measurements of the physical and built environment. We explore the relationship between these physical and environmental properties and the variance of our provenance metrics.

Chapter 7. Chapter 7 is divided into three statistical investigations of provenance metrics data. In Section 7.2, we derive summary, proxy measures of map quality/maturity using visual survey based on comparison with satellite data, and from the output of an automated error detection engine. We assess the extent to which our maturity metrics relates to the real-world notion of map maturity. In Section 7.5 we carry out an exploratory factor analysis to understand latent variables representing themes in OpenStreetMap contribution practices. In Section 7.7 we use 2011 census output area classification supergroups to perform a MANOVA to identify differences between output area supergroups and use post-hoc discriminant function analysis to understand the themes which differentiate these demographic groupings.

Chapter 8. In Chapter 8 we evaluate the findings in the previous three chapters, outlining the themes which have been uncovered. We also make recommendations for future work.

Chapter 2 Literature Review

2.1 Introduction

In this chapter, we explore the literature and research dealing with OpenStreetMap contribution patterns. We examine research into OpenStreetMap contribution which profiles participation patterns, OpenStreetMap contributors and the techniques behaviours and practices by which they have created what has become one of the most important GIS datasets. We examine how these patterns relate to data quality and trust and review research into OpenStreetMap data quality.

The heterogeneity which characterises OpenStreetMap data is a characteristic of user generated content, which exists outside the more uniform, centralised scientific paradigms of scientific data generation. We examine modes of quality assessment using more traditional methods of comparison with authoritative reference data, and the intrinsic methods of assessment which have emerged for OpenStreetMap as a response to patchy availability of authoritative reference data. We explore how the scientific community has wrestled with the problem of quality assurance in OpenStreetMap which exists outside traditional scientific paradigms. This increases the importance of metadata and provenance data as tools for understanding and profiling OpenStreetMap contribution. In the second part of the chapter (Section 2.4) we review research into the study and use of provenance, examining the insights which can be gained from methods of studying it and evaluating the practical considerations of obtaining and using provenance data.

2.2 OSM, a World of Maps and mappers

The OSM contributor base have been the subject of research to understand how the dataset is produced and who produces it. Important areas of concern include:

- participation patterns: studying the motivation and dynamics of contributor engagement.
- contributors: looking at the demographics and the political and societal contexts of the contributor population.
- mapping behaviour: identifying characteristic patterns of behaviour among OSM contributors

2.2.1 Participation patterns

OpenStreetMap production has been shown to follow the characteristic 90:9:1 rule identified by Nielsen [63], that the amount of contributions to online user generated content tend to follow a *Zipf curve,* an exponential distribution in which approximately 1% of the user base

contribute most of the content; 9% of users contribute occasionally; and 90% of users are 'lurkers' who consume content but do not contribute. This phenomenon, also referred to as *participation inequality*, has been widely identified across the web and OpenStreetMap is no exception. Neis and Zipf [26] found that only 38% registered members carried out at least one edit and only 5% carried out significant sustained editing. These findings have been confirmed in more recent studies of OSM [33], [64], [65]. The significance of participation inequality was recognised by Nielsen in his original article, where he points out that the more highly contributing sections of the contributor base are not representative: *"**on any given user participation site, you almost always hear from the same 1% of users, who almost certainly differ from the 90% you never hear from**."* – Jacob Nielsen, 2006 [63]. In OSM for example, participation ratios and levels have been found to vary by gender [28], [30], [66], [67], nationality and ethnicity [34], [65] and socio-economic status [33]

Participation Bias. Haklay [33] explores this contribution profile in detail and regards it as one of the most significant features of both VGI and citizen science. He considers the time typically required to perform an edit in OpenStreetMap and looking at the editing volumes of prolific contributors, concludes that they devote a significant amount of time to OSM. Considering that men in well-paid occupations and people without major caring responsibilities typically have more leisure time, this is likely to cause a serious bias issue in OSM. Although he points out that this phenomenon is not uniform across all citizen science projects, other studies confirm OSM's male bias [28], [65]–[67].

Budhathoki et al. [30] and Gardner et al. [66] both investigated gender bias and gender-based participation inequalities in OSM using the results of a survey of active users, and were able to demonstrate that the OSM user base is heavily male dominated. Both studies were based on respondents to surveys of OSM users. Budhathoki identified 33,000 OSM contributors from OSM's internal messaging system and obtained a sample of 444 users who responded to invitations to a web monkey survey. Gardner et al used a similar method and obtained their sample by inviting OSM forum users. Although both studies use a non-random sample, selecting a small, English-speaking subset of OSM community, they used the 90:9:1 rule to check that they had a representative sample. Analysis of the contribution profile of their samples broadly mirrored the familiar pattern.

The effect of gender bias in OpenStreetMap has been described by Gardner et al. [66] who found different editing patterns between genders with women mapping more buildings and more new features and men mapping more highways and being more likely to edit existing features. Stevens [28] looks at the effect of "male gatekeeping" on OSM tags. The tags which add semantic meaning to features are created and then voted for by OSM users to become part of OSM's

convention-based tag model. She noticed a richer availability of tags for places of adult entertainment such as bars, strip clubs, brothels and swinger clubs when compared to tags available for childcare facilities. In 2011 there was only one option, “kindergarten” and several suggestions for useful delineations, such as “creche”, “playgroup”, and “nursery” voted down by male contributors.

Geopolitical factors have also been documented, which can provide distinctive motivations for contributors as well as mechanisms for exclusion. Quattrone [68] studied osm:node production across 43 countries, measuring the correlation of editing activity with national socio-cultural variables such as individualism vs collectivism; *Power Distance Index*, a measure of how concentrated decision-making power is; *Pace of Life Index*; and *Self-Expression Index.* They found these variables correlated with editing activity as did economic affluence. Bittner [34] examined socio-political aspects in their study of OpenStreetMap in Israel/Palestine which examined the reasons why most of the OSM coverage of the Gaza Strip has been produced by Jewish Israeli mappers. OSM uses a ground truth paradigm [69], which encourages contributors to produce a literal representation of features as they appear on the ground in such a way that they would be reproduced by another mapper visiting the locality. In Gaza this subjective view often represents a contested cartography. Features such as ruined farms, border fences and highways which are accessible to one community or another, represent a painful and contentious reality which acts to exclude Palestinian Arab contributors.

Participation Categories. Several studies have considered some type of classification or hierarchy of contributors, frequently based on judgements of seniority and expertise derived from activity levels or length of time active. Coleman et al [70] were one of the first to propose categories of VGI users. They drew on other related studies profiling UGC and produsage contributors/ contributions and extended various generic classifications to derive a similar profile for VGI. They used five categories reflecting the frequency and sophistication of contributions: neophyte, interested amateur, expert amateur, expert professional, and expert authority. Neis and Zipf [26] based their three categories on contributor activity level: senior mappers, responsible for more than 1000 osm:nodes; junior mappers, responsible for 10 – 1000 osm:nodes; and non-recurring mappers, who created less than 10 osm:nodes. Budhathoki and Haythornthwate [71] also used activity levels to define *serious* and *casual* mappers by measuring time active, osm:nodes created, and number of editing days. Bégin et al [64] studied the lifecycle of OSM contributors and considered six classes and an additional ‘lurker’ class of non-contributing users. They plotted a complementary cumulative distribution function for the contributor account lifetime and derived their classifications from the inflection points of the curve. They found an abrupt break in the curve at one hour, the point at which an idle changeset closes, representing the 15% of participants who stopped contributing after

minutes or seconds. The rate of cessation remains high for an hour and gradually slows until about 24 hours at which point, 60% of contributors have ceased. Fewer than 20% remain active beyond one year and less than 1% after five years.

Other studies use specific mapping practices to categorise users. Yang [65] used two categories: “professional” and “amateur”, i.e. a skilled, motivated user vs an amateur with less experience and dedication. The categorisation assessed edit intensity and software usage to estimate the proficiency of contributors. Mooney and Corcoran[72] defined a set of *edit actions,* i.e. operations which result in a new version of a feature. Types include *creation*; *created edit*, where a contributor creates a feature and then edits it; *node self*, performing an edit on the osm:nodes of a feature they created; *node*, an edit to a feature not created by the contributor; *tag self,* an edit to a tag of a feature they created; and *tag,* an edit to the feature not created by the contributor. Using K-means clustering of these action counts, and a naïve Bayes classifier, they derived four clusters: creators, geometry editors, taggers and geometry and tagging editors. They found that most prolific editors predominantly either edit geometry or edit tags. Steinman et al [73] also used K-means clustering to derive a set of 10 contribution profiles looking at contributor action types, feature types, total number of contributed actions and length of time active. Their results show that the groups of users they identified could be delineated by the extent to which they edited different feature types.

2.2.2 OSM Contributors

Understanding the OSM Crowd. Social scientists such as Yu-Wei Lin have examined the life of contributors using Social Worlds Theory [74]. Using a series of semi-structured interviews she showed how actors from the business, governmental and NGO sector work alongside more loosely connected individuals to shape OSM. Her work reveals a complex milieu of commercial and humanitarian interests and individuals with emotional attachments to their locales and their technologies. Wen Lin [75] carried out semi-structured interviews on contributors involved in mapping Newcastle upon Tyne. The respondents confirm other findings in that they were predominantly male, and university educated, with a wide range of occupations and some degree of comfort with technology. There was also a familiar longtail profile, i.e. a sharp disparity in the amount of edits carried out. Both studies allude to the influence of external bodies, many participants have links to local government agencies, the private sector and organised humanitarian mapping initiatives.

Our understanding of the demographic profile of OSM contributors is limited to studies of prolific users who have responded to requests for interview. Knowledge of the “long tail” of less

prolific contributors is more limited. Numerous studies have found, perhaps unsurprisingly, that OSM users are typically male, educated. Western European urban dwellers in the 25 – 35 age group [28], [29], [33], [66], [68], [73]. We have also seen that contributors preferred to map home regions, which tend to be areas they are familiar with. This is likely to be a factor which drives the preferential editing of urban and more affluent locations.

Budhathoki and Haythornthwaite [71] surveyed 443 OSM contributors and provided a detailed account of their motivation and demographics. The profile of their respondents broadly mirrored known proportions of contributors by region as well as Nielsen's participation profiles, although there was a greater proportion of prolific users as is often the case with OSM survey respondents. 64.6% were in the 20 – 40 age group, 61.2% in full-time employment with 72% of those in the commercial sector, 80.2% living in Europe and 78% university educated. Most contributors worked on OpenStreetMap at home rather than work, and most had some involvement with open-source projects such as Wikipedia or software. The respondents were classified as either serious or casual mappers based on a combination of number of nodes contributed, time active and contribution frequency. The authors carried out a detailed questionnaire to assess the motivation of their contributors. The results of factor analysis showed that contributors were motivated by

- Altruism: a perception of the value of the OSM community to wider society.
- Self-efficacy and the desire to improve the skill and knowledge.
- Enjoyment from learning, exploring local knowledge, map aesthetics and computer use.
- Learning and personal development.
- Personal need for map data.
- Anticorporate sentiments, i.e. all map data should be free.

It is worth noting that these motivations primarily apply to the mappers in northern Europe and to some extent North America. There were no respondents from South America, 3.6% from Asia, 1.4% from Africa.

Organised Mapping. From around 2014 – 15 we have seen a marked rise in the role of corporate editing. Anderson [76] has documented this, and in 2018, identified 954 user accounts associated with corporate editing. Organisations such as Apple, Amazon, Facebook, MapBox, Microsoft, Grab, Uber, Geofabrik and Stamen all employ teams of contributors, and some have provided data as bulk imports. Levels of corporate editing are not uniform either geographically or in terms of the data. Corporate editors are active in Europe, North America, and Southeast Asia, but do less editing in the Global South. Independent individuals are still responsible for about 70% of the total features in areas where corporate editing is present, but corporate editors are responsible for the bulk of road

network coverage. E.g. Apple is responsible for 80% of the road network in areas where its editing team is active. Anderson concludes that there is good evidence that this corporate editing provides benefits to OSM by providing access to data and innovation, for example Amazon has been able to use the GPS traces of its delivery drivers to improve turn restriction data. These developments have prompted the OSM foundation to produce organised editing guidelines [77], mandating transparency and cooperation with individuals and local communities.

OSM contributors have also been involved in coordinated humanitarian mapping initiatives, often facilitated by the OSM foundation's Humanitarian OpenStreetMap Team (HOT). These are generally either external mapping initiatives to provide coverage for a disaster struck area such as during the Nepal and Haiti earthquakes and the DRC Ebola epidemic, or local community led initiatives to provide a geodata resource for vulnerable and marginalised communities such as the Ramani Huria [78] and Kibera [79] initiatives.

The external mapping initiatives mobilise the global OSM community to focus on a disaster struck area. Kogan et al [80] carried out a network study of co-editing during the two weeks following the Haiti earthquake along with structured interviews. The mappers involved in the disaster relief mapping are generally individual OSM contributors responding to news coverage as well as the HOT initiatives They exhibit intensive mapping which is highly collaborative with contributors frequently editing each other's data.

The community humanitarian mapping efforts often mobilise local people who are trained to map their surroundings. The HOT staff and volunteers are also largely drawn from these communities. Partnership with international aid agencies, and local and national government bodies has also enabled the use of technology such as UAVs for high-resolution aerial imagery and the issue of GPS devices which result in comprehensive and high-quality map coverage. Whilst there is literature describing [78], [81], and investigating the impact [79] of these projects, there seems to have been little work done on understanding the contributor dynamics and practices of this mode of OSM creation. What we know about OpenStreetMap contributors and their characteristics are almost entirely based on studies in Western Europe which are unlikely to be generalisable.

2.2.3 Mapping Behaviours

Individual behavioural characteristics of users often leave patterns in the data which can tell us about the habits of OSM users. There are spatial, temporal and social aspects to contributor mapping behaviours. Spatial aspects include preferences for editing specific regions and features and other spatial patterns of map development. There are diurnal variations in editing intensity and

other temporal patterns governing edit frequency over an account lifetime, and there are also identifiable patterns in the way contributors interact with each other via the map data. Most the research identifying these patterns are of OSM in northern Europe and may not generalise to other parts of the world.

Spatial and Environmental Behaviours. Spatial preferences among contributors were described by Neis and Zipf [26], who used the geometry of created osm:nodes to delineate *home regions* for OSM contributors, i.e. polygons where the bulk of their edits occur. They found that prolific users had much larger edit regions and were more likely to edit data outside their home country. Zielstra [82] used clustering techniques on the frequency of set of core edit behaviours to delineate *home regions,* and *external regions* a contributor also edited but was less familiar with. They found contributors edited a greater diversity of features in their home regions. Napolitano and Mooney [83] also found contributors had "pet locations". Their study derived these from analysis of editing frequency and these results were then confirmed by semi-structured interviews. They found that the pet locations tended to be areas where contributors spent a lot of their time, either at work or home.

Mappers have also been shown to develop OSM coverage following a similar pattern to real world street network development. This elementary process of *densification and exploration*, was identified by Strano et al [84]. It describes the evolution of road networks whereby an undeveloped area is initially "explored" by creating major roads which then trigger further urban development. Corcoran [85] found this process also existed in OSM for the mapping of street networks. Exploration and densification patterns were also found by Arsanjani et al [86], [87] in their Cellular Automata (CA) modelling of the growth of OpenStreetMap coverage. In these studies, they modelled spatiotemporal processes of OpenStreetMap editing to produce accurate simulations of map development.

Some areas are more attractive to OSM mappers than others which seems to be a major driver of variations in completeness. Antoniou and Schleider [88] investigated levels of editing activity in the London area using Gettis-Ord Gi hotspot analysis. The hotspots they found were centred on popular and well-known areas. Another CA modelling study by Quattrone [89] used deprivation indices as one of the model features and their results imply that more deprived areas are less attractive to mappers. The explanation for this was that mappers tend to be more confident mapping familiar areas and tend to come from more affluent backgrounds. Arsanjani et al's [29] logistic regression analysis also confirms this. They found that highly contributed areas tended to mirror common demographic profiles of OSM contributors: educated, populous, high income,

culturally diverse, with a high rate of overnight stays, dominated by the 18 to 69 age group. They also found that proximity to built-up areas is an indicator of mapping intensity

OSM contributors also have preferences for specific land use and feature types. Arsanjani et al's [86], [87] CA model used these preferences. They found that OSM contributors had a greater propensity to edit man-made surfaces, transportation units, commercial and leisure facilities, and buildings. Natural features had fewer edits which the authors attribute to their lower density, slower rate of change and fewer human interactions. Bégin et al [90] also found that individual contributors often focus on "pet features" and are also more likely to work on other nearby features. They identified and used the feature preferences of prolific contributors to an area to predict variations in completeness. Neis et al [46] also attributed OSM's paucity of turn restriction data to the fact that this feature is less appealing to contributors owing to lack of visibility on the map.

Bégin et al [90] concluded that the interests of individual users directly affect OSM data through pet features and preferential editing. As well as density, rate of change and human involvement, Comber et al's [91] analysis of VGI from the geo wiki project shows how both national culture and expertise/self-efficacy can affect the way features are interpreted and viewed. GeoWiki is a crowdsourced land coverage database, and the authors assessed the way different land cover features interpreted by contributors from different countries with different levels of expertise and found marked differences. Bittner's observations [34] of OSM in Israel/Palestine regarding the meaning of features for different communities, and Gardener et al's [66] findings that men preferentially edit roads and women preferentially edit buildings also suggests that the identity of contributors plays a role in the selection of preferred features and areas and is likely to be a source of spatial and social bias.

Biases are already evident in OSM data. Several studies suggest a relationship between socio economic characteristics of an area. We have already noted Quattrone's use of deprivation indices for modelling completeness and this quality parameter seems to be the most affected with several studies [42] finding links with completeness and deprivation measures.

Temporal Behaviours. OSM users are also known to exhibit diurnal variation in their editing activity. Variations can include "shift" patterns where users edit during distinct time periods. This is a distinctive feature of corporate editing, where Anderson [76] noted a distinctive weekday 9-to-5 signature which varied by time zone. It also occurs in remotely organised humanitarian mapping efforts where Kogan [80] found shift editing at specific times of day, usually evenings, and again, varying by time zone. Temporal patterns can also be seen in editing intensity where Yang [65] found that "senior mappers" edited on consecutive days for longer periods, whereas less experienced

contributors had longer gaps between edit sessions. Wen Lin [75] also noted that many of her contributor survey respondents were likely to be 'lurkers' for some years before beginning to contribute. These respondents were initially put off by the difficulty of uploading GPS tracks, and this trend has reduced since the introduction of aerial imagery lowered the skill requirement.

Social Networking Behaviours. In addition to the spatial and temporal dynamics of mapping behaviour, contributors also exhibit social interaction patterns. Research has examined contributor interactions which occur when one contributor edits the work of another. These implicit "collaborations" have been the subject of network analysis often using graph theory, and analogies with social network analysis are commonplace. Collaboration networks are defined by co-editing behaviours identified in edit history [62] or by definition of OSM specific create, read, update, delete (CRUD) type operations.

Although Lin's ethnographic Social Worlds study [74] concludes that one of OSM's features is that it brings individuals from different social worlds together to coproduce the map, Mooney and Corcoran [72] were among the first to investigate OSM contribution as a social network activity. They identified a set of edit interactions which they used to build a network representation of contributor activity in London, and to categorise contributors into distinct categories by their contribution patterns. They looked at prolific editors and found that in all the cities they examined all prolific editors co-edited. They also found about 40% of contributors did no co-editing at all. The author's work raises the question of what causes these disparities in co-editing behaviour. Later research suggests that one factor is likely to be organised editing. Kogan et al [80] noted in their study of organised humanitarian mapping, that one of its distinctive features was the highly collaborative nature of the work.

Truong et al [62] modelled contributor activity using multiplex collaboration graphs with layers derived from spatiotemporal co-editing i.e. editing at the same time, or the same map location, and specific co-editing behaviours such as *completion edit,* which adds content while leaving the previous edit unchanged and *correction* which removes or changes part of a previous edit. By performing graph clustering analysis on the aggregated layers, she detected communities within these co-editing networks, i.e. groups of heavily connected vertices which are sparsely connected to vertices outside the community. She found some communities were organised around a single prolific user whose work was frequently co-edited. These communities were not studied in depth, but the author surmises different patterns: contributors whose work is frequently reused, and pioneers, who sparsely map larger areas and whose work is then densified by other contributors. Other studies have also looked at co-creation networks in both Wikipedia [92] and

OSM [31], [72], and they are commonly used for assessing contributor reputation for trust judgements.

2.3 Data Quality and Trust in OpenStreetMap

Writing in 2007, Goodchild [93] recognised the issues for VGI based web mapping caused by the shifting paradigms of quality assurance brought about by the rise of Web 2.0 and UGC. Chief among these according to Goodchild is the discrepancy between the "traditional top-down, authoritarian, centrist paradigm" in which professionals produce and amateurs consume, and the chaotic world of VGI. Goodchild called for a framework that could embrace the contributions of 6 billion amateur citizens observers while building similar levels of trust and assurance to national mapping agencies.

A climate of scepticism typified by Keen's 2007 "Cult of the Amateur" polemic [35]and more empirical investigations which identified a heterogeneity in VGI data quality [43], [94]saw OSM become the subject of a large body of research aimed at understanding and assessing its data quality and trustworthiness. Much of the work discussed in the following sections focuses on assessment against ISO standards and their quality parameters. Authoritative Comparison with reference data has been a common methodology, but this presents practical issues for OSM. Other strands of research address these by providing intrinsic methods of assessing data quality.

2.3.1 Parameter Based Quality Assessment.

Of the more traditional empirical approaches to data quality assessment, many rely on parameters originally identified by Veregin [40] and used in the ISO geographic data quality standard ISO 19157. This framework provides 6 parameters which have remained consistent across the various versions of ISO 1915, and which form a basis for most empirical studies in the literature.

Completeness. Completeness refers to the presence or absence of features and relationships and can either relate to excess or missing data. It is one of the most frequently investigated parameters for OSM, which by its very nature is an unfinished product. On a global level Barrington-Leigh et al [95] estimated in 2019 that OSM road coverage was just over 80% complete. On a more local level other studies have found completeness to be highly variable. Studies often focus on road/street networks and building footprints and most rely on comparison with authoritative reference data.

Haklay [43] compared road length computations from OSM and OS Meridian datasets in the UK and also performed visual inspections of tiles of areas in London. In France Girres and Touya [42]

evaluated road network completeness by comparing object density between OSM and BD Topo, a dataset produced by IGN, France's national mapping agency. Other studies have evaluated road network completeness by comparison with proprietary datasets from TelAtlas [47] and from Bing and Google maps [41]. In both studies some OSM data was found to have superior completeness so that in these relative comparisons it was difficult to give absolute statements about the completeness of either dataset.

Building completeness is often compared by building count or building area. Fan [96] and Hecht [97] both use ATKIS (German national mapping agency) data. Fan found OSM had high completeness and building footprints similar to those in the ATKIS data whereas Hecht found much lower levels and Dorn et al [98] found that completeness varied between land-use classes and geographic areas..

Positional accuracy. Positional accuracy is defined as the accuracy of the position of features either in relation to the accepted coordinates from an external authoritative source or for the accepted relative position to other elements in the dataset. Most analysis approaches use reference data and compare accuracy by superimposing layers with OSM data onto reference data. Analysis is then either carried out digitally or visually to arrive at an accuracy measurement.

Haklay's road network study [99] compared positional accuracy of OSM motorways to the OS Meridian dataset, and also looked at a set of tiles taken from the London area to investigate minor roads. They used vector data analysis techniques to measure overlap between roads in OS and OSM and carried out visual inspection of the tiles. They found that OSM was broadly of good quality but that this was not evenly distributed with a sharp drop apparent in rural and more deprived areas when they compared it against ONS deprivation indices. In the Republic of Ireland, Cipeluch et al's Bing and Google maps study [41] also looked at positional accuracy of Points of Interest (POI) and road networks. They converted OSM data into KML (Keyhole Markup Language), which can be overlaid onto Google and Bing maps. In their visual analysis they scored the data based on discrepancies between the datasets, which raised the issue of which to regard as the authoritative data, because the OSM data was often superior. For example, positional accuracy of roads in OSM was better than in Bing, because OSM had more recent updates, reflecting changes to a major road which had not been recorded on Bing or Google maps.

Thematic accuracy. This refers to the correctness of the classification and attributes of data items and in OSM is related to its tagging system which provides semantic meaning and classification to the data primitives which form the basis of OSM's data model (see Chapter 4, Section 4.1.2). Girres and Touya [42] examine attribute accuracy by comparing OSM attribute values with matching

values from the BD TOPO dataset using a Levenshtein distance algorithm. They also looked at specific tags, comparing the “highway” attribute from OSM to the corresponding attribute in BD TOPO to test the semantic correctness of roads classification. Thematic accuracy does not necessarily require comparison with reference data and some major work such as Mooney and Corcoran [100] have assessed it using spelling correctness as a measure of accuracy. Other non-cartographic reference data can also be useful, e.g. Bright et al used alcohol licensing data in the UK to verify attribute accuracy of buildings tagged as licensed premises[101].

Temporal Quality. Defined in the ISO standard as the quality of temporal attributes and relationships of features. There are three elements: accuracy, the closeness of time measurements to accepted values; consistency, the correctness of the ordering of events; and validity, the use of a correct time format [39]. Veregin [40] makes a distinction between temporal accuracy and *currency*, pointing out that temporal accuracy refers to an objective agreement between actual and encoded values whereas currency is an subjective judgement about temporal accuracy. There are numerous examples in the literature and news coverage illustrating OSM’s potential to provide timely geographic data. Fan [96] noted that German national mapping agency data had missing buildings that were present in OpenStreetMap because of its speedier update frequency Humanitarian efforts such as the Haiti and Nepal earthquake initiatives, or the role of OSM in the Arab spring uprisings [102] demonstrate the speed with which OSM users can band together to provide a rapid GIS response to emergencies . In traditional reference data, comparisons, OSM can have superior temporal quality, Cipeluch et al [41] even regarded it as the reference data when temporal quality was compared with Bing and Google maps.

Logical consistency. This is a measure of whether the data is structured according to logical rules. Some of the elements of logical consistency specified within the ISO standard are less applicable to OSM. i.e. adherence to rules of the conceptual schema and adherence of values to the value domains, as the OSM data model is semantics are governed by convention rather than schema. Topological correctness can be assessed without reference data. Instead, investigations check consistency with sets of logical or topological rules, for example in the network graph representations of street networks by Jilani et al [103]. Neis et al [46] also looked at topological errors. They identified a list of errors to test for: junctions not sharing common nodes, duplicate ways/nodes and streets which overlap rather than crossing.

Usability. This parameter differs from the others, in that rather than map characteristics, it is primarily based on user requirements, which emphasise other quality parameters depending on those requirements [39]. There are relatively few studies which specifically address usability of OSM data. What discussion there is, either looks at interface usability or treats it as a synthesis of other

quality parameters, particularly those which can be intrinsically assessed without using reference data[104]. Another aspect of usability was examined by Mooney et al [45], who looked at the use of metadata in tags, i.e. source, attribution and description tags. Although this is also a completeness metric, it is mentioned here as it affects the usability of the data

Elsewhere, in academic GIS, fitness-for-use has also been studied in terms of usability of data sets. Harding addresses its subjective qualities in a study [105] of the usability of geodata. She conducted semi-structured interviews with geodata professionals to understand fitness-for-use from the perspective of user's needs. In her results, GIS application design and interface interaction issues were raised by interviewees more than the quality of the actual data set. She found that the data issues highlighted were broadly in line with ISO guidelines: provenance, currency, positional accuracy, attribute accuracy, logical consistency, and completeness. In other related work [106], published in the same journal issue, Brown et al describes the outcome of a workshop attended by 20 experts in Geospatial information and HCI to identify core issues and research challenges in GI usability: VGI data quality, metadata, standards and user behaviour. They considered a change in focus from application usability to data usability.

The role of stakeholders, traditionally conceived as end users, developers, and data producers, was seen as crucial to understanding GI usability and the group noted that the relatively recent advent of VGI has blurred and altered this stakeholder model. The basic data quality attributes identified by Harding include data provenance, and the group noted that in VGI, traditional methods of obtaining data lineage are problematic. The workshop group identified VGI as one of the key future research challenges, and one that raises issues affecting all the others raised.

2.3.2 Intrinsic assessment.

The data quality assessments discussed so far in this section mostly rely on comparison with authoritative reference data from national mapping agencies from the UK, Republic of Ireland, France, and Germany. In many parts of the developing world, where free geographic data provides the greatest societal benefits, suitable reference data is either absent or prohibitively expensive. Some researchers have responded by looking at methods for predicting the quality of data using the characteristics of the data itself. These ***intrinsic*** methods often focus on events in the lifecycle of the data, such as user behaviours and editing patterns. Most ISO parameters are not amenable to direct intrinsic assessment and rely on internal data characteristics to provide predictive estimate of the parameter measurement. Another approach is to use intrinsic measurements to provide some trustworthiness value as a proxy for data quality parameters.

ISO parameter based intrinsic assessment. Apart from usability , which is a subjective parameter not amenable to internal measurements, logical consistency is the only other parameter that be directly measured without reference data. Intrinsic analysis techniques for the other parameters generally involves using properties of the data to estimate a parameter value or using a score or measurement of some intrinsic data property to provide a trust rating as a proxy for data quality. Examples of the first category include Neis at al's work [46], which evaluated completeness and attribute accuracy by looking at proportions of unnamed streets and roads. Girres and Touya's study [42], also examined the number of contributors per km^2. They found an exponential positive relationship with the number of contributions, which they regarded as an estimation of completeness. They also estimated temporal accuracy using the mean capture date and found that this correlated with number of contributors. Logical consistency was investigated by Jilani et al [103] , who made graph representations of street networks and used their topological characteristics to create feature vectors to train machine learning models. Neis at al [46] also assessed logical consistency by identifying topological errors in road network data. Intrinsic positional accuracy estimates are provided by Mooney et al [45] who used the density and distribution of polygon nodes – high density objects with a small average distance between nodes signifying more accurate polygon representation in land-use cover.

Baron et al [107] provides a comprehensive study of intrinsic analysis with their application framework for intrinsic quality assessment, which generates report on OSM coverage from a user defined area. The reporting is organised around a range of use cases used to categorise results. Within these they hypothesise 25 intrinsically derived quality indicators. The framework uses information extracted from an OSM edit history file, and some of these heuristic measurements use aspects of edit history. Update frequency is used to estimate road network completeness and combined movement patterns in multiple features are used to estimate positional accuracy. They make few absolute predictions of data quality but do provide useful assessments of the development of OSM coverage over time, illustrating the utility of provenance information for intrinsic analysis.

Proxy Methods. Kessler et al analysed OpenStreetMap edit history to produce trust ratings based on editing behaviour[61], [108]. They identified the set of possible edits and created reputation metrics for users based on other edits made to their work and edits made to the neighbouring features. E.g. a deletion of a feature or part thereof signifies a correction or rollback, lowering the user's reputation, whereas an uncorrected edit to a feature surrounded by other often corrected features bolsters users' reputation. They use this and the "many eyes principle" identified by Raymond [18] and described in OSM by Haklay [44] as a means of estimating quality. The many eyes principle

implies that a feature edited by large number of users will tend to be of high quality. Kessler and DeGroot extend these ideas to use provenance data [60] which they extracts from edit history to derive their trust and reputation values. A similar tool was proposed by D'Antonio et al who also defined a set of editing types to produce formal trust calculations based on user reputation[58].

2.3.3 OSM Quality Heterogeneity

All the studies discussed here which report on data quality in OSM have a common theme, which is that of heterogeneity. OSM data is often comparable to other commercial offerings, and in some cases can even rival national mapping agency data in temporal quality [96], but there are sharp variations in the spatial distribution of various measures of OSM data quality. Some studies have found systematic variation, for example Haklay's comparison of road lengths in OSM and OS Meridian datasets in the UK [37] found that when compared with socio-economic UK census data, completeness varied according to deprivation indices and was less complete in deprived areas. Other studies have found relationships with population density [46], [47], [109], and others found distinctions between urban and rural areas [47], [97]

OpenStreetMap contributors are a major source of this heterogeneity. Their individual characteristics and variations have a profound bearing on the nature and quality of OpenStreetMap coverage. Traditional approaches to quality assessment such as the ones outlined in the preceding sections have been conceived to work in a more regulated, centrally controlled environment within the contexts of academic geography or national mapping agencies, which do not experience the same level of variability in terms of their contribution practices and contributor profiles.

Intrinsic modes of data quality analysis have arisen to address the practical problems posed by assessments using comparison with reference data and many of these uses edit history and provenance information to make predictions about aspects of data quality on the map. However, as Goodchild [110] points out, a framework for VGI quality assurance will need to account for variations in contribution practice. Whilst qualitative investigations can make a valuable contribution to that effort some analysis the scale at which these investigations can be conducted is limited. Approaches which use provenance to understand OpenStreetMap contribution practices at scale have a valuable contribution to make to the development of VGI quality assurance frameworks.

2.4 Provenance

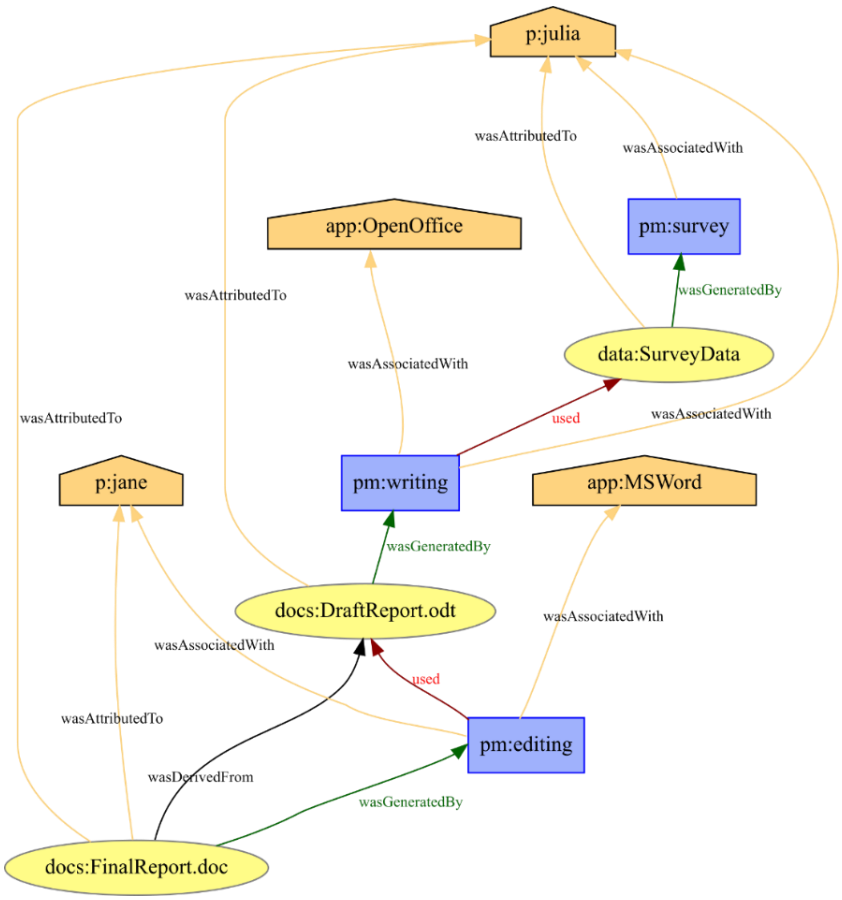
Although the term provenance, from the middle-French come forth, arise, originate, has been with us since the late 18th century, ideas about data provenance are a much more recent response to changes in the nature of information brought about by the adoption of databases and

the Web, and the decline of paper documents and files [111]. This section looks at some of the important domains of concern for provenance research. Earlier it was noted that provenance data is naturally expressed as a directed graph. What emerges from the work discussed here is how many possibilities exist for different graph expressions of the same provenance.

2.4.1 Provenance Graphs

The W3C defines provenance data as a record of the artefacts, agents, and activities which had a role in the creation of a piece of data or thing [112]. This data is naturally expressed as a network graph. The vertices in the graph represent the components and events which led to the creation of the item, and the edges represent relationships and dependencies between them. For example, the provenance graph in Figure 1 describes the creation of a report document. Juliet produced a draft report document, DraftReport.odt. She did so in a writing activity with the OpenOffice application. The writing used some survey data, which was generated by a survey activity also carried out by Julia. Jane then produced a new version, FinalReport.doc, in an editing activity with Microsoft Word. This was derived from the draft report document. This graph uses the W3C's PROV data model, PROV-DM, which provides a set of definitions for various vertex and relationship types. These allow provenance information to be encoded in a complex network structure. The definitions used in PROV are described in Subsection 2.4.8.

Figure 1: An Example Provenance Graph



2.4.2 Database Provenance

Early provenance research was focused on databases and the tracking of queries to explain how data tuples have arrived in a result set as well as tracking the movement of data through and between databases. This remains an important strand of provenance research.

Buneman et al [111], from 2000, addresses issues about the provenance of data in scientific databases. They discuss previous work on provenance, or lineage [113]–[115], which focused on the provenance of data from query operations on input databases, trying to formalise provenance in terms of the tuples which contributed to the output, i.e. a tuple which, if changed, affects the presence of the result tuple. This category is known as '*why*' provenance. It can be discovered by inverting the query to discover the contributing tuples which explain the presence of a result tuple in the output.

Where provenance is also identified by Buneman et al, i.e. provenance that answers the question of where in an input database a value present in a result tuple came from. In [116], Buneman et al describe a formal model for *why provenance* and *where provenance*, which trace query results as a database path, an early notion of a provenance graph. These concepts of why and where provenance were expanded upon by Cheney et al [117], who developed the idea of how provenance to describe the make-up of the manipulations which produce a result triple, e.g. by showing how many times an input tuple contributed to the presence of a result tuple.

Chapman et al [118] then built on this by describing algorithms for computing why not provenance, explaining why a given tuple is not in a result by tracing the operations involved in evaluating a database query. As discussed in [113], [114], database provenance differs from object provenance because it operates within the constraints of SQL and relational algebra, which makes it possible to reverse engineer queries, i.e. if a source tuple is known, we can compute an output tuple, and vice-versa. The provenance of things, objects and non-relational data poses different problems and needs to be captured rather than computed. Issues surrounding the way provenance data are captured are an important theme in provenance research.

2.4.3 Scientific Workflows

The proliferation of computing resources, sensor techniques and data storage and transfer technology has increased the role of data in science. As a result, scientists are increasingly working with collections of databases rather than direct observations. The increasing use of in-silico experimentation in research has resulted in the emergence of scientific workflow management systems (SWfMS) to document and reproduce these procedures [119].

These systems can be divided into two categories: management frameworks for grid computing middleware, such as Askalon [120], and MyGrid [121], and discrete applications such as Taverna [122], Keppler [123], Chimera [124], Pegasus [125] and VisTrails [126], which work within computation grids and provide an interface for creating and editing workflows. Some of these are geared towards specific tasks, e.g. VisTrails, which produces data visualisations [126]. Some of them have in-built provenance recording architecture, e.g. VisTrails [127] and Chimera [124]. Zhao et al [128] describe the use of semantic web technologies for provenance capture in Taverna/MyGrid, which is an interesting case. Taverna is a workflow management system that works within the MyGrid environment. They describe four views of provenance, based on provenance use cases. Provenance metadata for these views are logged as RDF using a MyGrid/Taverna schema ontology and scientific domain specific ontologies. Using ontologies makes the provenance data usable by a wider community and the flexibility of RDF facilitates query, enrichment, and visualisation.

SWfMS are concerned with documenting the lifecycle of their data throughput and as such, are all fundamentally records of provenance, even if they are not ostensibly 'provenance aware'. This information still has to be logged as provenance data if it is to be of use in answering provenance related queries. Unlike databases, SWfMS do not have the relational algebra underpinnings mentioned in Section 2.4.2 and are determined by steps in the workflow, which are a series of 'black boxes' with potentially diverse internal behaviours such as web service calls, or script invocations [129], so we can no longer rely on computation to precisely calculate a provenance trace. Instead, some aspects of the provenance become dependent on the decisions made by the SWfMS developers. For instance, the raw provenance information produced by many workflow management systems can be exhaustive and is not always fit for use. Alper et al [130] identify use cases for workflow provenance and discuss requirements and strategies for encoding graphs of provenance for different audiences. These fitness for use decisions can produce graphs with different topological characteristics, which describe the same provenance.

Some attempts have been made to formalise and compute provenance in SWfMS, taking advantage of the fact that in this domain we are dealing with a chain of computational steps, which although black boxes, can still be categorised into types whose behaviour is sufficiently understood to allow researchers to devise taxonomies of workflows and their components [131], [132], or to build formal models of workflows [133], thus allowing some aspects of provenance to be computed. E.g. Bowers Et al [129] have formally specified dependency relationships by adapting the computer programming concepts of control dependency and data dependency [134]. They used these ideas to show how dependency annotations can be inferred within a workflow environment to generate provenance data.

2.4.4 System Provenance

Another area of concern for provenance researchers is provenance capture from operating systems. There have been several systems proposed and implemented, e.g. PASS [135] , OPUS [136], SPADE [137]. These whole system provenance capture systems can be used in major operating systems including Linux [135], [137] , Android [137], MacOS [137] [138] and Windows [135], [137]. They work by modifying or extending the operating system to record system events such as disc read and writes, file operations, function calls, etc. This type of provenance in its raw form is difficult to interpret because it tends to be large scale and fine grained [139]. Because the information originates from low level system events, it also lacks the high-level semantic information needed by most users [140], so these systems abstract it to a higher-level, deciding which elements are useful in a provenance graph, much like the fitness for use decisions discussed previously in [130].

Chan et al [139] point out that system provenance frameworks often lack the homogeneity required for domains such as intrusion detection. Some use their own model of provenance, and all are modelling data from disparate systems. Some use standards such as W3C's PROV recommendations [112] or the Open Provenance Model (OPM)[141], [142], to establish a common vocabulary for representing provenance, but these standards do not provide mappings between the domain, recorded provenance and the actual behaviour of the system being observed. This uncertainty is compounded by the lack of a universal and formally defined relationship between the content of system logs and the behaviour of operating system kernels. Reconstructing provenance from OSM history data is affected by similar issues. The OSM XML data model has no official schema and no formally defined model other than XML [143]. There is also no mapping between a provenance data model and the editing behaviour we infer from the version history, and we are reliant on the subjective decision making of data analysts during capture in order to interpret the history as provenance data. This uncertainty distinguishes the database provenance discussed in [113], [114], [116]–[118] from other strands of provenance.

2.4.5 Application Provenance

As well as whole-system-based provenance some strategies have been devised to make standalone applications provenance aware, i.e. able to log their processes and communicate with provenance recording middleware. These are often engineering approaches to building provenance enabled applications or modifying existing ones. PriMe [144] is a an approach in which the architecture of an application is analysed to discover the discrete actors involved in the passage of data through the application, so that it can be modified to log provenance. PriMe is also a

provenance design approach, in which the expressions of provenance are specified based on use cases and interaction with application users.

NoWorkFlow [145] generates provenance from the execution of Python scripts, providing tools for analysing and expressing provenance graphs. It uses the Python profiler to log the execution of user defined functions, producing data that is exhaustive and fine grained. The system summarises it in a provenance graph by aggregating the node by function call site (line number) number of activation and function name, merging nodes from the same loop. Edges are created from function calls, events where functions are called in sequence and from control flow returns. This capture approach produces a topographically altered graph to make it fit for use.

YesWorkFlow [146] is another script based system which captures prospective provenance, i.e. information required to describe and replay the program execution. The capture policy here is specified by the developer using a system of annotations in comments, much like those used by a Javadoc parser [147]. The resulting graph can vary topographically depending on how these annotations are specified [89].

2.4.6 Open World Provenance

The work on provenance systems discussed so far deals with closed world provenance, describing capture systems operating within the confines of a known domain, whether it be databases and SQL, scientific workflow systems, or a specific application.

Allen et al [148] introduce Open World Provenance. In government and commerce, many use-cases require provenance capture from disparate and often non-provenance aware systems, often in organisations beyond the control of the system collecting the provenance data. This makes invasive strategies such as retrofitting applications or operating systems impossible. Those systems from which provenance can be extracted often produce it at different levels of abstraction which cannot be integrated. Allen et al tackle this by proposing provenance collection at communication coordination points. They describe a system which monitors traffic on an enterprise service bus (Mule), from which it gathers provenance data using the PLUS system described by Chapman et al [149]. PLUS is a business workflow system designed for government and industry, which is designed for the distributed provenance capture described in [148]. This can be used by a variety of actors for different use-cases and Chapman et al [149] use node abstraction and aggregation to alter provenance graph representations to address data protection, confidentiality and security issues.

The problem of uncertainty mentioned in Section 2.4.4 , which makes the computation of provenance outside of databases problematic, becomes even more acute in Open World

Provenance. Causality is difficult to infer and tends to be overestimated, based on an assumption that all preceding events in a trace are causally linked [150], a problem that also arises when trying to model dependency in [129] where it was observed that many workflow systems assume that all input data to a node contribute to the output data, leading to inaccurate provenance traces.

Whittaker et al have tried to address this [150] with their approach: *Why Across Time* (WAT) provenance. WAT provenance uses a combination of '*why provenance'* [116] and state machines to formalise provenance in distributed systems and is arguably the nearest researchers have got to a formalism of Open World Provenance that allows provenance to be computed from a set of inputs and outputs. However, WAT provenance computation only works in systems with elements that lend themselves to being modelled as a deterministic state machine, which limits the scope of this framework. In Open World Provenance we have human agents and wildly disparate systems, some of which are effectively non-deterministic black boxes.

2.4.7 Provenance Reconstruction

The computation of provenance for data across the web becomes even more intractable unless applications have been specifically tooled to be provenance aware, and we are left with provenance reconstruction as a means of obtaining provenance data. Some efforts at provenance reconstruction address the problem of capturing provenance from applications in distributed systems, where application instrumentation is problematic because of the computing overhead and/or lack of access to source code. Ghoshal and Plale [151] proposed 'scraping' provenance from log files and built a framework which parsed system log files and used an XML rules engine to write provenance statements.

Other researchers have reconstructed provenance from files and file systems: Magliacane and Groth [152] used various semantic similarity metrics alongside file metadata to generate provenance graphs from clinical guidelines documents in a DropBox folder. Aierken et al [153] and Vasudevan et al [154] have developed a framework which uses a range of techniques which they term 'funnels' to categorise documents and news articles based on unsupervised machine learning and semantic similarity. The documents are first categorised using topic modelling and then these categories are increasingly refined by semantic similarity until provenance relationships can be inferred within each category.

Provenance has also been reconstructed from version control systems in software engineering. Git2Prov, the system proposed in [155] accesses commit logs via the Web using GitHub's API. Git, being a software version management system is, in many ways a provenance

framework in itself, containing implicit provenance: information about the entities activities and agents in the production of code. This information is designed to facilitate the rolling back and recreation of edits to code. The provenance in GitHub is made explicit by the Git2Prov framework, which translates the commit records into W3C PROV-DM (see Section 2.4.8) statements.

There has also been some interest in reconstructing provenance from social media posts. Whilst a certain amount of provenance is published by social media providers, i.e. timestamps, creator id, shares, likes, etc, there is a great deal more implicit provenance information contained in social media content. The recent concerns over fake news have fostered this interest as journalists and ordinary web users alike need to gauge the authenticity of online content by being aware of who produced and influenced it. Taxidou et al [156] built a framework which reconstructs provenance using their work on models of information diffusion on social media along with social media graph information and semantic similarity metrics. They also produced PROV-SAID, an extension to the W3c PROV framework to model and capture some of the more domain specific concepts.

As the provenance community begins to take up the challenge of provenance reconstruction for data on the web, an area which remains under researched is that of provenance reconstruction from edit history, and a substantive part of this work falls into that category. Edit history has been a feature of many GIS systems and databases, e.g. PostGIS [157] , Arc-GiS [158], GRASS [159] and Oracle [157]. Two important sources of spatial data on the web maintain edit history: Open Street Map and the Ordnance Survey.

Keßler et al[60], [61] have analysed OSM edit history to produce trust ratings based on editing behaviours, creating reputation metrics for users based on other edits made to their work and edits made to neighbouring features. e.g. a deletion of a feature or part thereof signifies a correction or a rollback, lowering user reputation; and edit to a feature surrounded by other often corrected features, without having needed correction bolsters trust and reputation. They also incorporate the ‘many eyes principle’, as described in OSM by Hacklay [44] i.e. a feature edited by a large number of users will tend to be of higher quality. Keßler and de Groot extend these ideas to use provenance data, which they adapt to OSM History [60] using an ontology based on a predecessor to the W3C’s PROV-O ontology. They also introduce the idea of provenance patterns based on the editing practices they have identified, which they extract from the edit history to derive their trust and reputation values.

D’Antonio et al [58] propose a model which builds on Keßler and De Groot’s work by defining a set of editing types as a basis for formal trust calculations. They have recently published

some implementation details in [59] where they propose a java tool to convert OSM data to RDF, performing a similar function to the XSLT tool we use in this project. Their tool produces RDF that respects the vocabularies they propose for their framework: an ontology based on the OSM provenance vocabulary from [108], which handles VGI edit history data. Although they borrow many concepts from [108], their ontology is designed to work across any VGI application which has a versioned edit history. They also use an ontology defined by the OGC to manage feature geometry and the FOAF ontology to manage human agents. Using their model, which derives trustworthiness scores from user actions and reputations as well as taking into account geometric interactions of features and thematic and geometric variations between versions of a feature, they were able to compute trust scores which correlate to similarity of those features with authoritative reference data. This suggests that patterns found in provenance graphs can be used to predict data quality metrics.

The authors of these works do not discuss in detail the way their provenance data are modelled, or how they extract it for OSM history data, but their use of a purpose-built provenance ontology would indicate that their graphs are designed specifically for their experimental use-cases. They identify provenance patterns based on prior knowledge and assumptions, such as the 'many eyes' principle identified in OSM by Haklay [44] and then test the trust score they derive against similarity with reference data, which validates the role of the pattern as a predictor of trustworthiness. They are, however, relying on their judgement to extract provenance from the history data, and this is a problem faced by any attempt to record Open World Provenance, i.e. to generate provenance metadata in the wild we must rely on provenance capture and reconstruction systems made by data analysts, which often rely on their subjective decision making, informed by the use-case they are satisfying.

2.4.8 Provenance Standards

As provenance usage becomes more widespread and the range of use-cases more diverse the need for the Open World Provenance discussed in [148] has led to the development of standards for provenance encoding. There has been work on technology independent models of provenance within specific architectures, such as PreServ[138]], a set of web service protocols for managing and collecting provenance, which has been implemented in scientific workflows, but not universally adopted. Open World Provenance sharing on The Web requires a standard and vocabulary which can be agreed by all the actors who make use of it.

This open provenance vision is outlined by Moreau in [142] who identifies the need for a provenance model that includes human agency, and describes the community initiative to create the

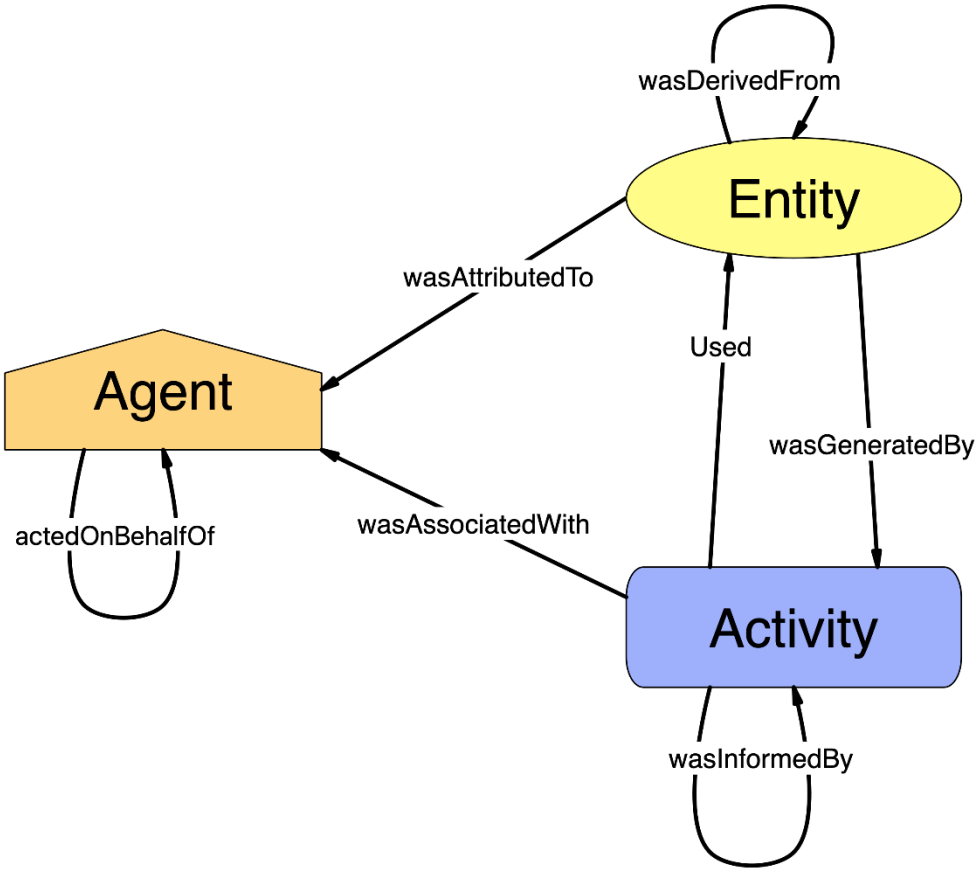
Open Provenance Model (OPM). Moreau outlines some principles for open provenance on the web, which are enshrined in the OPM, a model of provenance as a directed acyclic graph which can be RDF based and is capable of including real-world physical artefacts, as well as describing human agency. Use of RDF also means this provenance can be queried and reasoned over. The OPM is the predecessor to the W3C's 2013 PROV recommendation [160], which retains many of these features and principles.

In Geographic Information Systems, the widespread sharing, re-use and conflation of data led to the first international provenance standardisation in 2003 when the ISO 19115 geographic metadata standard was published with data lineage as a core component [161]. The standard defines a schema for encoding metadata for use with geographic data. It includes an XML Schema based specification for serialising the data and includes packages for citation to encode attribution and responsibility information, and lineage for describing events and processes in the lifecycle of the data. Interest in provenance from the geospatial community is evidenced by two engineering reports produced by the Open Geospatial Consortium [57], [162] which result from engineering initiatives aimed at producing or updating geospatial standards. Their OWS-9 Cross Community Interoperability Conflation with Provenance Engineering Report [162] looks at the role of data provenance tracking and measuring data quality in conflation workflows using web services that implement the OGC WPS geospatial web services standard. The OWS-9 work uses the XML based ISO 19115 standard, but after the publication of the W3Cs PROV recommendation, the OGC's focus shifted to this new model and their Testbed-10 Provenance Engineering Report [112] recommends adopting the W3C model.

The PROV family of documents, defined by the W3C, and forming the PROV recommendation [102], describe a conceptual framework and graph data model consisting of a set of defined nodes and relationships between them:

- **Entitie**s: the subject of provenance, the things we record the provenance of, e.g. books, articles, map features.
- **Activities**: events which create, destroy or cause entities to change state.
- **Agents**: things which act, initiating activities which create, destroy or change. The crucial thing about agents is that they have responsibility and have things attributed to them. They are usually people or organisations but can also be software.

Figure 2: PROV-DM: The W3C Provenance Data Model



These are connected by relationships which define attribution, derivation, association, and usage

Groth and Moreau [163] provide a detailed treatment of PROV and some example usage scenarios. The PROV framework provides RDF, XML and JSON serialisations as well as a native syntax called PROV-N, and the PROV-O ontology, an OWL encoding of PROV. The OGC Testbed-10 Provenance Engineering Report [57] contains a discussion about the advantages of PROV and rationale for choosing it over ISO 19115:

- The ISO model is better suited to dataset level provenance and would become highly verbose when used with GML (Geographic Mark-up Language).
- W3C PROV is a conceptual model and as such, is much less prescriptive than the ISO models, offering greater flexibility for linking geospatial elements, provenance, and their semantics.
- The ability to use RDF for serialisation, allowing easy updating of provenance information, combined with the option for using other serialisations for interchange of data.

Although W3C PROV seems to be overtaking the ISO Lineage model, particularly with geospatial data in the web domain, ISO lineage is not without its advantages. Jiang et al have evaluated both models in their study [164]. They observe that ISO 19115 is more expressive within the Geoscience domain, having been designed specifically for spatial data. W3C PROV has the advantage of being cross domain and interoperable but is not always fit for purpose in specific domains. This is borne out in Simmons et al's comparison study [165] of PROV and the VisTrails SWfMS in sport performance analytics - where analysts track the antecedent factors in sporting events such as injuries or goals. Whilst they found that PROV had advantages when trying to document ad hoc manual processes, it lacked the structures needed to document some of the more fine-grained aspects of automated parts of the analysis.

In Jiang et al's study [164] ISO19115 lineage and W3C PROV from earth sciences and geospatial datasets were compared using a set of typical provenance questions. They found gaps in both the PROV and ISO data and proposed a set of SPARQL rules and an OWL ontology to map ISO 19115 data to PROV getting the benefit of the richer instance data from the frame-based ISO model and the ease with which causal relations can be queried from the directed graph based PROV model. They also found that the domain agnostic PROV and RDF make it possible to incorporate external data sources. Evidence from both of these studies is indicative of PROV for use in a web context but recognise that it has limitations and potential for extension, and the W3C recommendation documentation [112] also envisages that PROV is extended for advanced use in specific domains. The mapping technique between ISO lineage and PROV [164] is also likely to be applicable to Ordnance Survey data.

2.4.9 Using Provenance

Many provenance aware applications produce raw data that are too large to be fit for use and we have seen approaches aimed at making improved provenance representations. Provenance can be used for some tasks, such as tracing attribution, by using specific queries which examine a tiny subset of the data. However, there are many tasks for which exploring an entire provenance dataset is necessary, including the formulating of queries, debugging workflows and provenance aware applications[166].

Projects such as PROV-O-Viz [167] and Prov Viewer [168] provide methods of visualising provenance graphs, but the size and complexity of many provenance datasets makes visualising them in their entirety impractical. Strategies have been employed to make visualisations more usable such as node summarisation and aggregation and tools to zoom into summarised nodes [166] but these still rely on concentrating attention on specific parts of the data. Other approaches have

tried to deal with the size of the dataset by automating the analysis, e.g. by applying weights and values to specific edges and nodes in a graph and then propagating those values across the graph to make a calculation [169], [170]. This requires knowledge of the application provenance, making the approach domain specific. It also relies on the judgement of the person applying those values [170], introducing opportunities for error.

2.4.10 The Art of Provenance Modelling

This section has described ways in which the shape of a provenance graph for an item of data can vary depending on the capture policy, which is affected by the use-case of the provenance, the constraints of the application and its capture mechanism. Examples of capture strategies that can alter the values of the graph metrics used in PNA from the same provenance include:

- Node summarisation for visualisation [166].
- Node abstraction for confidentiality and security[55] or for usability [139].
- Node aggregation for usability and efficient use of computing resources [145].
- Selective capture for publishing to different audiences (fitness for use) [135].
- User judgement[147].

Other work on provenance capture provides further examples. Ikeda and Widom [171] survey issues in data lineage, focusing on database provenance and identify strategies to compress very large provenance data. Pasquier et al [172] describe their system for capturing whole system provenance on Linux. They observed that capturing comprehensive system provenance can produce extensive provenance data with an unacceptable system overhead. Their system deals with these trade-offs and allows users to tailor their provenance record using capture policies. Coe et al [173] show how different types of capture agents have a different view of the data and processes which can result in different graphs of the same provenance. Goshal and Plale [151] tackle the problem of extracting relevant provenance using a rule based engine to allow users to specify and refine provenance capture. Yue et al[174] look at issues concerning provenance capture from geographic data at different granularities. Missier et al [175] describe ProvAbs: tools for abstracting parts of PROV graphs to provide access control, or simplified representations for specific audiences.

Provenance data is created and used in the domains outlined in this chapter, and which can be seen as spectrum based on the extent to which they can be computed. In databases and SQL, provenance can be computed, i.e. as long as we know the input and output tuples, the provenance is computable and the model depends on what questions we are asking, i.e. How[117] , Why [116], Where [116], Why Not [118]. In system and application, and workflow provenance this is increasingly less true, although we are still able to instrument elements of these systems to extract

inputs and outputs record accurate provenance, often dealing with elements that can be treated as state machines, allowing a degree of formalism and computation as described in [129], [150] , but also potentially producing a range of disparate provenance graphs depending on the provenance used case and capture policies devised to satisfy them.

At the other end of the spectrum are distributed systems with disparate nodes which generate data in a range of formats and at different levels of granularity. Many of these nodes are beyond the control of the provenance recording system and not provenance aware. At this end of the spectrum, where much of the data on the web and in VGI are situated, we are often reliant on provenance reconstruction, facilitated by human developers who must make subjective decisions about what implicit provenance exists in data and what that should look like when reconstructed as explicit provenance data.

2.5 Provenance Network Analytics

While the techniques described above are valid approaches to provenance analysis, some researchers have identified a need for an automated way to understand large scale data provenance that provides insights into the data by making use of the provenance in its entirety in a rigorous and principled manner [55]. Huynh et al have proposed such a method which they call Provenance Network Analytics[176]. They built on work by Ebden et al [176] who studied provenance data collected from CollabMap, a crowdsourced mapping initiative for disaster relief planning, which uses a micropayments crowdsourcing platform. The provenance from CollabMap was a collection of dependency subgraphs, capturing the details about how specific map features were created and edited. Their data was encoded using the Open Provenance Model and each graph was characterised by generating a series of graph theoretic measurements with some provenance specific variations:

1. Maximum Finite Distance(MFD), a measure of the longest distance between each of the OPM vertex types, ignoring unreachable nodes.
2. Diameter, the longest distance between two vertices in the graph.
3. Densification, the ratio of edges to vertices from which Edge to node correlation from which (ENC), a provenance specific measurement is derived by specialising densification by node and edge type and noting the Pearson's product- momentum correlation coefficient.
4. Degree distribution, the number of connected edges of different OPM types.

They also estimated degree distribution functions and discovered parallels with other types of network graph data such as that from social networks or the world wide web. A primary motivation was to predict issues in provenance capture, but they also identified some potential for

identifying malicious behaviour, by detecting communities of agent vertices to spot collaboration among users. They concluded that provenance data from CollabMap shows similar properties to other network graph data that have been studied, including social networks and the World Wide Web and that they can be used for inference in the same way. They also identify provenance graph network metrics as potential feature vectors for training machine learning classifiers. Huynh et al [70, 69, 68] built on this work by creating feature vectors from CollabMap feature subgraphs encoded using PROV-DM, the successor to OPM. The measurements they used were

- MFD.
- Number of edges.
- Number of vertices.
- Diameter.

Collabmap uses the find-fix-verify pattern for data creation, in which users are assigned the task of verifying the work of others, flagging a feature as trusted, or uncertain. Huynh et al used these to train a decision tree classifier to predict the values of these trust flags for each map feature, which they were able to do with around 95% accuracy. The use of a decision tree classifier also provides a human readable output (the decision tree) which can be interpreted to find the most significant metrics. In subsequent work [54] Huynh et al showed that PNA is suitable for use in multiple domains. They tested the method in the Radiation Response Game (RRG), a disaster response simulation, classifying messages sent between actors in the game scenario; and on the ProvStore, a web based PROV document repository, where they predicted the authorship of documents. Interestingly, in this later work, the authors observed that the classifier accuracy in the RRG was lower than the other applications, showing that not all provenance graphs contain the same level information required for predictions. By parameterising the graph to limit the dependency depth, i.e. the maximum allowed path length between two vertices, they were able to improve the accuracy. This suggests that different expressions of the same graph have the potential to affect the process.

2.6 Summary

Provenance is an area of active research in data science. In this chapter we have described some of the main themes: database provenance, scientific workflows, application and system provenance and Open World Provenance. We have examined some standards and models and discussed a novel method for characterising provenance graphs using graph theoretic measurements. We have also discussed research which shows how they can be used to train machine learning algorithms to predict characteristics of the data (PNA). Throughout this review, we

have seen situations in which the topology of a provenance graph can depend on the decisions made about the method of its capture and that it is possible to have graphs describing the same provenance that look different. We have also seen some evidence that suggests that varying the topological characteristics of a provenance graph can impact the accuracy of PNA [54].

Other research which has reconstructed provenance graphs from edit history has uncovered patterns in provenance graphs which can be used to estimate trustworthiness. It seems reasonable to ask if these findings might be exploited in OpenStreetMap. Can we use provenance network analytics to discover patterns which can be used to learn about the data and the modes of its creation? Provenance Network Analytics has the potential to predict data trustworthiness but is it reasonable to suppose we might implement it in OpenStreetMap? We have already explored some of the complex issues which drive OpenStreetMap's heterogeneity. Their scale and complexity greatly exceed any previous implementation of provenance network analytics. Before any such implementation could be considered, there is much groundwork to do to develop and evaluate appropriate methods of measuring OpenStreetMap provenance graphs and understand what insights can be gained from them.

Chapter 3 Methodology

3.1 Preamble

Provenance Network Analytics (PNA) suggests a methodological approach which harnesses the predictive power of provenance graph metrics. A common aim of PNA is to label data with the predicted value of some quality/trust metric. However, the use of quality metrics is fraught with epistemological, ethical, and practical difficulties, particularly for a global project such as OSM (see Chapter 1, Section 1.1 and Chapter 2, Section 2.3). These practical issues are a cornerstone of this thesis. A useful implementation of PNA in OpenStreetMap therefore requires considerable groundwork. Any such implementation would require a deep understanding of the nature and significance of the metrics which are to be used to measure provenance graphs captured from OpenStreetMap.

Webber, considering strategies for working with 1971 census data in Merseyside [177], distinguished two analytic approaches.

- P*redictive analysis*, designed to discover an optimal course of action for a specific use case.
- D*escriptive analysis*, which summarises and contextualises data so that analysts can use it as the basis for a predictive model.

He regarded descriptive analysis as a prerequisite, providing the necessary conceptual framework for informing predictive procedures. Predictive procedures are usually clearly defined one-off tasks, whereas descriptive analytics have a wide range of use cases. Geodemographic area classification was the descriptive strategy he used in his work "***…to summarise as economically as possible the varied types of residential environment found within the city.***" [177]. The work in this thesis falls into the category of descriptive analytics. It is designed to lay the groundwork for PNA by providing a conceptual framework based on insights gained from studying the relationships between provenance metrics and OpenStreetMap coverage.

The research questions outlined in Chapter 1, Section 1.3 are designed to explore the nature and interpretation of OpenStreetMap provenance graphs: what strategies are available for quantifying and measuring them and what insights can be gained from those measurements. This chapter outlines a descriptive analysis of OpenStreetMap provenance graphs in a series of experiments designed to understand the network graph metrics which can be used to enhance our understanding of VGI map production, laying the groundwork for more targeted PNA investigations for improved processes of documentation and quality labelling.

3.2 Research Questions

3.2.1 RQ1: How can approaches to the measurement of a provenance graph produce useful insights into the nature of VGI/UGC/OpenStreetMap ?

We define three approaches to the measurement of provenance graph data which enhance our understanding of provenance graph measurement. These are described in Section 3.3. We use them to identify provenance metrics and go on to implement a data pipeline that uses these approaches to obtain measurements. We also investigate novel capture methods for the analysis of geospatial provenance graphs and explore the implications of spatially aggregated provenance capture.

3.2.2 RQ2: What insights can be demonstrated about contributor editing behaviour and the mapped environment using provenance from VGI/UGC/OpenStreetMap?

We address this question by investigating the hypothesis that the nature of provenance graphs from OSM map data varies systematically because of distinct types of contributor editing behaviour and characteristics of the mapped environment. In Chapter 2, Section 2.2, we explore research into the nature of OpenStreetMap data, which suggests that mapping, and particularly aspects related to data quality, are likely to vary with properties of the physical, social, and built environment. Using census data from the UK 2011 census, and our analysis of the physical and built environment from the Ordnance Survey's topography layer we investigate these relationships with our provenance measurements to explore potential drivers of variance in our provenance metric data using three approaches. We examine:

- **Spatial variation** by examining thematic maps to investigate any spatial patterns in provenance variation and whether these correspond with spatial patterns in measurements of the physical, social, and built environment
- **Network variation** by inspecting examples of provenance graphs and comparing and contrasting those with high or low measurement values for a particular metric so that we can gain a visual understanding of the causes of variation.
- **Statistical variation** using
    - MANOVA procedures to investigate whether our metrics vary between geodemographically aggregated groups
    - Factor analysis to discover emergent themes among our metrics
    - Corelation between estimated proxies for data quality/maturity

3.3 Measuring Provenance: 3 approaches

In Chapter 2, Section 2.4, we looked at use cases for provenance data, most of which involve either answering provenance questions, or propagating values across a graph. We also looked at Provenance Network Analytics (PNA) [54], a more recent application of provenance data which uses measurements derived from a provenance graph as a machine learning feature. PNA uses graph theory derive these measurements from graphs encoded using the PROV data model. Decision tree output is used to provide insight into causal factors. We examine and further develop this idea that measurements from provenance graphs have the potential to explain the creation of large volumes of data. We propose three approaches to provenance graph measurement, outlined in Table 1.

Table 1: Provenance Measurement Approaches

	Abstract Metrics	Semi Abstract Metrics	Concrete Metrics
Description	Depend on the abstraction of provenance as a network graph	Still graph theoretic, but specific to provenance graphs	Uses domain knowledge to ask provenance questions by querying provenance graphs
Domain specificity	Any network graph data	PROV-DM Provenance graphs	OSM provenance graphs using the PROV-DM
Knowledge requirement	Graph theory	Graph theory, provenance data framework (PROV-DM)	OSM data model and understanding of VGI editing practices
interpretability	Very hard. Some insight possible with manual inspection of data	Hard. PROV-DM constructs help with manual interpretation	Easy. The metrics are self-explanatory

3.3.1 Concrete vs Abstract Metrics

The terms *abstract* and *concrete* and the distinction between them are concepts which originate in metaphysics, a branch of philosophy which deals with the nature of reality. There are numerous philosophical viewpoints on the defining traits of concreteness and abstractness, although there is broad agreement as to which objects are abstract and which are concrete [178]. Although many philosophers would regard all metrics as abstract objects, we use certain characteristics from Lewis [179] who identified four principal *ways of explaining* abstractness and concreteness. His *way of abstraction* defines an abstract entity as a specification of a concrete one; where that specification is a subtraction of some characteristic that all the concrete objects have in common.

For abstractions concerning real world objects this frequently involves its spatiotemporal characteristics.

An abstract object is therefore something which cannot be thought of in terms of space or time but is part or characteristic of an object which has a spatiotemporal dimension. Using this idea, we take the view that an *abstract provenance metric* is one which does not measure space or time, and instead focuses purely on the structure of a provenance graph. A *concrete provenance metric* is one which takes account of aspects of the physical world and requires specific understanding of the nature of the provenance subject and its spatial and temporal dimensions. Midway between the two are *semi-abstract metrics* which still rely primarily on the abstract graph structure, but also make use of theoretical constructs such as the W3C's PROV-DM (see Chapter 2, Section 2.4.8). These are in many respects, abstract metrics but the requirement for a specific data model puts them in their own category.

3.3.2 Abstract provenance metrics

Abstract metrics rely on an abstraction of provenance as a network graph. These purely graph theoretic measurements focus only on the network structure of the provenance, and as such, require no explicit understanding of specific data models. They are entirely domain agnostic, and examples of their use for analytics can be seen in in educational psychology [180], cyber security [181], computer networking [182], genomics and neuroscience[183]. In these applications, this type of measurement is used primarily for graph comparison or predictive analytics. Interpreting these metrics requires exhaustive examination of the graph data to understand what drives their variation. A full treatment of the graph theory underpinning these metrics can be found in Newman[184], [185]. These were calculated using the Python NetworkX Library [186]. Provenance graphs are normally directed and acyclic, but for several of the graph metrics this poses practical problems either because metrics such as distance can become infinite, or because using a directed graph does not produce sufficient variation. Other provenance network analytics studies [54], [56] dealt with this by conversion into an undirected graph, and we use this strategy. There follows a brief description of the different graph theoretic metrics used in this study.

Average Clustering Coefficient. The clustering coefficient of a vertex is a measure of the extent to which its neighbours are connected to one another. It can also be seen as a count of the number of triangles in the graph. We measure the mean clustering coefficient of the nodes in a provenance graph.

Power Law Exponent. The degree of a vertex is a measure of the number of connections it has with other vertices. This metric is a measure of the degree distribution in a provenance graph., Degree distributions obey a power law [185]. This metric is the exponent of that power law.

Density. This is the ratio of the number of possible edges in the network to the number of actual edges.

Average Rich Club Coefficient. The rich club coefficient is a measure of the extent to which a node is connected to other well-connected nodes. It is calculated for each degree of a vertex in the graph and is the ratio of the number of actual to the number of potential edges for vertices with a greater degree. The network algorithm the network X algorithm returns a dictionary of rich club coefficient values, keyed by degree and we use the mean average

Transitivity. Transitivity is the ratio of actual to possible triangles in the graph. It is calculated using the ratio of triangles to triples (potential graphs)

Assortativity. This is a measure of the extent to which vertices are connected to other vertices with similar degrees. Full details of the calculation are available in [187], and these are the equations used in the network X library.

3.3.3 Semi-Abstract Provenance Metrics

Semi-abstract provenance metrics include many of the metrics used in PNA [54], [56]. This category of metric still uses graph theory but distinguishes between PROV-DM vertex types. Their interpretation still involves detailed inspection of a provenance graph but using the PROV-DM provides additional insights into what drives variation in these metrics. These metrics are in many cases similar to the abstract versions but use PROV-DM specific vertex types.

Agents, Activities, Entities. These are counts of the PROV-DM node types in the graph.

Type Specific Clustering Coefficients. Node specific clustering coefficients are calculated for each PROV-DM node type.

Type Specific Degree Distributions. Power law exponents are calculated for each PROV-DM node type

Maximum Finite Distance. This measures the longest distance between vertices for each PROV-DM type. It is a PNA specific metric originally proposed by Ebden et al [56]. We calculate this using our own purpose-built algorithm, using NetworkX to calculate distances.

Type-specific degrees. Degrees are calculated for each PROV-DM vertex type.

3.3.4 Concrete Provenance Metrics: Maturity

As outlined in Chapter 3, Section 3.3, a concrete metric measures some tangible aspect/s of the physical world. Values for such metrics are obtained by evaluating answers to provenance questions. For such a metric to provide meaningful insights, a conceptual is needed for the formulation of provenance questions. To be useful, it should provide some constraints which isolate some real-world value which enhances human understanding of the data.

Some studies have used provenance data to model trustworthiness by identifying and recording specific edit actions carried out by contributors [58]–[60]. The provenance network graph is crucial for these types of analysis because it combines a view of the state of the feature with the sequence of events and prior states that led to it. This permits the measurement of those edit actions to produce concrete provenance measurements by querying provenance data. Most of this work is from OSM and Wikipedia and use provenance data derived from editing history logs recorded by both platforms

Wikipedia has parallels with OpenStreetMap in that anyone can edit it, leading to scepticism among the academic community and wider credibility issues despite studies which find its quality comparable to commercial offerings [35], [36], [188]–[190]. We have investigated various metrics derived from Wikipedia's article provenance which have been used by researchers to investigate the maturity of an article and derive automated predictions of its quality rating. Wikipedia articles develop through a series of life-cycle stages particularly defined by quality criteria, and we contend that OSM data also undergoes a similar maturation process. Using relevant literature looking at quality analysis for OpenStreetMap and Wikipedia, we have identified metrics which are related to data maturity and have the potential to be used an automated trust labelling as well as providing insight into the development of OpenStreetMap data and the practices of its creation.

Maturity. Maturity is broadly defined in the Oxford English dictionary as being "complete in natural development or growth" [191]. Ontologically it is defined a thing which has attained an advanced and settled state [191]. In economics it describes an economy that is developed to a point at which substantial expansion and investment no longer occur [192]. In Wikipedia, the ultimate stage in quality assurance is when an article attains *featured article* status which signifies a high level of quality and trustworthiness [193]. One of the criteria for featured status is stability, i.e. having content which does not change significantly from day to day. OSM has no equivalent to Wikipedia's quality assurance processes and no internal definition of maturity, although researchers have identified some characteristics of it which suggest that OSM data also undergoes a maturation process and that mature data is likely to be more trustworthy and of higher quality [194].

Maturity in Wikipedia. Wikipedia has an internal quality assurance mechanism which includes a peer review and editorial process during which articles are rated, nominated, and voted through a series of life-cycle stages defined by various quality criteria. They begin life as a "stub" article progressing through various stages until reaching "good" and finally "featured article" status [195]. This progression happens as part of an article' s life-cycle and represents a process of maturation and improvement. The effectiveness of Wikipedia's quality control processes are not entirely evenly distributed and the effectiveness of peer review depends on the nature and edit frequency of its editorial group, so it is likely that quality ratings of Wikipedia articles are a function of their usage and edit frequency [196].

Like OpenStreetMap, data quality in Wikipedia is an active research topic which has devised numerous strategies and metrics for providing quality labelling of articles. Wikipedia also has large volumes of data and the English edition currently hosts over 6 million articles, with an average of 572 new articles created each day [197] . This means there is a similar need for a principled automated quality labelling process using data metrics to gauge information quality and trustworthiness.

The Wikipedia article metrics shown in Table 2 have been calculated in studies investigating potential automated quality assessments of Wikipedia articles. They use machine learning techniques either to predict quality flaws [198] or the QA status of an article, i.e. whether it has achieved featured article status [92], [199], [200] , or to predict whether an article is about a recognised high-profile topic, which is perceived as an indicator of high quality [201]. These studies all make use of metrics which measure some dimension of an article's edit history. Some use specific editing patterns such as reverts to edits, others count edits and unique editors or focus on temporal dimensions such as the percentage of edits taking place in the last three months, edit frequencies over time, and variations in revision rate and revert rate.

Other studies have carried out a more detailed profiling of articles [198], [199], categorising metrics in by readability, structure, style, text metrics, linkage to other content and edit history. These metrics were used to train classifiers to predict article ratings and quality flaws. Both studies achieved good classification results with some of the highest information gain [199] and precision scores [198] coming from edit history related metrics.

Table 2: Wikipedia Article Metrics From Edit History

Author	Metric	Base calculations
Lih [201]	Linus's Law	total number of edits, total unique users
Stvilia et al [200]	Linus Law, volatility, currency	number of unique editors, total edits, median revert time, reverted edits
Dalip et al [199]	Lifecycle, Currency	percentage of reviews over the last three months, Revision rate by time and user, editor count
Anderka et al [198]	Linus Law, currency, Lifecyle,	days between creation and now, days between last update and now, number of editors, number of edits, percentage of edits in last three months, edit frequencies over time.
Wilkinson and Huberman [92]	Linus's Law, collaboration	edits and distinct editors
Wohner and Peters [202]	Volatility, Edit characterisation,	Transient edits

Maturity in OSM. Wikipedia has a baked in quality assurance process through which an article matures in a defined series of life-cycle stages. This makes maturity an intrinsic property of Wikipedia content. In contrast, OSM quality assurance is not built in and relies on external actors and processes, with no definitive life-cycle. Because of this, explicit discussions around data maturation are much less apparent in OSM research. Notable exceptions are the work by Gröchenig et al [203] which explicitly discusses maturity and identifies phases in the life-cycle of OSM feature representations based on edit intensity. Maguire and Tomko [194] also conceptualise maturity, defining it as "a convergence of feature representations to a tacit and consensual format". They observed the evolution over time of buildings from point to polygon representations, drawing a distinction between maturity and completeness. The other studies Table 2 do not explicitly reference maturity but do identify characteristics which change uniformly over time and so could represent a process of maturation.

Table 3: Table 2: OSM Article Metrics From Edit History

Author	Metric	Base calculations
Haklay et al [44]	Linus's Law	Editor count.
Kessler et al [60]	Linus's Law, volatility	editor count, revert rate and edit count.
Arsanjani et al [87]	Linus's Law	revision counts, editor counts
Gröchenig et al [203]	Lifecycle maturity	Ratio of edit intensity between 3 lifecycle phases - late-stage stability
Maguire and Tomko [194]	maturity	Convergence of features into a universally agreed format
Rehrl et al [204]	Linus's Law, Volatility	Transient edits, Edit count
Quattrone [205]	Linus's Law	Maintenance edits
Mooney and Corcoran [206]	Linus's Law, Volatility	Edit counts, editor counts, tag reversions

3.3.5 Maturity Metrics

Based on provenance analysis research literature for OSM and Wikipedia, we identify four dimensions to maturity: Linus's law, currency, lifecycle, and volatility, seen in tables 1 and 2 and defined in this section.

Linus's Law Maturity. We define Linus's law maturity as a measure of how many people have "seen" a feature. It is characterised by Raymond [18], who uses the maxim, "many eyes make bugs shallow" to explain the stability of the Linux operating system, which harnesses the wisdom of the crowd to overcome bugs and become highly trusted.

Linus's law maturity metrics are based on edit and editor counts. Being derived from studies of both OpenStreetMap and Wikipedia there is potential for different approaches to their measurement. In Wikipedia the atomic unit is an article whereas in OSM we are measuring multiple features within an area, which means we have two potential approaches measuring on a per feature basis, which broadly corresponds to strategies used in Wikipedia, or in an average per feature basis which seems more suitable for spatially aggregated map data. In map data there are subtle differences to each approach. Per feature measurements focus on activity surrounding specific features in an area, whereas simple counts are a more general measure of activity, focusing purely on all OSM primitives and providing a more area based, rather than feature-based measurement. We also use a specialised metric from OSM which reflects the amount of maintenance editing which was found to indicate activity by experienced editors [205]. From Table 2Table 1 and Table 3, we define five measurement methods.

Edit Count: the count of the number of versions of a primitive in the grid cell data, normalised by the number of OSM primitives.

Average edits per feature: the number of feature versions divided by the number of features.

Editor Count: The number of prov:Agents who influenced data within the cell.

Average editors per feature: the number of agents who have edited any version of the feature or any child features of that feature divided by the number of features in the cell.

Maintenance ratio: the number of Maintenance Edits in a cell divided by the number of Creation Edits in that cell.

Currency. Currency is a measure of how recently editing activity has taken place, and how "up to date" data is. Currency is not a measure of editing intensity. E.g. if data has only been edited once in five years, but that edit is recent, the cell will have a high currency value. It is generally a metric of Wikipedia articles, some of which use currency as a feature, but do not explicitly say whether high or low values indicate maturity/trustworthiness[198], whereas others regard recent editing as an IQ indicator [200]. We use two estimators of currency, both from Wikipedia studies [198], [200]. Days since last update and a count of new edits. There has been little work looking at currency in OSM, but in contrast to Wikipedia, a study has found that a recent stable period of low editing intensity after an earlier, more active period indicate data maturity [205]. We adapt days since last update to spatially aggregated data by dividing it into two metrics, one for the most recently edited primitive within the cell, and an average for all primitives within the cell.

Days since last update: The difference between the timestamp of the most recent version of a data primitive within the cell, and now.

Average days since Last Update: The average difference between the timestamp of the most recent version of all primitives within the cell

New Edits: *the number of edit versions within a cell with timestamps within one months of the edit history file download date*

Lifecycle Maturity. Changes in edit frequency over the life-cycle of data have been studied in Wikipedia by Li et al [207] who examined edit frequency and magnitude in Wikipedia articles at 3 lifecycle phases, including in the run up to nomination and attainment of featured article status. They found that featured articles had much greater increases in activity prior to nomination than articles which never became featured. They did however note that the nomination process has a direct effect on edit activity. In OSM, Gröchenig [203] modelled activity stages in the lifecycle of data

to estimate data completeness. He identified start, characterised by low edit rates; growth characterised by very high edit rates and saturation, a final stable phase with a low editing rate. We take a simplified approach, calculating the percentage of edits to data in a cell which occur in the final 20% of the cell's lifetime, i.e. the time elapsed between the first edit to data in the cell and the last. A very low value indicates saturation phase, i.e. a high level of maturity. This leads to a definition of:

> ***Life-Cycle Edits:*** The number of edits that occurred in the last 20% of an artefact's life divided by the total edits.

Volatility. Volatility is a measure of the rate at which edits are retained. This has been studied in Wikipedia by identifying transient edits as those reverted within one month [199], [200], [202], and the reversion rate [202]. In OSM similar studies have looked at reversion of edits to tags [60], [206], and this is the approach we adopt in order to simplify the process of measurement. In both OSM and Wikipedia studies high rates of reversion signify mature high-quality data the community quickly return to its previous state when attempts are made to change it.

> ***Tag Revert Count****:* the number of tag reverts, defined as a tag edited and then returned to its previous state in a subsequent edit.
>
> ***Revert Rate*** the average number of tag reverts per feature
>
> ***Transient Edit ratio*** the number of edits to tags reverted to their previous state within one month

Miscellaneous. We also include a maturity metric used by Quattrone [205], which can be readily calculated from OSM provenance data captured in census output areas

> ***Quattrone maturity:*** The ratio of number of features mapped and the population of the area

3.3.6 Other metrics Considered

Several of the provenance metrics were considered and their values calculated. Several matrix-based metrics were considered, such as the determinants of the Laplacian adjacency and rank matrices. Average In-degree, out-degree, and average degree centralities; edge and vertex counts; the chromatic number and number of components were also calculated. In order to refine the dataset some variables were removed as they were found to have very strong correlations with other variables ($\boldsymbol{\rho} > 0.9$). Corelating variables where selected for removal based on compute time and variance. Diameter is also a common metric for provenance network analytics but does not scale well and became intractable for larger provenance graphs and so was not used.

Figure 3: Removed Variables

Discarded variable	Correlating variable	Removal reason
Laplacian rank	Nodes	Compute time
Nodes	Density	Compute time
Edges	Density	Compute time
Upper average RCC	Average RCC	Compute time and variance
Edits per cell	Interactivity	Compute time and variance
Largest component fraction	components	Compute time
Average degree centrality	density	Compute time
Laplacian determinant exponent	density	Compute time
Adjacency matrix rank	density	Compute time
Entity-activity/Entity-agent MFD	Entity-entity MFD	variance
Revert rate	Revert count	Compute time, variance

3.4 Data Acquisition

Before we proceed with the measurement of a provenance graph, we need some principled means of defining those graphs. This raises the issues of granularity and identity: do we capture the provenance of individual OSM features, or do we capture the provenance of data in an area? …and whatever granularity we use, how do we define our units of aggregation?

3.4.1 Granularity and Aggregation

Provenance data can end up being many times the size of the original data and can easily become unmanageable, especially when dealing with big data [151], [208]. This has led to work on graph summarisation [209], using provenance node types to summarise part or all of a provenance graph and the refinement of provenance into *coarse-grained provenance* which "black boxes" transformations within a provenance graph and *fine-grained provenance* which describes data flow within those transformations [210]. Working with Geospatial provenance brings an additional set of granularity issues specific to map data.

For map data we can capture attribute level, feature level and dataset level provenance [57]. Features are the principal unit of a vector-based geographic data model. They are an abstract representation of a thing in the real world. These can be physical objects, such as streets, buildings and watercourses or intangible objects such as administrative boundaries or traffic routes[211]. These features are distinguished from one another in OpenStreetMap by using attributes, tags which provide semantic meaning to data primitives. As with many vector data models features can also be compositions of other features and not the atomic data unit which complicates decisions about what to capture.

A central question when recording a provenance graph, is how is the subject defined and what is its identity? In vector map data, entities frequently nest inside and make common use of other entities. This can make it difficult to see where one entity ends and another begins. A related issue also affects scientific workflow provenance. Workflow procedures can nest inside others in a broadly similar way and solutions have been proposed which involve aggregating provenance from multiple workflow runs in a process of normalising and integrating provenance graphs [212]. The computing overheads of this normalisation effort do not scale well, and this is likely to problematic for OSM provenance . OpenStreetMap contains over a billion *way* and *relation* primitives, almost all of which are some sort of feature, and a further 9 billion nodes, many of which are also features. Many are recombined and reused in a complex hierarchy which complicates the capture of feature level provenance graphs. It also means that for geospatial provenance there is another dimension to granularity: that of *area level* vs *feature level* provenance.

Geographic provenance data present a unique approach: *spatial aggregation*, the capture of discrete provenance graphs defined by a spatial zone. OSM features are seldom mapped in isolation because map editing is typically done over a region with several features being mapped together. This suggests that data assessment may be more meaningful if data is captured within a specific area. This, however, raises another set of problems for provenance analysis because this mode of provenance capture has the potential to introduce geospatial statistical bias

3.4.2 The Modifiable Aerial Unit Problem (MAUP)

Definition. The modifiable aerial unit problem is a type of statistical bias which arises when data is aggregated by geographic areas, also referred to as *aerial units* i.e. polygons which denote a two-dimensional area. It is closely related to the *ecological fallacy* [213], i.e. the notion that it is possible to generalise between aggregated data and data recorded at an individual level. Statistics often deals with the distribution of data, i.e. the frequency with which a given value occurs across a sample space. Our understanding of distributions and their properties allow us to perform formal statistical analysis. When our sample space is a geographic area, distinctly spatial phenomena come into play which change the nature of those distributions, giving rise to the MAUP. The MAUP has two components [214]:

The scale problem: This occurs when data are aggregated by grouping into different sized units. Although there is as spatial component to the scale problem in that it is affected by spatial autocorrelation, this type of effect also occurs in non-spatial data at different subdivisions. Most investigations look at regression slope coefficients and correlation coefficients, both of which can vary considerably at different levels of aggregation.

The zonation problem: This also affects correlation and slope coefficients and occurs with variations in the shape and location of the zones used to aggregate data, i.e. the zoning scheme used. It is more closely related to the concept of gerrymandering and differences between correlations and slope coefficients for different variables are more pronounced. It is sometimes also referred to as the aggregation problem in the literature.

History. The scale problem was first described in 1934 by Gehlke and Biehl [215] when studying spatially aggregated U.S. Census data. When they combined aggregation areas, they found fewer, larger zones increased correlation coefficients. This only occurred with contiguous groupings, so they concluded that the effect was spatial. Yule and Kendall confirmed these findings [216], [217]. They understood correlation as a relationship between variables with a systematic causal factor attenuated by other unrelated components. Grouping individuals in larger zones magnifies the effect of the causal factor while the unrelated effects gradually cancel one another out, thereby increasing the correlation. The scale effect was also found on bivariate linear regression with similar variations in slope coefficients [218]–[220]. These studies called into question a large body of research that is relied on the use of geographically aggregated data.

Openshaw was one of the first to describe the zonation problem in detail [221], [222]. In a study of correlation in spatially aggregated housing data he noticed that the shape and location of the zones produced pronounced effects independently of scale. He proposed the AZP algorithm [221], [223], which generates zones from aggregated data using *basic spatial units* which are exchanged between already adjacent zones to optimise an objective function. Using this procedure, he was able to design zone schemes which maximise correlation coefficients. Although gerrymandering was a well-known phenomenon, Openshaw's work showed the profound impact it had on statistical procedures in geographic data, even referring to his AZP algorithm as "applied gerrymandering" [221].

3.4.3 UK Census Output Areas: Demographic Data Aggregation

In the 19th century, UK census data was mainly used for reporting statistics of local authority areas, and these were used for census publication. The 20th century saw increasing demand for computer readable small area statistics, and enumeration district data became available in the early 1960s [224]. These small, originally hand drawn areas were designed to organise efficient data collection in the 19th century when households were visited by census enumerators. As geospatial data analysis became more sophisticated there was a demand for output areas standardised by characteristics such as population and social composition. A new output geography was designed for the UK 2001 census that was independent of the enumeration district [224]. This new output area

geometry was generated by an automated digital process. Each output area polygon is composed of a group of adjacent postcode polygons. These Thiessen polygons are generated from a list of geo-referenced addresses. They were merged with other addresses sharing the same postcode, and clipped to waterways, roads, and administrative boundaries to create polygons nested within parishes and wards. These polygons were grouped together using a zoning algorithm which swaps postcodes geometry between output areas to maximise the value of a target function. The resulting polygons were designed to be as internally homogenous as possible in terms of population, accommodation type and tenure [225].

Output Areas and the MAUP. UK census data is published in aggregated form to preserve the anonymity of respondents. It is widely used by social scientists and policymakers to understand demographic trends, test theories and evidence policy decisions. Data is released at several scales from census output areas up towards, districts and larger middle layer super output areas. This is often the case in other countries and so the MAUP is a potentially serious issue for statistical analysis of census data. Its effects were explored by Flowerdew [226], who studied UK 2001 census data at output area, ward and district level. He found evidence of MAUP effects, although some of these were smaller than other studies. There were no variables which went from positive to negative correlation at different scales. When assessed using the Fisher's Z test, most of the correlation variations, including the smallest, were found to be significant. Flowerdew was unable to predict which variables would be affected, so although in many cases the MAUP effect was not severe, its effects are variable and unpredictable

The MAUP is clearly an issue for any research involving the aggregation of spatial data. It has been shown to affect several statistical procedures and is likely to affect the aggregation of provenance data. Although the mechanisms driving the MAUP are still poorly understood, Blalock showed that homogeneity within aerial units reduces the role of potentially confounding variables by reducing their variance [218]. The MAUP is a form of statistical bias which means that the results of the statistical analysis of data aggregated by an aerial unit should not be generalised to other scales and units. Although heterogeneity within units may help ameliorate the effects, Flowerdew recommends choosing geographically meaningful rather than arbitrary units, which is likely to reduce the need for such generalisations [226].

In view of this we propose to use 2011 census output areas as our aerial unit for provenance analysis. They are generated using an algorithm based on Openshaw's AZP procedure [225], designed to homogenise demographic characteristics. Because they are units of census data publication, we are also able to access a lot of information about the physical and demographic characteristics of the environment they enclose. The UK Office for National Statistics (ONS) provides

a complete geometry for output areas, which each have a unique output area code. Despite these advantages it is important to note that these facts do not solve the MAUP, and caution should be used when generalising analysis results to individual OSM features or to other aerial units.

3.4.4 The Output Area Classification (OAC)

Another advantage of using census output areas is that they have been classified according to their demographic characteristics. The OAC is a set of classification groupings based on the demographic properties of output areas, derived from their census data. It is available for all output areas in England and Wales. The 2011 OAC is a nested hierarchy consisting of 8 supergroups, 26 groups and 76 subgroups. This provides an opportunity to investigate whether provenance graphs vary according to the demographic characteristics of the coverage area.

Area classifications have long been a key tool for revealing information about people based on where they live, using census data. Its roots go back into the 19th and early 20th centuries. Early proponents were Charles Booth in the early 19th century, who produced colour-coded maps of the spatial distribution of poverty in Victorian London indicating demographic properties categorised by poverty, industry, religion, and morality. Sociologists of the Chicago School in the 1920s and 30s developed ideas about human ecology and began to model the spatial and temporal relations of populations in terms of their size, shape, industries, transportation, buildings, etc [177], [227][177].

In the later 20th century, the growth of computing resources and sophisticated GIS applications saw a growth in area classification and geodemographics, which became increasingly important for commercial marketing. A range of geodemographic projects appeared, such as ACORN (a classification of residential neighbourhoods) in the UK and PRIZM (Potential rating index for zip markets) in the US [228]. In the UK there were several academic/commercial collaborations to produce general-purpose classifications using census data. Many also used other, often sensitive data, such as credit histories, private surveys, and product registrations. As a result processes and variables used were not always published, leading to constraints on the extent to which these classifications could be validated [228]. in 2001 and 2011 the ONS undertook geodemographic classification of census output areas using an open-source approach. All the data used came from the UK census and the methodological details were published. This is a significant advantage over other classifications because they can be critically evaluated or extended [229].

The 2001 and 2011 classifications used the K-means algorithm to cluster output areas [229]. The K-means algorithm starts with a specified number (K-seed) of random cluster centroids. Data points (output areas) are then assigned to each cluster based on the value of a distance measure from the centroid. A new cluster centroid is then calculated and used to reassign data points. This

process is iteratively repeated until the clusters stabilise. The methodological steps of the classification process are outlined in [230]. The process starts with exploration and variable selection, where strongly correlated variables are removed or merged. Euclidean distance was chosen as a basis for defining the usefulness of clusters produced using different K values. This is followed by interpretation, testing and replication to assess the significance of cluster structures within the data.

Two crucial inputs to this process are the K-seed, i.e. the initial number of clusters, and the selection of variables used. Both parameters were chosen after careful data exploration, consultation with the ONS and other stakeholders and using prior experience from previous classifications. For the 2001 OAC, 41 variables were selected [229], and for 2011, 60. Additional variables were used in the 2011 OAC to reflect the changing demographic landscape. For example, the 2011 OAC has enhanced indicators for Britain's ageing demographic and include some communal establishment variables to reflect older citizens living independently versus those living in care homes [231].

After a consultation exercise with the ONS and various end users, a 3-tier nested structure with six groups at the highest level followed by 20 mid-level groups and then 50 at the bottom level was planned for the 2001 OAC. This was used as a starting point for testing a range of K-seed values to minimise the average within cluster distance from the mean. The 2001 procedure identified seven top level supergroups and 52 subgroups [229]. The 2011 census used more variables and after further analysis, produced a similar structure but with eight supergroups, 26 groups and 76 subgroups [231] see Table 4.

Gale et al [231] refer to the names and descriptions of cluster grouping as the "user interface of geodemographic classification" each cluster is assigned a name and short "pen portrait" description. Gail and Vickers et al noted the sensitive and potentially contentious nature of this task which must address the possibility of introducing and reinforcing negative stereotypes and bias. Descriptors were selected which provided unambiguous links to the nature of the underlying data, avoiding the use of overtly positive or pejorative terminology. This process was finalised after a review conducted by the ONS and is presented in Table 4 .

Table 4: 2011 OAC structure (from geogale.github.io/2011OAC/)

SUPERGROUPS	GROUPS	SUBGROUPS
1 - Rural Residents	1a - Farming Communities	1a1 - Rural Workers and Families
		1a2 - Established Farming Communities
		1a3 - Agricultural Communities

		1a4 - Older Farming Communities
	1b - Rural Tenants	1b1 - Rural Life
		1b2 - Rural White-Collar Workers
		1b3 - Ageing Rural Flat Tenants
	1c - Ageing Rural Dwellers	1c1 - Rural Employment and Retirees
		1c2 - Renting Rural Retirement
		1c3 - Detached Rural Retirement
2 - Cosmopolitans	2a - Students Around Campus	2a1 - Student Communal Living
		2a2 - Student Digs
		2a3 - Students and Professionals
	2b - Inner-City Students	2b1 - Students and Commuters
		2b2 - Multicultural Student Neighbourhoods
	2c - Comfortable Cosmopolitans	2c1 - Migrant Families
		2c2 - Migrant Commuters
		2c3 - Professional Service Cosmopolitans
	2d - Aspiring and Affluent	2d1 - Urban Cultural Mix
		2d2 - Highly-Qualified Quaternary Workers
		2d3 - EU White-Collar Workers
3 - Ethnicity Central	3a - Ethnic Family Life	3a1 - Established Renting Families
		3a2 - Young Families and Students
	3b - Endeavouring Ethnic Mix	3b1 - Striving Service Workers
		3b2 - Bangladeshi Mixed Employment
		3b3 - Multi-Ethnic Professional Service Workers
	3c - Ethnic Dynamics	3c1 - Constrained Neighbourhoods
		3c2 - Constrained Commuters
	3d - Aspirational Techies	3d1 - New EU Tech Workers
		3d2 - Established Tech Workers
		3d3 - Old EU Tech Workers
4 - Multicultural Metropolitans	4a - Rented Family Living	4a1 - Social Renting Young Families
		4a2 - Private Renting New Arrivals
		4a3 - Commuters with Young Families
	4b - Challenged Asian Terraces	4b1 - Asian Terraces and Flats
		4b2 - Pakistani Communities
	4c - Asian Traits	4c1 - Achieving Minorities
		4c2 - Multicultural New Arrivals
		4c3 - Inner City Ethnic Mix
5 - Urbanites	5a - Urban Professionals and Families	5a1 - White Professionals
		5a2 - Multi-Ethnic Professionals with Families
		5a3 - Families in Terraces and Flats
	5b - Ageing Urban Living	5b1 - Delayed Retirement
		5b2 - Communal Retirement
		5b3 - Self-Sufficient Retirement
6 - Suburbanites	6a - Suburban Achievers	6a1 - Indian Tech Achievers
		6a2 - Comfortable Suburbia
		6a3 - Detached Retirement Living
		6a4 - Ageing in Suburbia
	6b - Semi-Detached Suburbia	6b1 - Multi-Ethnic Suburbia
		6b2 - White Suburban Communities
		6b3 - Semi-Detached Ageing
		6b4 - Older Workers and Retirement
7 - Constrained City Dwellers	7a - Challenged Diversity	7a1 - Transitional Eastern European Neighbourhoods
		7a2 - Hampered Aspiration
		7a3 - Multi-Ethnic Hardship

	7b - Constrained Flat Dwellers	7b1 - Eastern European Communities
		7b2 - Deprived Neighbourhoods
		7b3 - Endeavouring Flat Dwellers
	7c - White Communities	7c1 - Challenged Transitionaries
		7c2 - Constrained Young Families
		7c3 - Outer City Hardship
	7d - Ageing City Dwellers	7d1 - Ageing Communities and Families
		7d2 - Retired Independent City Dwellers
		7d3 - Retired Communal City Dwellers
		7d4 - Retired City Hardship
8 - Hard-Pressed Living	8a - Industrious Communities	8a1 - Industrious Transitions
		8a2 - Industrious Hardship
	8b - Challenged Terraced Workers	8b1 - Deprived Blue-Collar Terraces
		8b2 - Hard-Pressed Rented Terraces
	8c - Hard-Pressed Ageing Workers	8c1 - Ageing Industrious Workers
		8c2 - Ageing Rural Industry Workers
		8c3 - Renting Hard-Pressed Workers
	8d - Migration and Churn	8d1 - Young Hard-Pressed Families
		8d2 - Hard-Pressed Ethnic Mix
		8d3 - Hard-Pressed European Settlers

3.5 The Experiments

3.5.1 Interpreting Provenance Networks

In this section we carry out a detailed inspection and interpretation of individual provenance graphs. Focussing on those with high and low values of abstract and semi-abstract graph theoretic provenance metrics, we aim to understand what drives this variation. The visualisations were implemented with the Cytoscape software [232], which allows us to carry out a detailed inspection of the graph's edges and vertices. The graphs are encoded in RDF format (see Chapter 4, Section 4.1.1) with each edge and vertex represented by a URI. This means that each vertex can be resolved to an osm:node or osm:Way in OpenStreetMap. These can be viewed and inspected via a web browser using the OpenStreetMap point and click query tool, which provides easy access to the feature's metadata. It retrieves editing history, changeset comments, contributor details, and the changeset bounding box, representing the area where editing in that changeset took place. Bounding boxes for a single contributor's changesets are also available. Cytoscape also provides easy access to other graph metrics for individual nodes and provides options for colouring edges and vertices to aid interpretation. The geometry for the output area used to capture the provenance graph under investigation is loaded into the QGIS software [233] to produce a representation of the OpenStreetMap coverage.

We use this information to compare and contrast the characteristics of provenance graphs with high and low values of provenance network metrics. Examination of the map coverage using QGIS along with closer inspection of individual map features using OpenStreetMap's query tool also illustrates the effect that the physical environment and the way it is represented in OSM influence provenance graph network properties. The assessment considers the provenance graph alongside the OpenStreetMap coverage and its associated meta data to provide a detailed interpretation of the graph structure. Identifying factors which relate the provenance to the network properties of the graph explains how graph theoretic provenance metrics can provide reveal patterns of OpenStreetMap contribution (research question one)

3.5.2 VGI Provenance as a Geospatial Variable

Most, if not all, provenance can be linked in some way to a place. Either the creation event, or some entity or agent involved with it is linked to a location. In that respect provenance is geospatial data. There has been research examining the provenance *of* geospatial data, e.g. [57], [234]–[236], but to the best of our knowledge, none explicitly examining provenance *as* geospatial data. Some work has looked specifically OpenStreetMap provenance with the aim of building trust metrics by evaluating users and their edit actions [58], [59], [61]. However, these approaches mainly consider temporal variation. They cannot account for the spatial variability of provenance captured from data creation processes with a spatial dimension as is the case with VGI map editing. Studying the spatial distribution of provenance variables has the potential to reveal these spatial patterns in OpenStreetMap contributing and their potential drivers.

It is immediately apparent from inspection of thematic maps created using provenance variables that they represent deterministic and spatially variable phenomena. In this chapter, we unpack this to identify and interpret distinctive spatial patterns in our study area. We investigate some potential drivers using ONS 2011 census output area classification pen portraits [237], which provide information about the characteristics of output areas in OAC groupings. This is considered alongside detailed visual map inspection and correlations with measurements of the physical and built environment extracted from the Ordnance Survey MasterMap Topography Layer. Several themes emerge from the results of this examination which can be used to understand OpenStreetMap contribution patterns in terms of their relationship to the mapped environment, and to the editing behaviour of OpenStreetMap contributors (research question two).

3.5.3 Metric Analysis

Chapter 7 is divided into three sections describing a series of statistical experiments. The first addresses research question two by investigating concrete provenance measurements and what relationship they bear to the tangible real-world dimension that they represent (see Chapter 3, Section 3.3.1) . The concrete metrics we use in this thesis have been conceived to represent *data maturity* (see Chapter 3, Section 3.3.4) as a possible proxy for trust/data quality. This experiment addresses research question two by asking whether concrete metric strategies might predict aspects of data quality. The second experiment seeks to address research question two by uncovering latent variables in provenance data using exploratory factor analysis. These represent phenomena which cannot easily be directly measured but can be inferred from direct observations. The third experiment examines the role of demographic characteristics of the map coverage area, seeking to understand with OpenStreetMap contributions differ according to demographic area classifications. A post-hoc discriminant function analysis is carried out to understand what factors, might distinguish any differences found.

Evaluating Data Maturity. In Chapter 7, Section 7.2, we used two proxies for data quality: a score based on a comparison with OpenStreetMap and satellite imagery, and a score based on output from an automated OpenStreetMap error detection application. We assess relationships between the schools and our maturity metrics using Spearman's rank correlation coefficients.

Exploratory Factor Analysis. Many of the insights available from provenance analytics are not necessarily connected with single variables. In Chapter 5 and Chapter 6, several themes emerge representing insights into the patterns and variations of OpenStreetMap contribution. The identification of these themes has required detailed interpretation of thematic maps, network graphs and the physical environment. These have been used alongside measurements of provenance metrics to infer, rather than directly measure insights. Variables which cannot be directly measured or easily observed are sometimes referred to as *latent variables*. In this section we continue to address research question two by using statistical methods to reveal latent variables.

The notion that the phenomena we observe have underlying causes which cannot be directly observed is well-known and forms the basis by which humans understand aspects of our daily lives, based on unseen concepts and intrinsic knowledge [238]. The inference and modelling of latent variables derived from direct observations goes back to the work of Spearman in 1904 [239]. . He recognised contemporary weaknesses in experimental psychology, derived from reductionist approaches of the physical sciences. Spearman took the view that these failed to account for the complexity of the human psyche and modelled a factor of general intelligence based on a range of

observations. Thurstone extended this work to account for multiple factors [240] and the practice of factor analysis has since become widespread. This is particularly true in psychology and the social sciences which often deal in complex phenomena which are constructs rather than simple observations found in the physical sciences [241]. Since then, scientists from numerous disciplines have gained valuable insights into complex real-world constructs by building mathematical models from observations. These are used to infer distinct characteristics and dimensions of real-world constructs such as personality [242], intelligence [240], [243], human personality [244], pathologies of developmental disorders [245] and psychiatric conditions [246], and the demography of sexuality [247]. VGI mapping is also a real-world construct which cannot be easily understood in terms of single, simple observations. In this section we use factor analysis to understand the themes and insights which can be gained from studying and modelling provenance data (research question two).

Factor analysis reduces the dimensionality of a dataset by transforming a set of variables which have linear relationships with each other into a smaller set which accounts for as much of the dataset's variance as possible. This is done with one of two objectives in mind. Factor analysis can be used for *dimensionality reduction* to simplify and avoid overfitting a model. Alternatively, *exploratory factor analysis* can be used to understand the underlying structure of a phenomenon by inferring latent variables from observed variables [248], [249]. The techniques involved differ somewhat between the two approaches.

Factor analysis for dimensionality reduction is generally known as Principal Components Analysis. In this technique we decompose our dataset into a set of components which we assume will account for all of the variance in our dataset. This means all of the variance from the input variables is considered. Exploratory factor analysis reverses this assumption. We assume that there are underlying factors which contribute to the variation in our phenomena, but there is also error variance, and variance which is unique to each individual variable and not shared with the factor. Exploratory factor analysis accounts for this unexplained variance and only considers variance which is shared between variables [249].

Some writers of statistics manuals, such as Field [248] highlight the mathematical similarities between the two procedures, both methods which decompose a correlation matrix into eigenvectors to find linear combinations [248], [249]. Field does not emphasise the distinction between the two methods, citing a literature review from Guadgnoli and Velicer [250] which claims that both procedures give broadly similar results in many situations. Others are much clearer on the distinctions. Tabachnick and Fidell [249] recommend choosing between the two methods based on research goals. For a theoretical solution uncontaminated by unique and error variability, factor analysis is their recommended choice. Guadgnoli and Velicer [250] also note that their experiment,

the results show that the methods differ where there are variables with low shared variance (communality). Any study of spatially aggregated geographical data naturally has sources of unexplained variance, as per Tobler's second law of geography: "*the phenomenon external to a geographic area of interest affect what goes on inside*" [251], [252], and VGI provenance data is no exception. Few of our variables have strong correlations and the presence of unique and error variance would be likely to distort the results of PCA. In the previous two chapters we have identified themes within the data which provide evidence for the existence of latent variables. In this study we are seeking to identify and understand the structure of these latent constructs and so exploratory factor analysis, which isolates and studies common variance is our chosen method.

Assumptions. Exploratory factor analysis requires the following assumptions to be met.

Multicollinearity. Multicollinearity is a multivariate version of linearity, i.e. correlation relationships between variables. In multivariate analysis, moderate correlations indicate a degree of common variance, but very strong correlations are wasteful because variables are essentially carrying the same signal. Multicollinearity also includes multivariate correlations, i.e. relationships between one or more variables [248]. Multicollinearity is a serious problem in exploratory factor analysis because it obscures the unique contribution to factor from variables or groups of variables which have strong correlations [249]. One option is to inspect a correlation matrix for highly correlating variables, but this will not detect correlations between more than two variables. Another heuristic is to examine the determinant of the correlation matrix, which should be greater than 0.00001.

For accurate identification of variables which are responsible for multicollinearity we can use the Variance Inflation Factor (VIF) method which is a by-product of the regression procedure in SPSS [248]. The VIF is a component of variance of the standardised slope in a multiple regression and is a measure of the extent to which this variance is inflated by relationships between predictors [253]. VIF values can be obtained by running multiple regression in SPSS. A VIF greater than 10 indicates multicollinearity issues with that variable [248].

There Should Be a Linear Relationship Between Variables. Some guides recommend using scatter plots to assess this , but as we have a large number of variables, we assess this using inspection of a correlation matrix to ensure each variable has some degree of correlation ($r > .2$) with at least one other variable. There should also be some correlations greater than .30 [249]. Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is an identity matrix, i.e. having one 1.0 in the diagonal and 0.04 other values, indicating that no relationships exist between variables. This should also be significant, although in reality this is almost always the case

[248]. Another test which can determine the suitability of a dataset for factor analysis is the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. This statistic is a measure of the amount of common variance present in the dataset. KMO scores above 0.5 indicate some degree suitability for factor analysis [254].

There Should Be a Large Sample Size. Sample size should be greater than the number of variables [248], [249]. Sample sizes in excess of 1000 are excellent and allow interpretation of small factor loadings.

Communality. Communality is a measure of the amount of shared variance associated with each variable. It is computed by summing the squared factor loadings of each variable. These are obtained from an initial run of a factor analysis procedure. Communality represents the extent to which a multiple regression analysis can predict the value of a given variable [248], [249]. If variable has low communality values, this suggests that it contributes little to any other factors and should be removed. We use a value of less than 0.2 as a criterion for removal [249].

The Procedure. Exploratory factor analysis is an iterative procedure which is first run to test assumptions as detailed above and then to select factor rotation method and achieve a simple structure.

Number of Factors. Among the outputs of the SPSS factor analysis procedure is a scree plot. This is a line chart which plots each factor against its eigenvalue. The eigenvalue of a factor provides an indication of the amount of variance it accounts for. An eigenvalue of one represents the amount of variance accounted for by a single variable. Kaiser's criterion suggests that the number of factors to be extracted is simply to use all factors with eigenvalues greater than one. This has been criticised as inaccurate as a feature selection method [255]. Factors can be selected using the elbow method, i.e. selecting factors above an inflection point on the scree plot, which indicates a point of diminishing returns in terms of variance accounted for [248]. Kaiser's criterion can still be considered as a lower bound, below which inflection points were disregarded as these factors would account for less variance than one of the original variables [255].

Factor Rotation. After the generation of factors from the eigenvectors of the correlation/covariance matrix, the strongest factor will dominate the factor loadings as it accounts for most of the variance. Variables will tend to load more strongly on this factor and weakly on others which makes the structure difficult to interpret [248]. To aid interpretation of the factor loadings, the eigenvectors from which the factors are derived are rotated to maximise variable loadings on each factor. Where the initial factors have weak or no correlation, an orthogonal rotation is used, which maintains perpendicular factor axes, producing an uncorrelated factor

solution. Alternatively, an oblique rotation can be used which allows factors in the solution to correlate with one another [248], [249]. This decision is based on whether factors from an initial run of the procedure are correlated. Absolute correlations greater than 0.3 indicate use of an oblique rotation method [249].

Simple Structure. The end result of a factor analysis procedure should be a factor loading matrix which has Thurstone's simple structure. This means that all variables should load strongly on at least one factor, no variables should load strongly on more than one factor and all factors should have zero or minimal loadings for one or more variables [256], [257]. This sometimes involves eliminating variables from the analysis which have strong cross loadings.

3.5.4 Analysing and Comparing Variance: MANOVA

In Chapter 5, Chapter 6, and Chapter 7 we find evidence that the individual characteristics of OpenStreetMap contributors leave measurable signals in provenance graphs, as do the physical and topological properties of the environment and its representation in OpenStreetMap. Other researchers have found evidence that some demographic characteristics of an area affect the extent to which contributors engage with it [29], [32], [33], [97], [258]. This is partly due to a propensity of contributors to map areas they are familiar with and which therefore reflect their own demographic profile [26], [82], [83]. This explains why studies such as [29] find that heavily contributed areas have similar demographic profiles to those of OSM contributors. These studies rely on questionnaire surveys of OpenStreetMap contributors, analysed with regression and clustering techniques. If we can find evidence that these patterns are reflected in provenance graph metrics this will identify further insights that can be gained from studying OpenStreetMap provenance graphs (research question two). The ability to distinguish the demographic characteristics in an area using its OpenStreetMap provenance would also be a valuable finding.

Demographic data in the UK is published by the ONS in spatially aggregated form using census output areas. These are classified according to their demographic characteristics in the 2011 census Output Area Classification (OAC). For more details, please see Chapter 3, Section 3.4.4. The provenance graphs we have recorded are captured using census output area geometry and so we have an output area classification for them and can compare provenance for different OAC groupings. One of the central aims of the investigation in this section is therefore to ascertain whether and to what extent demographic classifications delineate separate groupings within our data. Do provenance measurements from different OAC supergroups represent different

populations with their own intrinsic variance are the provenance graphs in our study area demographically homogenous?

There are two approaches which can be used to identify differences between groups of observations: univariate and multivariate analysis. We could simply examine individual variables using univariate ANOVA to compare their means among the output area supergroups [248]. The number of variables in our dataset would make the interpretation of individual ANOVA results a difficult and convoluted task. The factors which drive variance in our data are also highly complex, and many are not directly measurable as we see in Chapter 7, Section 7.5. The existence of these latent variables suggests a multivariate approach which can account for differences in variance which may not be noticeable in individual variables [248], [249].

Multivariate Analysis Of VAriance (MANOVA). MANOVA is a technique which tests for differences in the combined means of several dependent variables. These variables are considered as a linear combination rather than individually as in ANOVA investigations. This has the potential to reveal latent differences between OAC supergroups which may elude univariate approaches. The MANOVA procedure is essentially an ANOVA on a linear combination of three or more dependent variables.

Assumptions. Data used for a MANOVA procedure should meet the following assumptions:

There Should Be an Adequate Sample Size in Each Group. The sample size should be larger than the number of dependent variables [248], [249]. Uneven sample sizes for each group of the independent variable (OAC classification) can also distort results, however the SPSS MANOVA procedure carries out corrections for this [249].

There Are No Univariate Outliers Within Groups. For the MANOVA procedure we need to inspect each group for within group outliers. A simple heuristic for doing this is the Tukey method, which can be carried out using the inspection of box plots. Tukey recommended that data points with values larger than 1.5 times the interquartile range (1.5 IQR) be regarded as outliers and values three times larger as extreme outliers [259]. Later research found the 1.5 IQR criterion to be too restrictive and recommended 2.2 IQR [260]. There are a lot of extreme but not anomalous values in our data, and so we use the 3 IQR criteria provided by SPSS. In a normal distribution 99.7% of all data points should not be flagged as outliers. In view of the variance in our data we adopt a common-sense rule that no more than 1% of the data will be treated as an outlier.

Because of the skewed characteristics of our data, most of the variables within groups are likely to have an unacceptable number of outliers. In this situation Tabachnik and Fidell [2] recommend transformation of skewed variables as a reasonable option. This is particularly true our data where the variable scale is less meaningful. We are more interested in establishing general

patterns based on OAC supergroups, and under these circumstances, transformation can often improve the results of analysis. This is regarded as particularly true in situations where some variables are highly skewed and others not [2]. Transformation strategies include log 10, square root and reflection techniques which can be carried out using functions in SPSS. The presence of outliers increases the risk of type I errors in MANOVA with uneven group sample sizes [249] After transformation, each distribution needs to be reassessed and if more than 1% of the data points remain flagged as outliers we propose to remove the variable from the analysis.

Remaining outliers require examination to understand why these values are so extreme. However even if such values are not obviously anomalous their presence is likely to have a disproportionate effect on the modelling and so the removal of a small number of data points is justified.

There Are No Multivariate Outliers. Multivariate outliers occur where individual data points have extreme combinations of values for dependent variables. To identify multivariate outliers we need to measure the position of a data point in multivariate space and then measure its distance from the centroid of the remaining variables in that multivariate space. This measurement is known as the Mahalanobis distance [249], [261]. In SPSS it can be calculated using the regression procedure [262]. The dataset must be split using the grouping variable and then and then the dependent variables used to predict the output area code which is a unique identifying variable for each data point [249], [262]. This has the side-effect of calculating the Mahalanobis distance for each variable. The larger the value the more unusual the data point is in multivariate terms. A cut-off value can be obtained by cross-referencing the degrees of freedom, i.e. number of dependent variables on a chi-square table. Mahalanobis distances greater than this cut-off value identify multivariate outliers. Unfortunately, examination of individual data points to diagnose the cause of multivariate outliers is impractical and the only solution is often to delete the offending data points. Although MANOVA has some robustness to multivariate outliers their presence is not desirable and can reduce the power of the procedure [262]. Another option is to repeat the procedure with and without the multivariate outliers to see what material effects they have on the results [263].

Normality. Although the MANOVA procedure assumes normal distributions within groups, these procedures have some degree of robustness to minor normality variations [249], [264]–[266] and so minor deviations from normality are acceptable. Normality can be assessed by inspection of normal QQ plots and histograms.

Homogeneity of Variance/Covariance Matrices. The MANOVA procedure assumes that the dependent variables have similar variances and covariances across groups. Where group sample

sizes are equal, MANOVA is fairly robust against violations of this assumption [249] however this is not the case with our data. In SPSS, the Box's M test tests the assumption that the covariance matrices are similar. If the Box's M test is significant then the assumption is violated. Box's M is extremely sensitive to deviations from normality and is also sensitive in large samples [248], [249]. This means that in our data it is likely to be significant even for very small deviations. MANOVA also has a certain degree of robustness to violations of this assumption when group sizes are larger than 30 [249]. Where Box's M is significant, a common recommendation is to require stricter significance criterion and to choose the more robust Pillai's Trace test statistic [249], [267], [268] (see MANOVA test statistics, below). Homogeneity of variance can also be examined with Levene's test which compares the variance of dependent variables and is significant if this assumption is violated. In that situation it is recommended to avoid the use of post-hoc ANOVA and to require stricter significance levels for MANOVA test statistics [249].

Dependent Variables Within Groups Should Have Some Linear Relationship. The recommended way to assess this by visual inspection of scatter plots [248], [262]. This would need to be carried out for all of the variables, separately for each group. This is not practical using SPSS, so instead the dataset was split by OAC group in the correlation matrix generated for each group. An absolute correlation value of 0.2 represents a weak linear relationship [269].

Multicollinearity. Although MANOVA requires some degree of linear relationship between dependent variables strong correlations are not desirable and highly correlating variables are better assessed using separate ANOVA procedures [249], [262]. Relationships between combinations of variables are also problematic. There are more complex procedures for detecting multicollinearity such as the one used in the previous section. In exploratory factor analysis, the VIF inflation method also has the advantage of providing communality values. Multicollinearity is more serious in factor analysis as it can distort factor loadings. For MANOVA we simply use a heuristic calculation of correlation coefficients as recommended in [249], [262]. Correlations with an absolute value in excess of .09 indicate problems with a variable which will need to be removed from the analysis.

MANOVA Test Statistics. In MANOVA, the null hypothesis is that there is no difference between any of the groups of the independent variable. Four commonly used multivariate test statistics are provided by SPSS, which can be used to evaluate it. They can be used as Cohen's D values to assess the effect size [269] and their associated *P* values can be used to test the null hypothesis.

At a high level, MANOVA works by linearly combining the dependent variables from each group of the independent variable and assessing the difference between those linear combinations, or variate. They also used to derive the MANOVA test statistics. A more detailed account of these

test statistics can be seen in Field [248]. *Pillai's Trace* is the sum of the proportion of explained variance for each variate, calculated using its eigenvalues. *Hotelling's T* is the sum of the eigenvalues for each variate, *Wilks' λ* is the product of the unexplained variance for each variate and *Roy's largest root* is the sum of the eigenvalues for the first variate, i.e. the one with the highest eigenvalues. These statistics have various strengths and weaknesses. Wilks' λ is generally preferred in large samples with even sample size and no assumption violations. Pillai's Trace is the most robust and is recommended for groups with uneven sample size and heterogenous variance [249], [265], [267], [268]

Post-hoc Analysis. The MANOVA procedure can show whether there are significant differences in a set of variables between groups of measurements. It cannot, however, provide insights as to which combinations of variables are different and in between which groups. We have detailed information about the characteristics of the different OAC supergroups from the ONS pen portraits document. Knowing which of the groups differ, and which combination of variables is the source of variation will provide insights into role of demographic factors in provenance variation that would address research question two.

There are post hoc procedures which can add more detail to a significant MANOVA result . These include supplemental ANOVA procedures which can highlight the role of individual variables. Unfortunately our data violates the assumption of homogeneity of variance/covariance matrices which contra-indicates the use of post hoc ANOVA procedures [249]. Discriminant function analysis is another post-hoc procedure which is more informative because it considers the role of groups of variables [248]. In contrast to MANOVA, which uses combinations of variables to identify differences between groups, discriminant function analysis uses those differences to predict group membership by generating discriminant functions. These functions are linear combinations of variables which maximise differences between groups. The results also provide variable loadings which can show the contribution each variable makes to a discriminant function [248], [249].

Chapter 4 Implementation

This chapter describes the data analysis pipeline used to carry out the experiments described in this thesis. The technique is related to methods of 'scraping' provenance from log files generated by an application as part of its instrumentation, such as [151], [155] . Some literature discusses provenance extraction for OpenStreetMap [58], [59], [61], [108]. However, these do not go into much implementation detail and only describe prototype applications conceived for their own research use cases. They do not produce interoperable provenance data that conforms to universally agreed standards. Our goal was to produce an application which ingests OSM XML history data and extracts the provenance information. This is then encoded using the W3C PROV-DM [112] and stored in a graph database from which provenance graphs can be extracted for analysis.

The nature of the extracted provenance depends on a capture policy defined by the queries which are used for the extraction. The provenance graphs are extracted as RDF files which are loaded into custom-built analysis applications which calculate the values of various provenance metrics. For this thesis we have used two applications for measuring metric values. One is built in Java and is designed to measure maturity metrics. This application can only measure graphs from OpenStreetMap provenance captured using the W3C PROV-DM. The other application is built in Python and deals with abstract and semi-abstract metrics using the network X graph analysis library [186]. It has a modular structure, such that the module which deals with semi-abstract metrics can measure RDF provenance graphs from any domain, so long as they respect the PROV-DM. The abstract metric module can measure any network graph.

4.1 Technical Background

In this section we discuss the rationale behind the development of the analysis pipeline for OSM data and provide an account of the main components.

4.1.1 RDF, Ontologies and OWL

The provenance log files ingested by this pipeline are transformed into RDF data which is inserted into a triple store. RDF (Resource Description Framework) is a W3C standard which defines a graph data model [270]. There are numerous syntaxes for encoding RDF documents, and in this thesis, we use RDF XML [271]. The data structure of RDF is common to all syntaxes, however. The basic building block is known as a triple, a data structure consisting of three components: the subject, predicate, and object.

The subject is a node represented by a URI which resolves to some resource or representation of it. In our case this is an OSM data primitive or version of it. The predicate is an edge represented by URI describing a relationship the subject has with the object, another node representing a resource. Resources can be real world objects and are represented by URIs which should resolve to a description of that object. If the resource is directly available via HTTP, then the URI should resolve directly to it. Otherwise it should resolve to a representation of that object.

Computers can reason over RDF data using sets of logical rules. These web ontologies are encoded using the RDF-based OWL language [272]. OWL ontologies represent knowledge about data using sets of axioms encoded in RDF which provide logical rules for a domain of discourse. This allows machine-based reasoners to make inferences using RDF data. The PROV-DM recommendation provides an OWL ontology called PROV-O, which provides rules for reasoning over provenance data and defines the various provenance data types and relationships [273]. The PROV-O ontology can also be used for checking the consistency and integrity of PROV-DM data.

The RDF specification also includes SPARQL, a query language designed to work with graph data [274]. SPARQL can query graphs by specifying subgraphs and then returning matching graph patterns. It can also be used to insert and update graphs according to a specified graph pattern. In our pipeline we use this to capture provenance graphs. Patterns defined in SPARQL form our provenance capture policies.

4.1.2 OSM Data

Data Files. OSM XML data are made available as 'planet dump' files from the OSM website. There are three types of file available: map data, history data and changeset data, which we describe in some detail below. The map and history files are very large and impractical for most purposes, so most researchers use files from Geofabrik, a geospatial technology and services company who specialise in OSM, and provide free geographic extracts of OSM data. Geographic extracts of all three types of OSM data can also be made with the Osmium command line tool.

Map Data. The OSM data model has been designed to be simple and flexible so that mappers all over the world can encode potentially unanticipated geographic features [275]. The building blocks are three primitive types, which in OSM XML data, are defined by an element:

- osm:Nodes provide positional information
- osm:Ways use osm:nodes to describe linear features
- osm:Relations use the other types to assemble more complex structures such as transport routes.

These elements have attribute metadata containing creator details, version number, a timestamp, and a reference (id number) to the changeset record of the edit session where it was created.

While the OSM primitives describe the geometry and topology of the map; arguably the most powerful feature of this model are the tags which can be applied to each primitive to provide semantic meaning. These consist of key/value pairs which provide semantics for the primitive describing how it should be interpreted and rendered. The tagging model is governed by convention and documentation rather than schema. The 'any tag you like' paradigm [143] reflects the global nature of OSM by providing enough flexibility to enable mapping of unforeseen geospatial features. There is no official XML schema for OSM data and the only thing that can be relied on in any OSM XML file is that primitives occur in blocks of osm:Nodes, osm:Ways and osm:Relations in that order.

History Data. OSM XML edit history comes in files with the extension .OSH. The structure of the file is the same as normal map data, except that every version of each element is included, ordered by version number, and deleted elements are still present. The elements have a boolean 'visible' flag which indicates whether that version of the element was created by a delete operation, in which case it will be set to 'false' on the latest version.

Changeset Data. Changesets are a structure describing an editing session, automatically created when a user edits the map, they record the number of edits, the software used, timestamps, source datasets, user id, imagery used, whether the user is a bot, and the bounds of the area edited. OSM publish a complete XML dump of all changesets.

4.1.3 GraphDB

Graph DB is a widely used commercial graph database (triple store) with an offering for community and educational use. GraphDB was formerly OWLIM [276] [16], a collection of RDF repositories that have been widely used in the commercial and research communities. It features integration with Apache JENA and has its own ontology and inference API. It is compatible with the RDF4J Java Library which allows Java code to interact with the database via repository objects. GraphDB has some advanced graph visualisation features, which can generate visual representations based on SPARQL CONSTRUCT queries. The GraphDB ontology API accepts OWL files via RDF4J as well as manual loading and has an inference engine which supports OWL EL, RL, QL and various other dialects such as owl-max, owl-horst as well as RDFS. It is also possible to load custom dialects [277].

4.2 OpenStreetMap Provenance Reconstruction and Modelling With XSLT

OpenStreetMap provenance in its raw form is available in an XML edit history document. The provenance in this XML dump is in an explicit provenance format. To extract provenance data in an internationally agreed interoperable standard, this history log can be converted to W3C PROV-DM data in RDF format using eXtensible Stylesheet Language Transformations (XSLT). XSLT is an XML-based declarative language designed to transform XML compliant documents into other XML formats. For example, it can be used to convert XML data into HTML for display as a web page [278]. XSLT stylesheets use XPATH expressions to reference elements in a source XML document. In XSLT processor then uses the style rules specified by the XSLT document to generate new XML output [279]. We use the XSLT stylesheets within a purpose-built Java application . In the next section we discuss the design of this stylesheet which is used to model the provenance data in the triple store.

4.2.1 Modelling the Data

One of the central issues in provenance reconstruction from an edit history is how to model and record derivations: the sequential series of versions of an entity. Bowers et al [129] examined this in scientific workflows and using computer programming concepts of control and data dependency. Using these ideas, they produced some definitions of the notion that worked well in the domain of scientific workflows. The W3C also have a definition for the web. Section 2.1.2 of the PROV-DM recommendation [112] defines derivation as ***"...a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.***"

The PROV documentation provides no explicit guidance as to how sequential versions of an entity should be modelled. It does observe however, that while seemingly simple, there are subtle nuances to the concept [112]. For instance, an activity must be involved in the generation of a derived version, which must use the previous version/s. Even so, we cannot assume that an entity generated by an activity is derived from all the entities is used. The PROV-DM documentation leaves it up to the developer to devise a strategy to identify an influence between versions, which is required, along with usage and generation to describe derivation. PROV-DM and the PROV-O ontology supply various structures which can be used [112]. Influence between versions is expressed with the *prov:wasInfluencedBy* object property, which describes a range of relationships which exist in a PROV graph, including a child property: *prov:wasDerivedFrom*. Further down the hierarchy is *prov:wasRevisionOf*, a sub property that would seem to best describe feature derivations in OSM History. This is at the bottom of the object property hierarchy and so allows the inference of triples

with the super properties without them having to be explicitly declared, so this is the approach we adopt.

Table 5: Provenance Attributes in OSM History Data

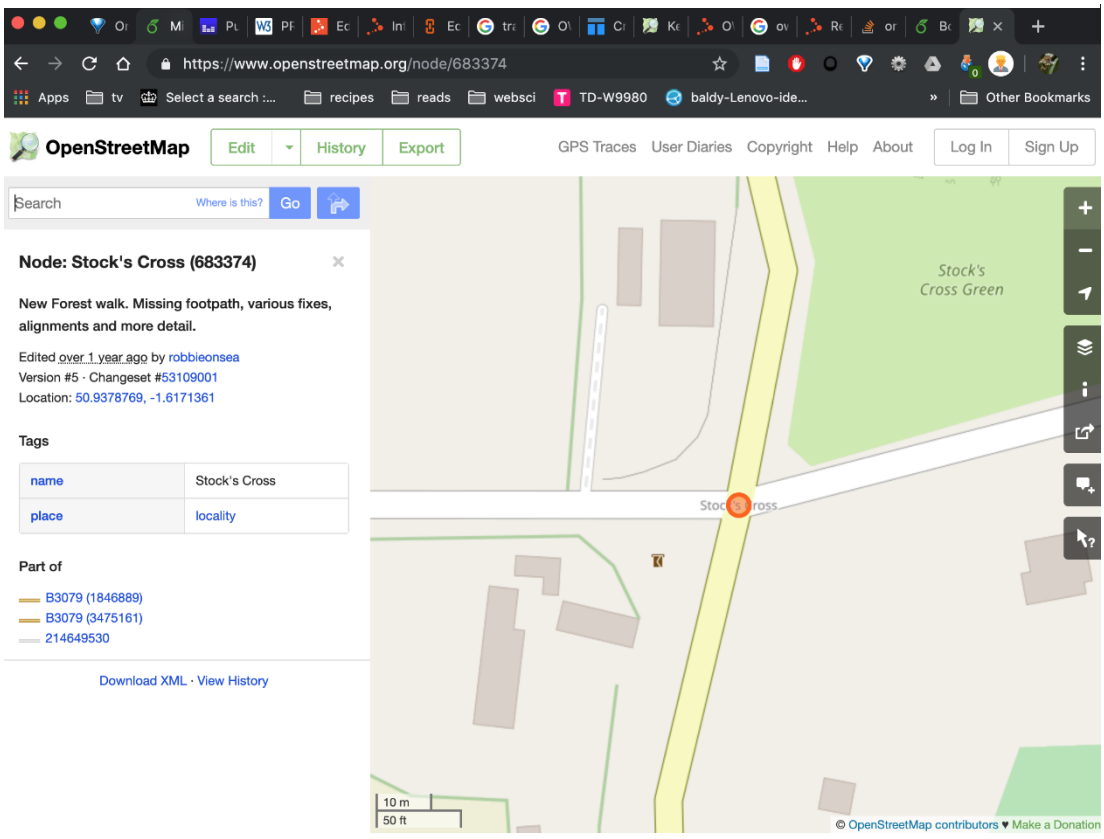
Attribute	Value	PROV-O Class
Id	the feature id number	prov:Entity
uid	the user account id number	prov:Agent
changeset	the changeset id number	prov:Activity
version	the version number	n/a

Other implicit provenance information can be found in attributes to the OSM feature elements including the id, creator, changeset, and version attributes in Table 5. These are used to mint the URIs used in the subjects and objects of the provenance triples.

The XSLT Script: Main Data Transformation. In designing the XSLT script, we originally hoped to use as little domain specific XML markup as possible, to allow use of the same XSLT with other datasets. This approach could use include statements to dynamically add any unavoidable domain specific code, while keeping the bulk of the XSLT code domain agnostic. The resultant RDF/XML data would contain a lot of literal values as subjects of the triples, and we would be left having to write some complex and domain specific SPARQL queries or OWL axioms to get RDF Provenance graphs. To simplify matters we built the XSLT specifically to reference elements in an OSM Edit history file.

The XSLT script generates a unique URI for each feature version. In most situations, RDF requires a unique URI for each subject and object of a triple unless they are literal values. These URIs also represent the vertices of the provenance graph in the triple store. They should also dereference to some meaningful representation of each version, and these representations must also have some logical consistency that the data reflects. The work of minting a dereferenceable URI of a live OSM feature has already been done for us by OSM, where all live map features already have URIs. Using the prefix 'http://www.openstreetmap.org' and appending the primitive type, plus the id, gives us a URI such as: ***http://www.openstreetmap.org/node/683374***, which references to that OSM location at OSM.org (see Figure 4).

Figure 4: Dereferencing a URI in OSM: http://www.openstreetmap.org/node/683374



To generate RDF which describes a sequence of versions, we need to generate a unique URI for each version. We might do this by adding the version number, producing: http://www.openstreetmap.org/node/683374v2.

Unfortunately, this URI produces an HTTP 404 error, and as the OSM.org domain is not under our control, we cannot do much about this. To generate a dereferenceable URI for each version we need to use a different URI scheme, one that is not ambiguous and is logically consistent with reality. We are representing the provenance of a feature that has a URI at OSM.org and need to represent all the previous versions of it. In our world, these versions are also important entities which form vertices in our graph data and so should all have unique URIs. However, they should also be distinguished from the live feature in OSM as they are elements in the history of a feature and not part of the live map.

This problem of how to semantically describe versions as entities that are both part of, and distinct from the thing they are versions of, has previously arisen in geographic data. Lohfink and McPhee [280] looked at version histories of Ordnance Survey administrative boundary data and used

RDF Containers, specifically the RDF sequence (rdf:seq) as a structure for holding the version history. In RDF Sequences, each version is represented by a *blank node* and contained within the rdf:seq. These constructs are a convention, and the sequence has no attached behaviour or validation constraints. Its purpose is to indicate to a human that the contents are an ordered list.

Blank nodes are a construct within an RDF graph that allow vertices to have no identity. This allows the simple linkage of complex information and abstract objects without having provide their identity. Using blank nodes to represent versions in our provenance model poses practical problems. GraphDB mints its own URIs to handle blank nodes. As we can not predict what these will be it becomes difficult to access them using queries. Adding new data also becomes problematic because there is no guarantee that duplicate triples will be overwritten, or that subjects and objects in new triples will match existing ones. As the main use-case for this dataset is to query subgraphs in a highly linked data environment, readability is not a primary concern. Therefore, there is little advantage in Lohfink and McPhee's approach. Instead, we mint separate sequential URIs using the feature id and version number.

To provide non-ambiguous and unique URIs for OSM feature versions, we created OSMP, an OWL ontology. This uses a different namespace and gives us the option to use this domain to publish web documents describing our versioned resources. Because each feature id is not unique in the OSM database and is part of a compound key formed by the feature type and feature id, we combine these along with the version attribute value, so that version 3 of our example osm:node from Figure 4 would have the URI:

'http://www.semanticweb.org/bernardroper/ontologies/2018/7/osmp#node683374v3'. We then link to the previous version using a prov:wasDerivedFrom predicate pointing at the same URI made with a decremented version number until we reach version 1.

Each OSM primitive type has a class in OSMP which sub-classes the appropriate classes in PROV-O, as shown in Table 6. This allows the osm:Way type to have member osm:Nodes.

Table 6: PROV-O, OSMP and OSM classes

OSM Type	OSMP Class	PROV-O Class
osm:Node	osmp:Node	prov:Entity
osm:Way	osmp:Way	prov:Collection

The XSLT parses the XML and generates an *rdf:about* statement for each version it finds, using the URI generated as above. The element's attributes are then converted to triples. These are not currently directly used, but may prove useful in future experiments, particularly in identifying changes between versions. These attributes can then either be expressed as child elements as in Figure 5, or as attributes as in Figure 6.

Figure 5: RDF With Child Elements

```
<rdf:Description rdf:about="osm:node254429">
    <osm:timestamp>2012-03-07T15:07:12Z</osm:timestamp>
    <osm:uid>14320</osm:uid>
    <osm:user>Michael Mouse</osm:user>
    <osm:changeset>10899959</osm:changeset>
    <osm:lat>0.9199448</osm:lat>
    <osm:lon>-1.3953573</osm:lon>
</rdf:Description>
```

Using attributes results in a smaller result file size. Processing a single OSM Node into an RDF/XML statement with child elements results in a file size of 514 kb, whereas using attributes produces 414 kb of data. This difference will be much more significant when processing thousands of nodes.

During processing, the ***id*** attributes seen in Table 5 are stored as XSLT variables and used to write other RDF statements which produce a set of RDF/XML triples such as those in Figure 7. The osmp:nd elements (14 removed for brevity) are references to member nodes, as this is an osm:Way element. osmp:nd is also an object property in the OSMP ontology, which is a sub-property of prov:hadMember, making the Nodes members of the PROV-O prov:collection class. The Tag

Figure 6: RDF With Attributes

```
<rdf:Description rdf:about="osm:node254429"
                 osm:timestamp="2012-03-07T15:07:12Z"
                 osm:uid="14320"
                 osm:user="Michael Mouse"
                 osm:changeset="10899959"
                 osm:lat="0.9199448"
                 osm:lon="-1.3953573"/>
```

elements have been generated from OSM tags and dereference to the tag description on the OSM Wiki. This will allow extensions to this work to include tags as prov:Entities and generate provenance for them. The prov:wasAttributedTo and prov:wasGeneratedBy predicates have subjects that dereference to a changeset display and an osm user account (not viewable without OSM administrator permissions) on OSM.org.

As the RDF feature is generated, so is the second element in Figure 7. This is another rdf:about statement that creates a prov:Agent vertex identified by a URI which is also the object of the prov:wasAttributedTo statement. Where there are multiple features edited by one agent, this record will overwrite any duplicates in the triple store, so that all the prov:wasAttributedTo predicates point at only one prov:Agent for each *userId* in the history data.

The final step in this process is to add a dc:isVersionOf triple to link all the versions to the current feature on the map. This is rather difficult to achieve in XSLT, because we have no way of knowing which version number is the latest one until all the versions have been processed. This triple is added in post processing using a SPARQL INSERT query, which uses a sub-query to find the latest version, i.e. highest version number.

Figure 7: A *prov:Entity* element in RDF describing an OSM Way and its responsible *prov:Agent*

```
<rdf:Description rdf:about="http://www.semanticweb.org/bernardroper/ontologies/2018/7/osmp#way3475161v14"
                 osm:id="3475161"
                 osm:version="14"
                 osm:timestamp="2012-11-20T11:41:43Z"
                 osm:uid="397946"
                 osm:user="untitled"
                 osm:changeset="13943478"
                 osm:visible="true">
    <osmp:nd rdf:resource="http://www.openstreetmap.org/node/683374"/>
    <osmp:nd rdf:resource="http://www.openstreetmap.org/node/1077731266"/>
    <osmp:tag>:ref\%3DB3079</osmp:tag>
    <osmp:tag>:highway\%3Dsecondary</osmp:tag>
    <prov:wasAttributedTo rdf:resource="http://www.openstreetmap.org/users/397946"/>
    <osmp:hasVersionNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#int"> 14 </osmp:hasVersionNumber>
    <rdf:type rdf:resource="https://wiki.openstreetmap.org/wiki/Way"/>
    <prov:wasRevisionOf rdf:resource="http://www.semanticweb.org/bernardroper/ontologies/2018/7/osmp#way3475161v13"/>
    <prov:wasGeneratedBy rdf:resource="http://www.openstreetmap.org/changeset/13943478"/>
    <dc:isVersionOf rdf:resource="http://www.openstreetmap.org/way/3475161"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.openstreetmap.org/users/397946">
    <rdf:type rdf:resource="http://www.w3.org/ns/prov#Agent"/>
</rdf:Description>
```

Adding Changeset Data. One of the strengths of RDF is that it is easy to add triples from other datasets. Once the data is in RDF it is compatible with other RDF. This allows us to seamlessly combine changeset data with edit history data in a way we could not in its XML form. The changeset data published by OSM (contains some useful provenance, but the schema of the file is different, making it more difficult to query information from both files when in xml form. Using XSLT to

transform this into RDF/XML triples overcomes this problem and adds more information to the provenance graph.

A separate XSLT script is used, which creates an RDF about statement for each changeset using the changeset id to generate a URI which dereferences on osm.org to a map tile representing the spatial extent of the changeset. A prov:wasAssociatedWith predicate is generated, pointing at a URI which resolves to the user account page on OSM.org, and connects with the prov:Agent records.

Some of the information in a changeset is in the form of tags and two types are processed by the XSLT: *Source* tags, which are used to generate a prov:Entity representing any external datasets used during the edit session, and *Created By* tags, which are used to generate a prov:SoftwareAgent representing the software which was used to create the changeset. This is linked to the changeset with a *prov:wasAssociatedWith* predicate. Using the content of these tags presented some problems, as they use user generated input.

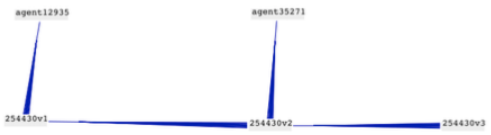
Initial Data Modelling Experiments. To explore our XSLT approach, an initial transformation was conducted using a specimen .OSH data file. This file was based on real OSM data which was anonymised and had a smaller number of elements in the same order and layout as a real .OSH file. It contained seven node, two way and one relation elements. The file contained the version history for each of these, resulting in 43 node, five way and two relation elements. An XSLT stylesheet was created to transform the OSM XML into RDF/XML, adding an import statement to add the PROV-O ontology and some axioms to map the OSM primitives to PROV-DM classes. This experiment modelled derivation and attribution relationships in RDF. Each version was assigned a unique URI created using a combination of its version number and feature id attribute. The derivations, expressed with prov:was:derivedFrom relationships, were added with a URI generated in the same way, but with decremented version numbers as an object of the triple. Attributions for agents (prov:wasAssociatedWith) were created using a URI containing the OSM user id attribute from the feature, which identifies the user account which created that edit.

The result file was imported into the Protégé OWL editor along with the PROV-O ontology. Running SPARQL queries in Protégé produced the results shown in Figure 8. These are very two different graphs produced from the edit history of the same node. They illustrate the effects of using SPARQL to define provenance capture policy. Figure 9 shows the results of running the reasoner in Protégé, showing the generation of extra triples. Careful choice of provenance expression in the XSLT transformation allows the inference of these extra triples which maximises the options for defining different expressions of provenance using our SPARQL based capture policy.

Figure 8: Two SPARQL Queries and Their Resulting Graphs

<table>
<tr>
<td>

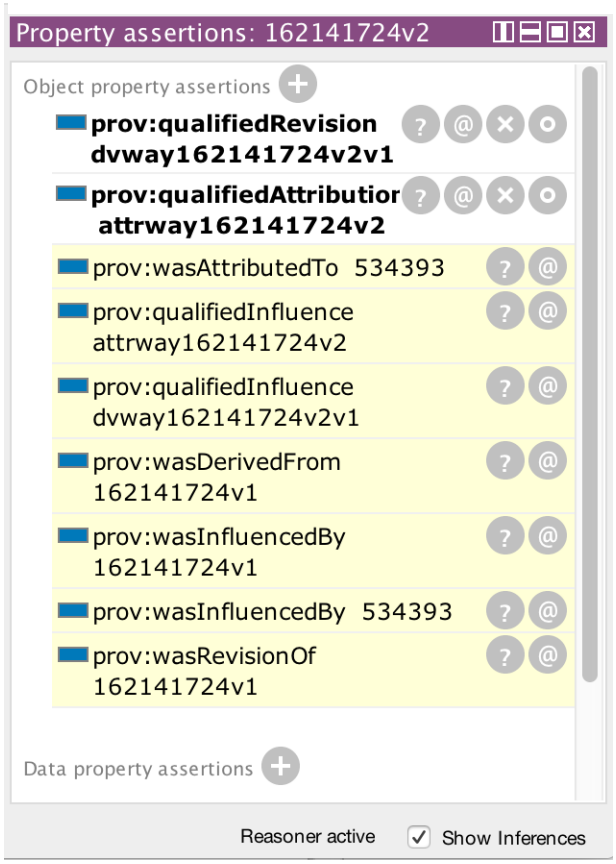
```
CONSTRUCT {
?version prov:wasDerivedFrom ?entity.
?entity prov:wasAttributedTo ?agent.
}
WHERE {
?entity provanalytics:versionOf osm:254430.
?version prov:wasDerivedFrom ?entity.
?entity prov:wasAttributedTo ?agent
}
```

</td>
<td> 
</td>
</tr>
<tr>
<td>

```
CONSTRUCT {
?entity prov:qualifiedAttribution ?attr.
?attr prov:entity ?entity.
?attr prov:agent ?agent.
?version prov:qualifiedRevision ?rev.
?rev prov:entity ?entity.
}
WHERE{
?entity provanalytics:versionOf osm:254430.
?entity prov:qualifiedAttribution ?attr.
?attr prov:agent ?agent.
?version prov:qualifiedRevision ?rev.
?rev prov:entity ?entity.
}
```

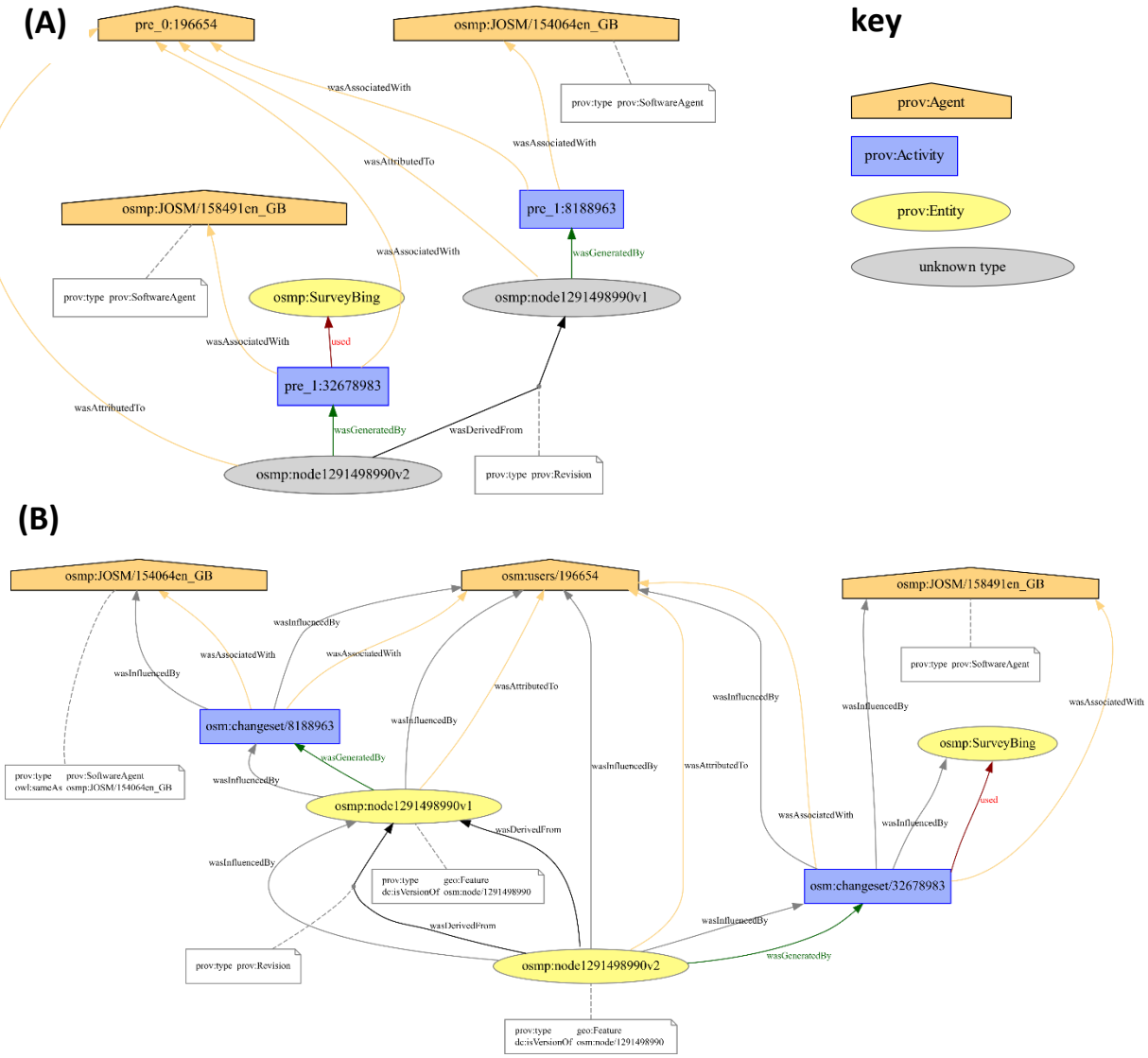
</td>
<td> 
</td>
</tr>
</table>

Figure 9: Running the reasoner in Protégé



In GraphDB, a built in RDFS reasoner infers extra triples. This simplifies the requirements of the XSLT transformation and maximises the options for provenance capture using SPARQL. Figure 10 shows an example from the GraphDB triple store after the transformation is complete. The two graphs shown, A and B are the provenance of a single OSM node which has two versions. The second version is the result of a revision created in a changeset that used Bing imagery. In graph B the PROV-DM types of the node versions have been inferred, along with extra prov:wasInfluencedBy triples.

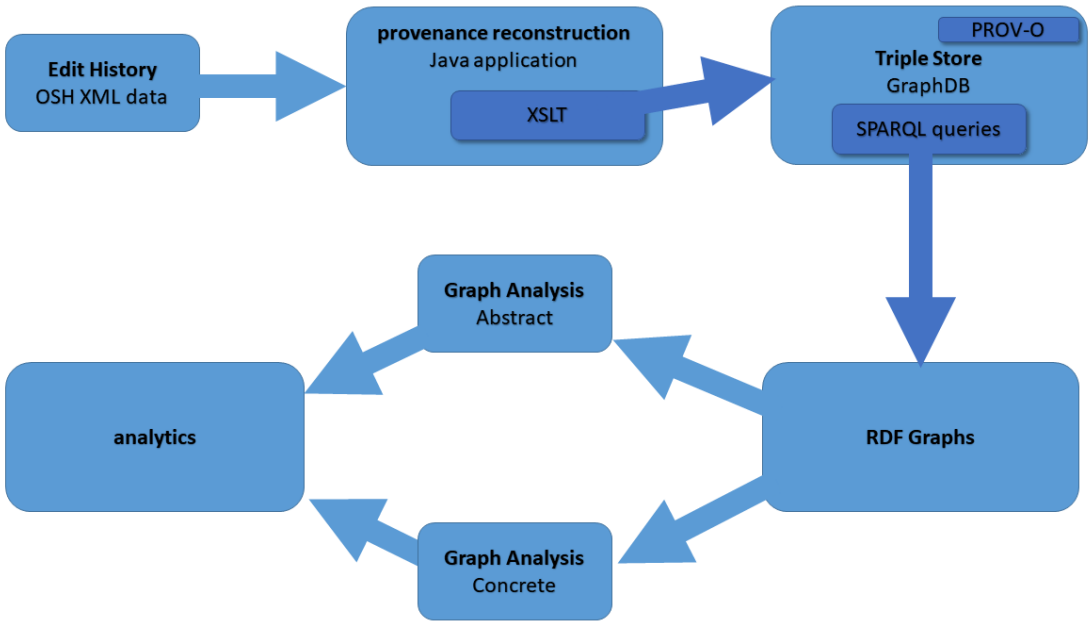
Figure 10: Provenance Graphs of a Single Osm:Node (Graph A). Graph B Shows the Effect of Running Graphdb's Rdf Reasoner (Image Produced Using the Prov Store at Openprovenance.org)



4.3 A Data Processing Pipeline

Figure 11 gives an overview of the data pipeline, which uses OSM history and changeset data. XSLT is used to transform the OSM data into an RDF/XML graph. This is imported to a Triple store (GraphDB) along with the PROV-O ontology and a mapping ontology called OSMP, which contains axioms assigning the OSM primitives to PROV-O classes, which are used to enrich the data set by entailing more triples to generate a comprehensive provenance graph containing an explicit representation of the implicit provenance information contained in the OSM edit history. After transformation and insertion to GraphDB, the triple store contains 28,057,135 explicit triple statements. The reasoner infers another 80,280,593, resulting in 108,337,728 triples in total.

Figure 11: Data Pipeline Overview



4.3.1 Data Processing

Geographic extracts of OSM history and changeset data are very large files, so we used the O*smium* command line tool to perform geographic extracts from these files using polygons generated by the OSM interface. To process the XSLT and OSM data into a transformed RDF graph, we built a Java application using the SAXON XSLT library [281] . This applies the transformation and outputs a stream of RDF XML data. This result can then either be output to a file or directly inserted into GraphDB repository. We opted for output to file and designed Java classes which output a series

of RDF/XML files created from the Saxon output stream. This enabled the program to fail gracefully in the event of an error and then resume processing.

Data Sanitisation. Software using the API can have illegal XML characters in its title, as can the names of source datasets. These are referenced by changesets and so appear in tags. It is possible to use regex in the XSLT code, but because XSLT is an executable form of XML it cannot run with any illegal characters, even inside a regex expression, so using regex inside an XSLT program to search for the ">" character causes the XSLT to fail. In OSM data, these XML characters are converted to XML entities, but the strict adherence to the XML standard by the XSLT processor means that these entities are treated as literal characters. In the end it was necessary to resort to *find and replace* in a text editor (VIM) to sanitise the OSH history file before processing.

GraphDB. The resulting RDF datafiles are imported into GraphDB along with the PROV-O and OSMP ontologies. GraphDB has an internal reasoner which uses the ontologies to infer extra triples resulting in a comprehensive provenance graph database. Provenance subgraphs relating to specific features can then be extracted for analysis using SPARQL queries. GraphDB also supports the GeoSPARQL protocol which allows geographic SPARQL queries using WKT geometry to perform geographic provenance extracts. Using GeoSPARQL along with output area geometries from the ONS we are extract provenance graphs for OSM coverage within specific census output areas.

4.3.2 Measurement

Graph data is extracted from the triple store (GraphDB) as a set of RDF files containing provenance subgraphs defined by SPARQL CONSTRUCT queries. For our analysis geometry files defining polygons were used to define GeoSPARQL queries, so that each graph is a geographic extract. We built two applications to perform the measurements: A Python program to perform the abstract and semi abstract measurements, and a Java application to conduct the maturity analysis.

The python program was based on the network X library and the PROV Python library [282], locally modified to support software agents. They calculate values for the abstract and semiabstract measurements described in Chapter 3, Section 3.3. Processing overheads for some of the graph theoretic measurements were too demanding for desktop computing in some of the larger provenance graphs and so these calculations were carried out on the University of Southampton's Iridis4 high-performance computing cluster.

We built a Java program to calculate the maturity metrics. This used the Apache JENA library to manipulate and query RDF data, and our own purpose-built metrics modeller. This loaded graph

was loaded into an Apache JENA RDF model. A set of SPARQL queries was run on the models which generated a set of base descriptive metrics. These were then used to evaluate the maturity metrics.

4.4 Summary

This data analysis pipeline has successfully ingested large volumes of OpenStreetMap data and then extracted and measured provenance graphs. It allows the specification of different provenance capture policies for different analytics use cases. It has a modular design, so that many of the software components can easily be reused and adapted for other types of graph and provenance analysis. The end-product of this procedure is a set of values for provenance metrics for OpenStreetMap. An important by-product is OpenStreetMap data in an interoperable format which can be used for the study of OpenStreetMap in a variety of use cases and adapted to other platforms.

Chapter 5 Interpreting Provenance Networks

To address research question 2: “what insights can be demonstrated about map contribution behaviour and the mapped environment using provenance from OpenStreetMap?” We need to understand what drives variation in provenance measurements between individual graphs. In the previous chapter we examined how variations are represented on thematic maps to study their spatial patterns and relationships with the human and physical geography of the OSM coverage area. In this chapter we focus on the network graphs with a visual examination of individual output area provenance graphs and their abstract graph theoretic metrics. We compare graphs with high and low values of specific provenance measurements alongside the map coverage they represent, examining individual vertices in graphs to build up a picture of how their characteristics interact with provenance measurements to drive variation.

5.1 The Graph Analytics Spectrum

Graph analysis applications can be thought of as existing on a spectrum. One end takes a holistic view of the data. Graphs are distinguished by using sets of metrics as “graph fingerprints”, either to identify anomalies or to divide graphs into groups in a classification or clustering task. Here we consider a dataset in its entirety, using all the available information to identify patterns or classify data. At the other end of the spectrum are more atomistic approaches, graph metrics are interpreted using expert domain knowledge to draw more specific, detailed insights into the real-world processes represented in the graph data, and there is more emphasis on the analysis of individual graphs, often visually.

At the atomistic end, Jamieson et al [180] analysed the network properties of mind maps created by students after a learning activity and compared them with mind maps produced by an expert in the field, using graph metrics to assess the state of the student’s knowledge after the activity. Here, classification is not the main task. Their objective was to understand how the mind maps change over time and whether their similarity to expert mind maps increases. Domain knowledge is used to draw insights, e.g. that topics with highly connected concepts produce graphs with high density values. The graphs being examined are relatively small and simple, with well understood vertices and edges that allow visual interpretation.

Further along the spectrum, Huynh et al’s work on provenance network analytics [54], and Gomes et al’s spam email detection studies [181] are more atomistic approaches used for classification tasks. Gomes et al’s graphs are structurally quite simple, but too large for visual inspection, while the graphs used by Huynh et al in provenance network analytics can provide

supplementary insights from inspection of some individual examples. The nature of the classification task requires consideration of the entire dataset and there is less emphasis on the examination of graphs at an individual level because of the volume of data involved.

Applications in neuroscience tend to be at the holistic extreme. Studies such Deuker et al [183], using magnetoencephalogram data to generate network graphs from electrical brain activity, potentially produce thousands of network graphs over a short period, with vertices representing positions on the brain surface and edges representing voltage difference. The aim was to establish that graph metrics are reliable indicators of brain activity for longitudinal studies of neurological change. Another study of multiple sclerosis (MS) patients [283] uses MRI imaging to model neural connectivity within the brain to classify patients into categories of disease progression. In both these studies there is little or no consideration of individual graphs because of the volume the data and nature of the available knowledge.

The OpenStreetMap output area provenance graph data sits at a midpoint between the two extremes. The 1178 graphs in our dataset vary in size between about 200 vertices and several thousand. Most are too large and complex for rapid and easy visual assessment of many graphs, but individual graphs can be visualised on a powerful desktop computer. There is a lot of existing domain knowledge from research into OSM and VGI, and because the provenance graphs are encoded in RDF format, each vertex has a URI which can be used to identify changesets, osm:Ways and osm:Nodes giving access to additional information via the OpenStreetMap API and a web browser. The statistical analysis in Chapter 7 and spatial interpretation provided in Chapter 6 are assessing the wider dataset. The careful examination of a selection of individual graphs provided here supplements those investigations to address research question 2 by shedding light on the structure of individual graphs and exemplifying some of the drivers of variation.

5.1.1 Domain Knowledge: Known Drivers of Variation in Contributor Activity

Contributor Driven Variation. OpenStreetMap contributors are a source of variation in the map coverage, and their activity has been an active research topic (see Chapter 2, Section 2.2.2). For example, variations in completeness and editing intensity have been found in numerous studies. Gender [27], [28], nationality [34], [65], [68] , socio-economic status [33], [68] and other demographic characteristics have also been shown to influence the types of feature and location that are mapped. Studies have profiled contributors often forming a categorisation based on level of experience or expertise [26], [65], [71], [284], or types of editing activity [72], [285] often with a view to predicting aspects of data quality or edit intensity. They have been shown to exhibit various individual characteristics which are covered in more detail in Chapter 2, Section 2.2.2.

Environment Driven Variation. Another driver for variation in the provenance metric values is the nature of the features being mapped, i.e. the physical environment. Variations in the built and natural environment will be reflected in provenance network graphs because of the structural characteristics of the OSM primitives used to represent them. This is complicated by the interplay between environmental and geodemographic properties of the region being mapped. For instance, we know that contributors have individual preferences for mapping specific features and regions and in Chapter 6, we found relationships between the physical and built environment and our provenance variables.

Our units of analysis are census output areas which are defined using a zoning algorithm designed to optimise demographic homogeneity within the output area by population size, number of households and housing characteristics. The occurrence of many physical features, particularly those of the built environment, will reflect the community living in those areas, which will in turn affect the provenance network properties. For example, a feature representing a small business premises is likely to change tags more frequently than a church and this will have a distinct effect on the provenance graph. The incidence of different feature types has been shown to vary demographically, e.g. Venerandi et al [286] used the presence of specific tags such as “golf course”, “gastropub” or “car wash” to compute urban deprivation indices.

5.2 Graph Theoretic Measurements

To better understand the nature of our metrics, in this section we explore the factors which drive high and low measurement values. For several metrics we look at provenance graphs with some of the highest and some of the lowest values (disregarding extreme outliers). For each case we also examine the OpenStreetMap coverage. This was complicated by the fact that the provenance data we use was downloaded in February 2020 and OpenStreetMap is being actively edited. This activity has been even more pronounced owing to the Covid-19 pandemic and the large number of people living under lockdown conditions. To make certain that the visual representation of the map is as it was when the provenance graph was recorded, we queried the data via the OpenStreetMap interface and examined the major changesets in the area. Any output areas which had seen active editing since the download date were disregarded. In some cases, edits which did not affect the visual representation of the map, such as the addition of some tags, were overlooked. In each case it was still possible to find areas with extreme values which have not undergone significant editing. Other graphs were also examined in detail, but the illustrations are all from OpenStreetMap content which has undergone no significant editing since the download of the provenance data.

The provenance graphs themselves were converted from RDF format into GraphML and then loaded into a graph visualisation program called Cytoscape [232]. This has a range of capabilities, including graph theoretic measurements, nearest neighbour selection and a range of colouring and visualisation options as well as several graph layout algorithms to facilitate exploration of the network graph. The graph vertices have been styled using the legend in Figure 12. The colours and shapes have been chosen for their similarity to the W3C PROV-DM standard. The following account describes some of the major network metrics and insights obtained from looking at several provenance graphs.

Figure 12: Legend Graph Vertices

5.2.1 Degree Distributions

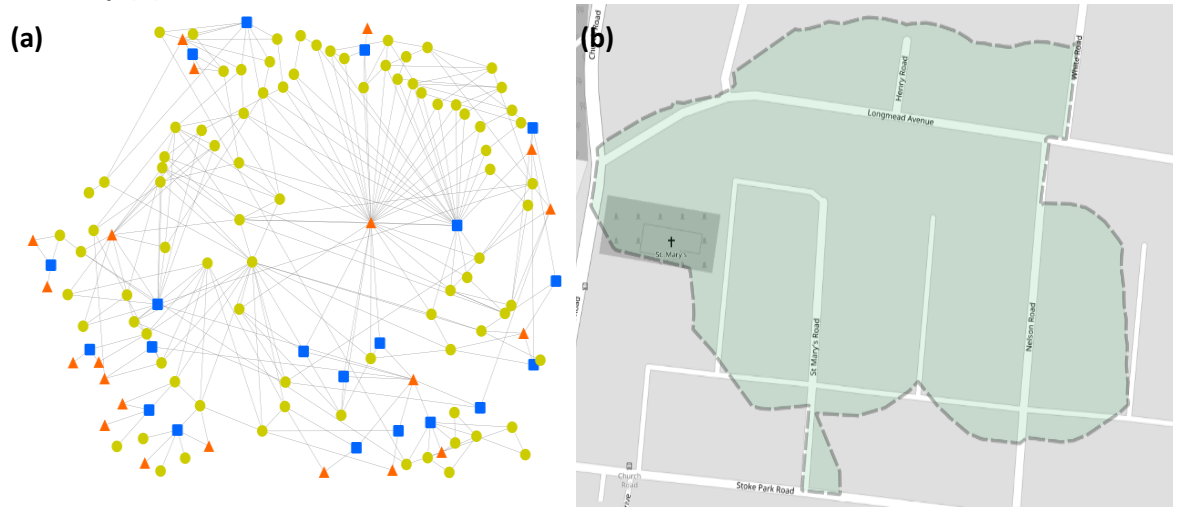
A common feature of real-world networks, including provenance graphs, is that their degree distributions follow a power law [56]. This means there are many vertices with small degrees and few with large degrees and the distribution increases exponentially. The power law metric we describe in this section is derived using the NetworkX library in Python [186]. The power law algorithm fits a power law to the dataset and the derived metric is the exponent of that power law [56].

Entity Power Law Exponent.

High Values. Figure 13 shows an output area and its associated provenance graph, which has a high entity power law exponent, caused by three prov:Entities (circles) with a degree significantly higher than the rest. Many of the prov:Entities in this graph have a degree of one and are attached with the prov:association relationship to a changeset (blue square shape). These are source datasets from which the edit was derived and represent satellite imagery from which the feature was traced. As can be seen from the OpenStreetMap rendering on the right, this area is

sparsely mapped, containing a simple street network with no building footprints, and a church with crude building outline and partially mapped churchyard.

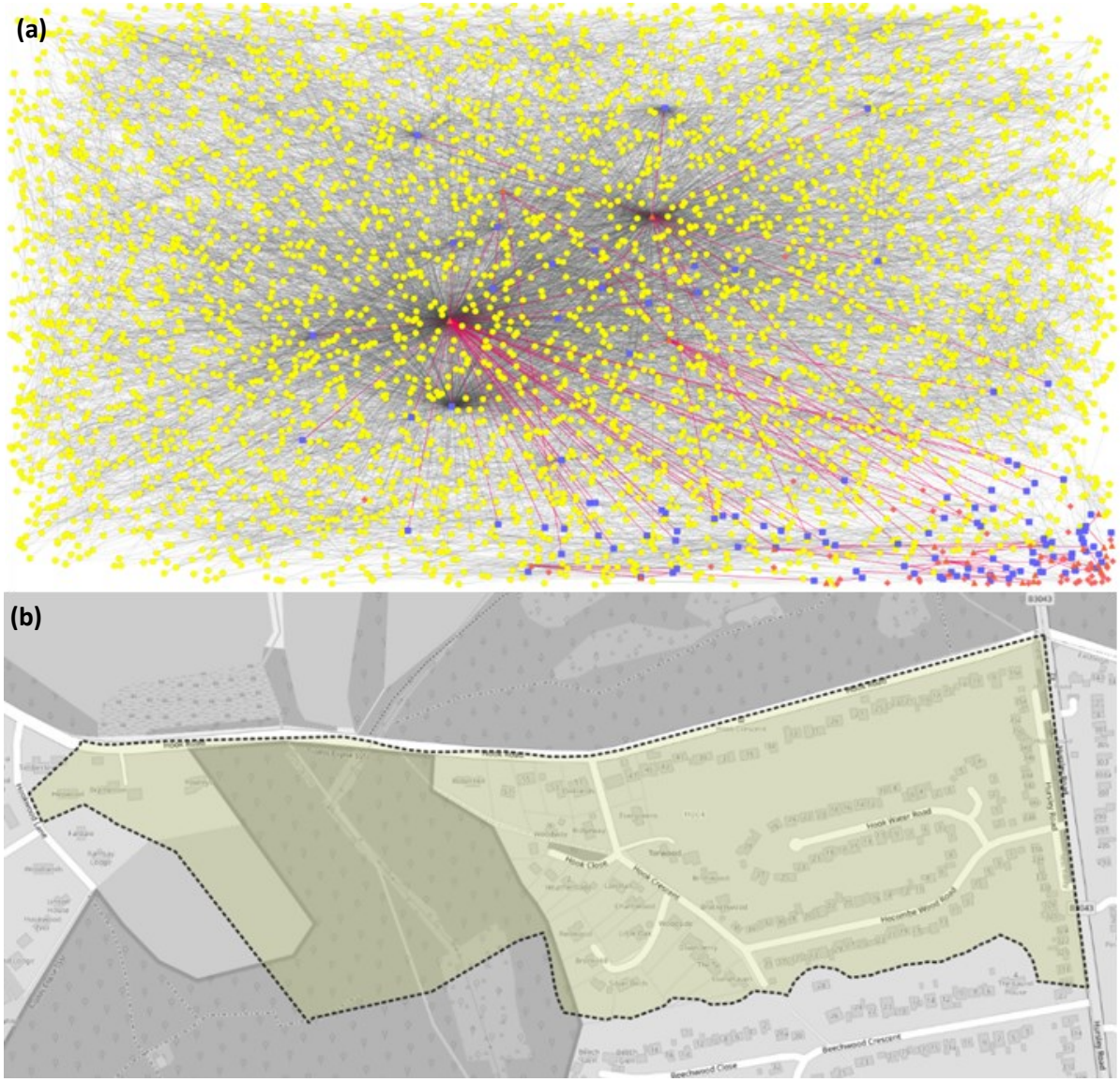
Figure 13: Provenance Graph (a) With High Entity Power Law Exponent and its OSM Output Area Map (b)



The contrast between high and low degree osm:nodes which gives rise to this high power law exponent is due to the small size and sparse mapping of the output area, and small side-streets being the prevalent feature. The highest degree osm:node is one which joins the largest side-street onto a main road and is shared with both features. The main road clips the edge of the output area and has been edited many times, resulting in a much larger degree. Other intersection nodes also have larger degree, and these cause the degree distribution to increase exponentially. Their effect is exaggerated by the paucity of other mapped features

Low Values. An output area with a low entity power law exponent, as seen in Figure 14, typically has a larger provenance graph containing more prov:Entities of varying degrees, with few or no extremes The example shown in Figure 14 contains just over 4000 prov:Entities and the map shows that the OSM coverage has a high level of completeness with accurate building footprints including house names and numbers and range of other features including an SSSI boundary, powerlines, garden polygons, a detailed street network and major roads, all of which are named. This coverage was mainly produced by a prolific mapper who has generated a lot of changesets in this area over many years and is likely to be local. The amount of detail in the map has lessened the impact of single features and produced a much more even degree distribution.

Figure 14: Provenance Graph With Low Entity Power Law Exponent (a) and OSM Map (b)



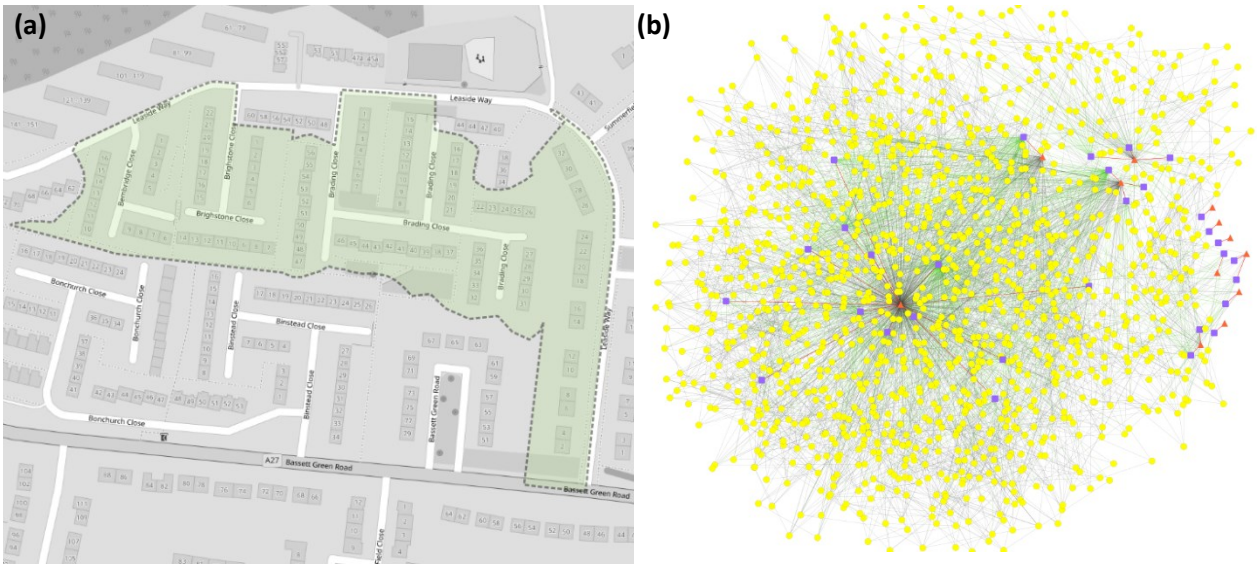
Summary.

- Low values of this variable can be associated with large, complex provenance graphs and detailed, mature OSM coverage. With a detailed OSM coverage, node degrees are generally higher and more evenly distributed, dampening the effect of extreme values.
- High values of this metric are often small sparsely mapped output areas including only a few simple street outlines. One or two nodes from heavily edited features such as those shared with road junctions skew the distribution and increase the power law exponent.

Activity Power Law Exponent.

High Values. A graph with a high activity power law exponent is shown in Figure 15. Edges between agents and activities have been coloured red and edges connected to activities have been coloured green to show the Activity degree.

Figure 15: OSM Map (a) and Provenance Graph With High Activity Power Law Exponent (b)



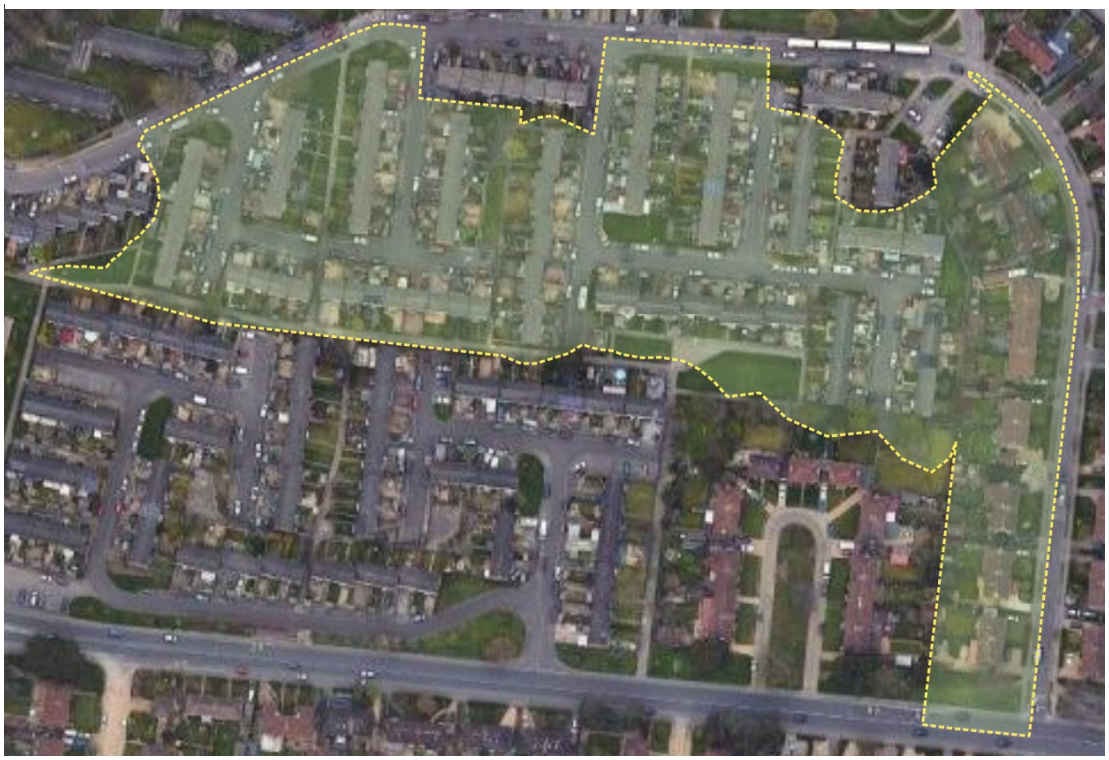
The characteristic long tailed distribution online contributions profile is apparent here, with much of the editing done by one contributor. They have done much of their editing in a single changeset as have the two others who have contributed significantly. Several other people have made minor edits in single changesets. This means most of the prov:Activities have small degrees with most editing done in only three, which skews the distribution resulting a high power law exponent.

The OSM coverage of this output area is of residential streets and semi-detached houses. Building footprints seem to be complete and accurate and all houses are numbered, and streets named. Other than some footpaths, no other features are present in the OSM coverage although inspection of satellite imagery shown in Figure 16 reveals that garden polygons, trees and other minor tracks and streets could be added.

In essence, this provenance graph suggests intensive activity by a single contributor concentrating on mapping building footprints, with several minor edits by other contributors. Most of the edits appear to have taken place on few occasions. Most of the changesets in the graph are minor edits by several different contributors. The largest changesets are all generated by one contributor who has been mapping building footprints, and from inspection of their changesets it

seems that this is their speciality. Other contributors have added addresses and other less numerous features such as footpaths, which generated changesets with smaller degrees.

Figure 16: Google Satellite Imagery and Output Area From a Provenance Graph With High Activity Power Law Exponent (output area boundary shaded)

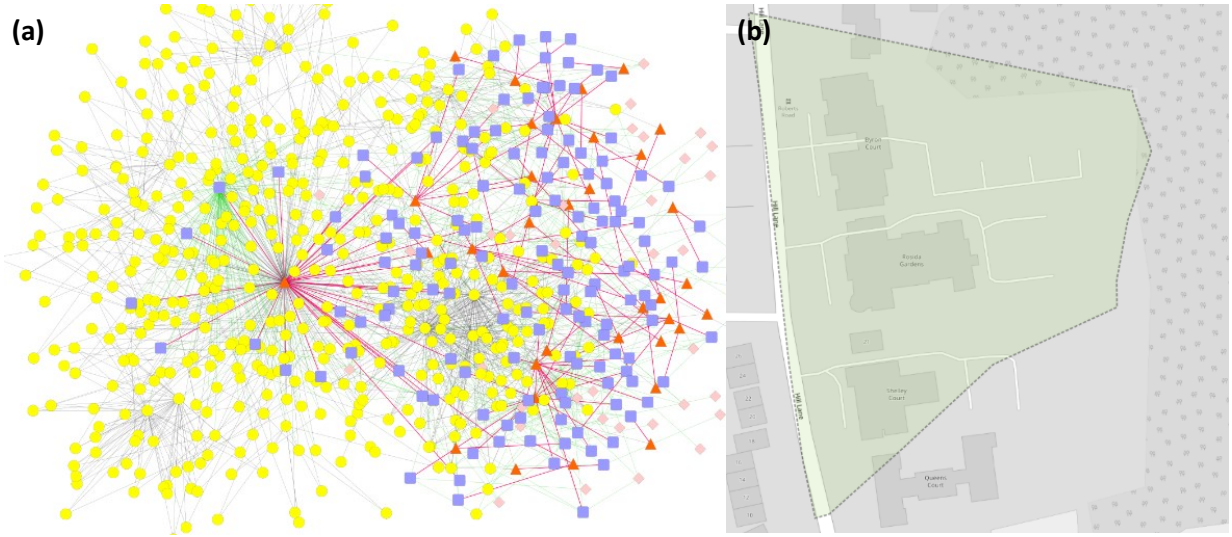


Low Values. Figure 17 shows a provenance graph from an output area with a low activity power law exponent. This output area has named building footprints for three large blocks of flats and simple road and track outlines. Inspection of satellite imagery shows that there are some features missing, such as car parks, wooded areas in the grounds of the flats and some additional pathways.

The most striking difference with the output area in Figure 15 is in the number of activities and agents who have contributed to this OSM coverage. Again, we can see that one contributor (in the middle of the graph) is the most prolific, but in this graph, there are many other contributors who between them have carried out much more editing than those in Figure 15. The editing of all contributors has been carried out in many more changesets (prov:Activities) with the degrees distributed between them. All the activities shown have had their edges shaded in green to show their degree. The prolific contributor had done much of their editing in one changeset which has a degree of 159. The skewing effect of this high degree changeset on the agent degree distribution is

attenuated by the other 145 changesets, which have degrees ranging from 3 to 8. This results in a lower power law exponent.

Figure 17: Provenance Graph With Low Activity Power Law Exponent (a) and OSM Map (b)



Many of these changesets cover large areas of Southampton and often focus on one type of feature such as surface water, cycle routes or footpaths. Also notable is the presence of a large number of software agents used in the activities. Many of these are different versions of the same software and indicate that the edits took place over a long period. The presence of only three buildings in this output area increases the significance of

Figure 18: Google Satellite Imagery and Output Area From a Provenance Graph With Low Activity Power Law Exponent (output area shaded)

these changesets. Edit intensity is generally higher in urban areas and so this effect will be more pronounced in central Southampton with areas here being more likely to be affected by these wide-area changesets.

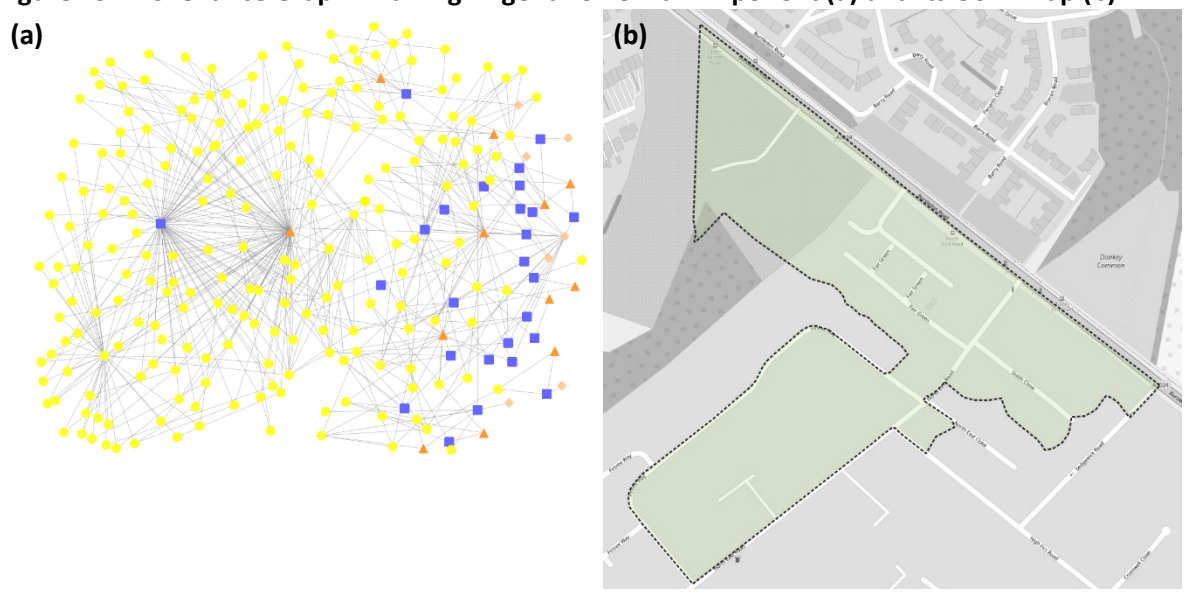
Summary.

- Lower activity power law exponent values can be associated with urban areas with a low building count such as those with blocks of flats.
- High values are a result of intensive mapping by a small number of contributors who focus on the dominant feature type, most commonly building footprints, with a wider variety of scarcer features added by other contributors.

Agent Power Law Exponent.

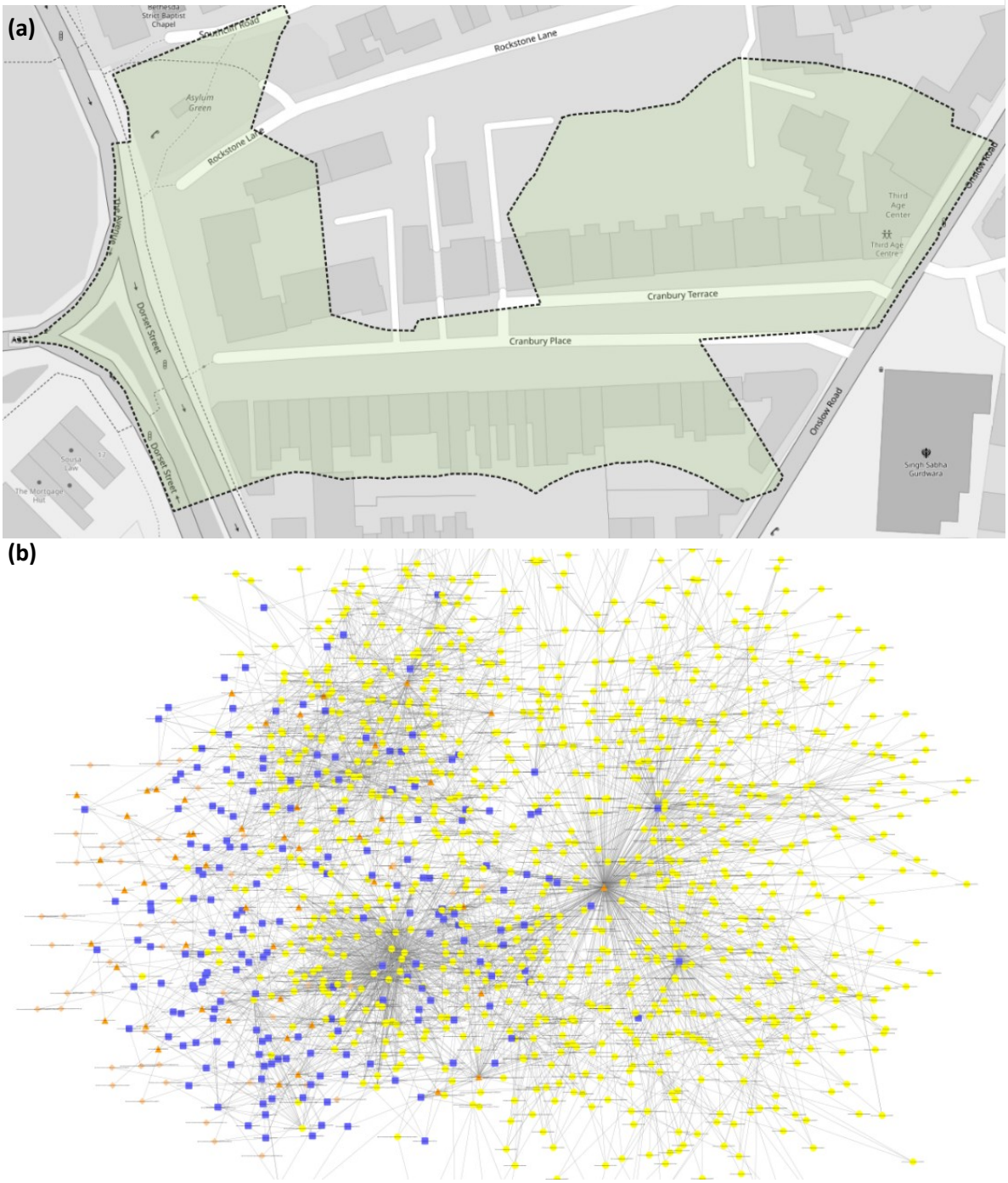
High Values. Higher values of this metric tend to be seen in sparsely edited graphs such as the one in Figure 19, which contains no building footprints and only a few streets in a residential area although these are named. The number of edits is quite small, and most have been done by a single agent. The degree of an agent is mostly determined by the number of edits they have made in the graph. Several other contributors have edited the graph only once or twice and there are also a lot of software agents in this graph. One of these, Potlatch 2, has a high degree and has been used by multiple contributors including the individual responsible for the bulk of the editing. It is an old editor, indicating that much of the editing took place some time ago with only minor changes made more recently. The other software agents are various versions of JOSM and ID. These editors are currently in use, but the range of different versions suggests that these more recent edits have taken place over a longer period.

Figure 19: Provenance Graph With High Agent Power Law Exponent (a) and its OSM Map (b)



Low Values. Low values tend to occur in larger provenance graphs. As seems to be the case in almost all graphs, there is one dominant contributor but, in the graph seen in Figure 20, several other contributors have also made significant edits, which suggests that the mapping of this output area was a much more collaborative effort. There are many other software agents with degrees distributed evenly among them, who have contributed significantly, which suggests that this graph has been built up gradually over time by many contributors.

Figure 20: OSM Map (a) and its Provenance Graph With Low Agent Power Law Exponent (b)



This is an urban residential area in central Southampton and contains lot of accurate building footprints which form the bulk of the map features and some quite detailed mapping including a major road with routing information, named business premises, a public telephone, and several footpaths. Although the built environment mapping seems to be of high quality, there are unmapped green areas and car parking. This suggests a great deal of work has been put into mapping one specific feature type.

Summary.

- This low agent power law value has occurred where one prolific editor maps a dominant feature type (building footprint) in detail with other editors subsequently adding other less numerous but important features. It can also indicate collaborative mapping efforts.
- High values can result if the bulk of the editing has been done by a single contributor, followed by a long period of only minor tweaks by other contributors.

5.2.2 Average Degrees

These metrics are derived by calculating the mean average degree for each of the provenance vertex types.

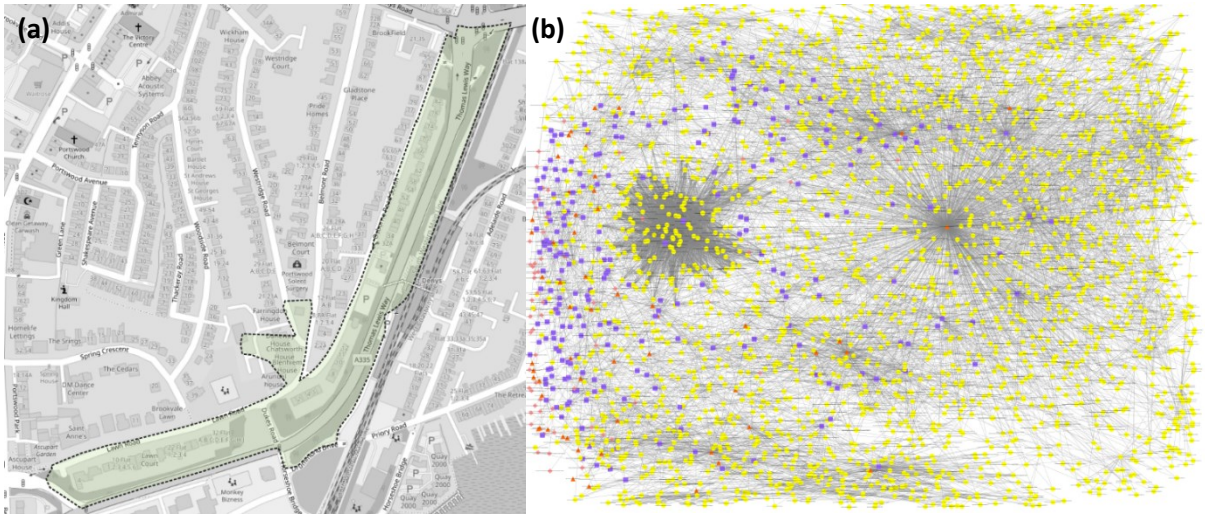
Average Entity Degree.

High Values. The average entity degree is strongly influenced by the number of osm:Ways in the data, and particularly those which describe linear features and have a high osm:node count. Multiple versions of these can lead to particularly high values. The graph and map in Figure 21 are from an output area with a high value. The dark cluster of edges visible on the right-hand side of the graph are osm:Ways and their member osm:nodes. These are both linear features: a main road and a cycle path. Both contain a lot of osm:nodes, many of which are shared by other osm:Ways, which increases their degree. This effect is compounded by multiple edit versions of these osm:Ways, some of which have been edited over 40 times. A node that is a member of a heavily edited osm:Way will have an edge between it and each edit version, and node reuse in other heavily edited osm:Ways can result in a degree of over 100, which greatly influences the mean entity degree value. This output area is ribbon shaped and follows both features, such that it contains a large number of their nodes which dominate the average entity degree.

In smaller graphs, imported data and the use of satellite imagery can also increase the osm:node degree. E.g. Microsoft made its Bing Satellite Imagery dataset available for use by OSM editing software for tracing features, and this appears as a source dataset our provenance model

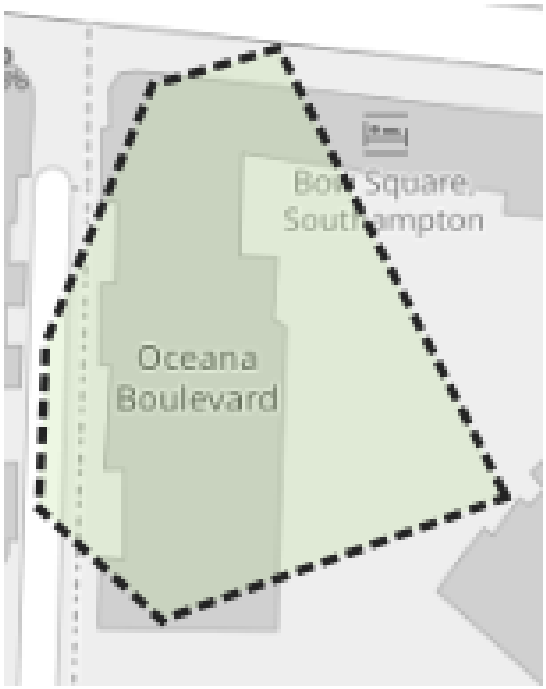
treats as an entity used by an activity. Where this is widely used in an output area, the resulting entity degree can be enough to raise the average.

Figure 21: OSM Map (a) and its Provenance Graph With High Average Entity Degree (b)



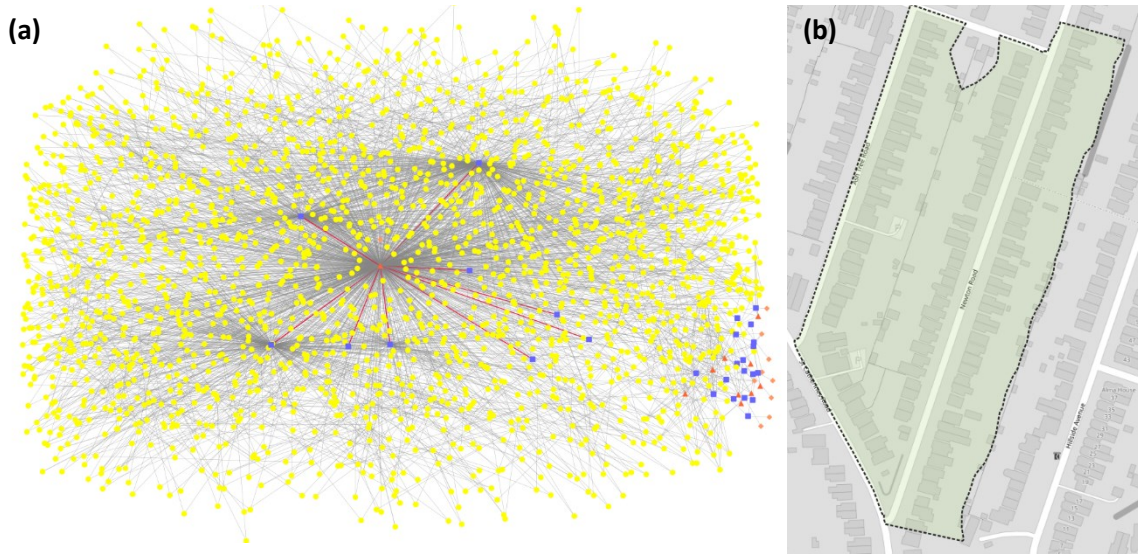
Low Values. Low values are often associated with blocks of flats which result in a high population density and a physically small output area which may only contain one building footprint. The building in the example shown in Figure 22 has had no edits since its creation. This means that the osm:nodes present members of only one version of a single osm:Way.

Figure 22: Output Area With Low Average Entity Degree



In larger graphs with low values such as the one seen in Figure 23, recently created data is a major factor. This is because fewer edit versions means that osm:nodes are members of fewer osm:Way prov:Entities. All the building footprints shown in Figure 23, were created recently, approximately two months before the download date.

Figure 23: Provenance Graph With a Low Average Entity Degree (a) and its OSM map (b)



Consequently, the majority of osm:Ways are at their first edit version. Their member osm:nodes will therefore have lower degrees which results in a lower mean value for entity degrees. The edges coloured in red in Figure 23 connect the contributor responsible for most edits with their changesets and we can see that most of the edits took place in two editing sessions which are likely to have been on consecutive days. Inspection of the changeset XML confirms this.

Summary.

High values can occur:

- In output areas with many linear features which have been heavily edited and where there is a lot of osm:node reuse. Output areas with a linear shape such as the one seen in Figure 21 are more likely to contain such features.
- Where there are a lot osm:Ways present.
- Where there is extensive use of satellite data sets for tracing.

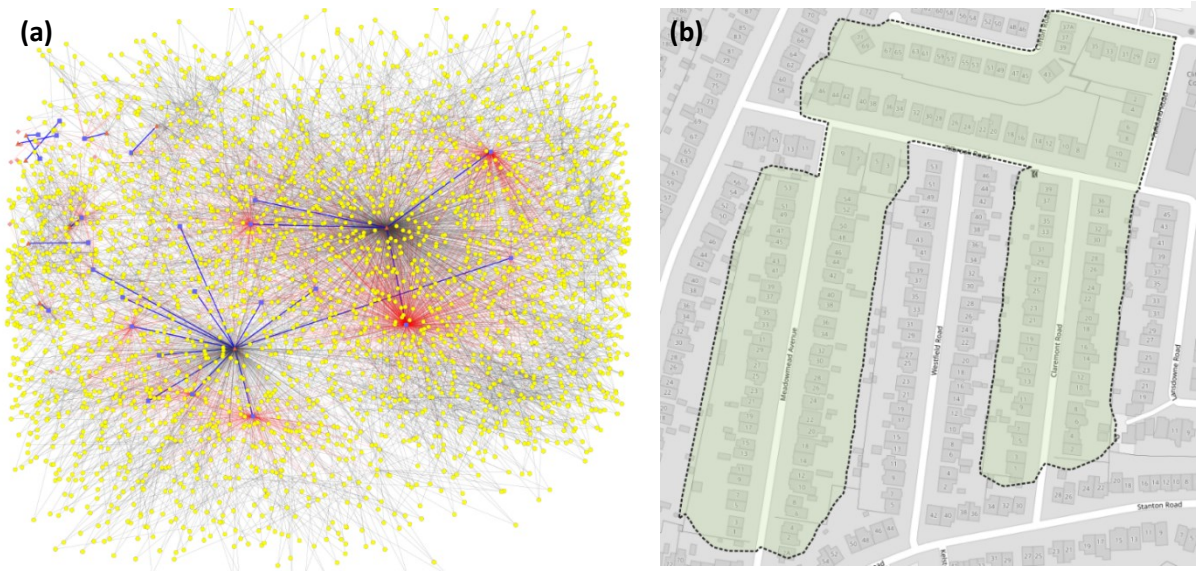
Low values can be associated with:

- Output areas containing tower blocks which house a high population resulting in a small output area dominated by only one building footprint.
- Larger graphs, with recent editing which means most ways are at their first edit version.

Average Activity Degree. Average activity degree and average agent degree have a strong linear relationship (ρ(1178)=.898, p =,<.001).

High Values. The average activity degree increases with the number of data items which are edited in a single edit session. The graph in Figure 24 has a high activity degree which has been caused by two prolific contributors who have dominated the editing. The blue coloured edges show the agent degree and those coloured in red, the activity degree. The two dominant contributors have created an accurate and complete set of building footprints which are the dominant feature of the map coverage. Most of the activity in the graph occurs in four changesets, one of which accounts for the bulk of the edits. This graph shows intensive and comprehensive editing by two experienced, expert editors. One has created a set of building footprints in an intensive editing session which appears to have covered the entire output area and then the other contributor visited later and added postcodes, house numbers, driveways, and other details in another intensive editing session, generating 3 changesets over a short period. The resulting map for this output area is a comprehensive coverage of the built environment generated intensively over a short period by expert contributors. This pattern also produces a high average agent degree because of the large number of features mapped by one or two contributors. The strong negative correlations for average number of creators per feature (ρ(1178) = -0.651, P <.001) and average number of editors per cell (ρ(1178) = -0.696, P <.001) also support this idea.

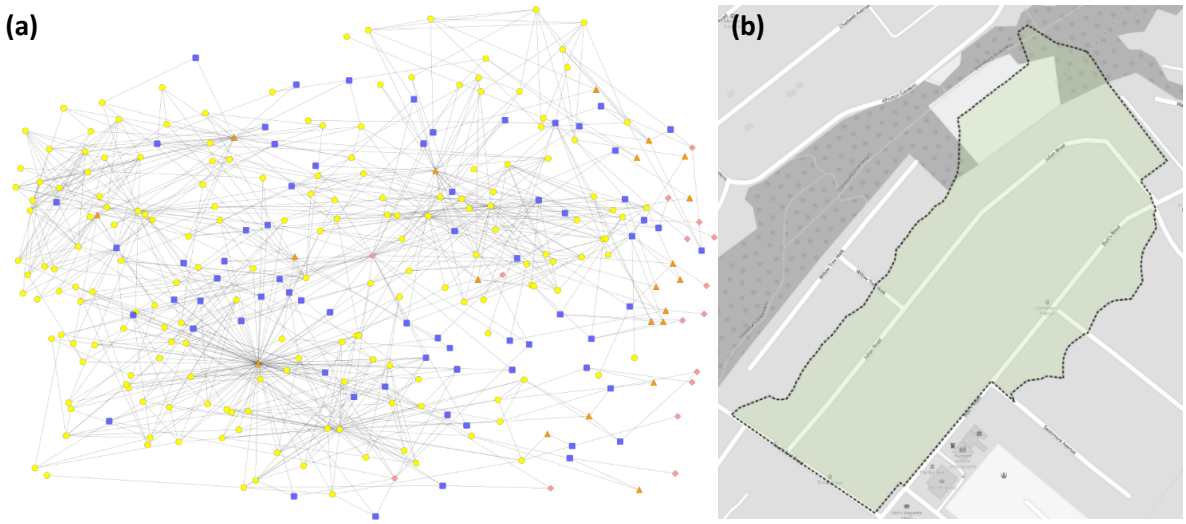
Figure 24: Provenance Graph With High Average Activity Degree (a) and its OSM Map (b)



Low Values. The provenance graph in Figure 25 shows a graph with a low activity degree. The accompanying map shows a sparse OSM coverage with no building footprints and the only

features present being named streets. As seems to be the case in most of the provenance graphs in this section, most of the editing has been carried out by one contributor with most others only carrying out single edits. Inspection of the changesets created by the dominant contributor show that they were not focusing on this output area and contributed substantial edits to the street network in the early days of OpenStreetMap approximately 14 years ago. Thus, much of the editing done in these changesets occurred outside the output area while the OSM coverage was in its exploration phase, covering wide areas in less detail. Since then, this output area has received no attention.

Figure 25: Provenance Graph With Low Average Activity Degree (a) and its OSM Map (b)



Summary.

- Low values of average Activity degree occur in areas that have not been mapped in detail or had close attention from an expert or prolific contributor.
- In urban areas, low values indicate that an area that has not been mapped for a long time.
- High values suggest intensive expert editing

5.2.3 Average Clustering Coefficients

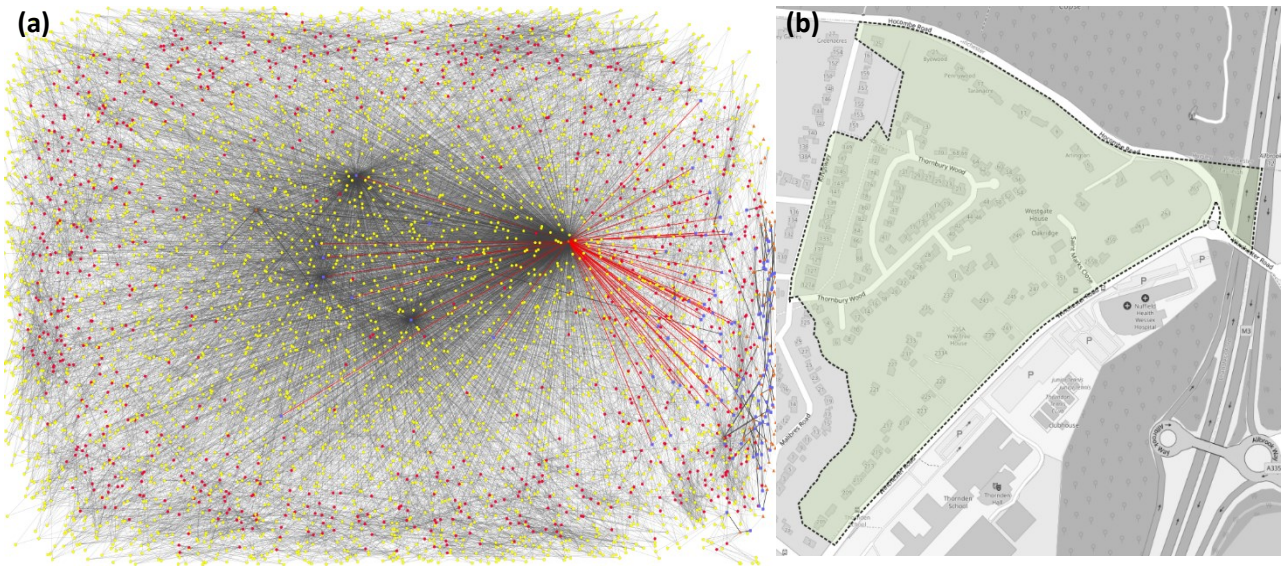
The average clustering coefficient of a vertex is derived from the number of triangles (i.e. complete subgraph of 3 nodes) that that vertex is a part of. It is calculated as the fraction of triangles that the vertex in question could potentially be a part of, which actually exist [184].

We calculate four values: the average clustering coefficient for all vertices, and the average clustering coefficient for each PROV-DM type.

Average clustering coefficient.

High Values. Figure 26 shows the OpenStreetMap coverage and provenance graph for an output area with a high average clustering coefficient. Edges between prov:Agents and prov:Activities have been covered in red, as have the osm:Ways in the graph. We can see the familiar pattern in which the bulk of editing is done by a single contributor. However, in this case the work has been carried out in many changesets.

Figure 26: OSM Map(a) and its Provenance Graph With a High Average Clustering Coefficient (b)

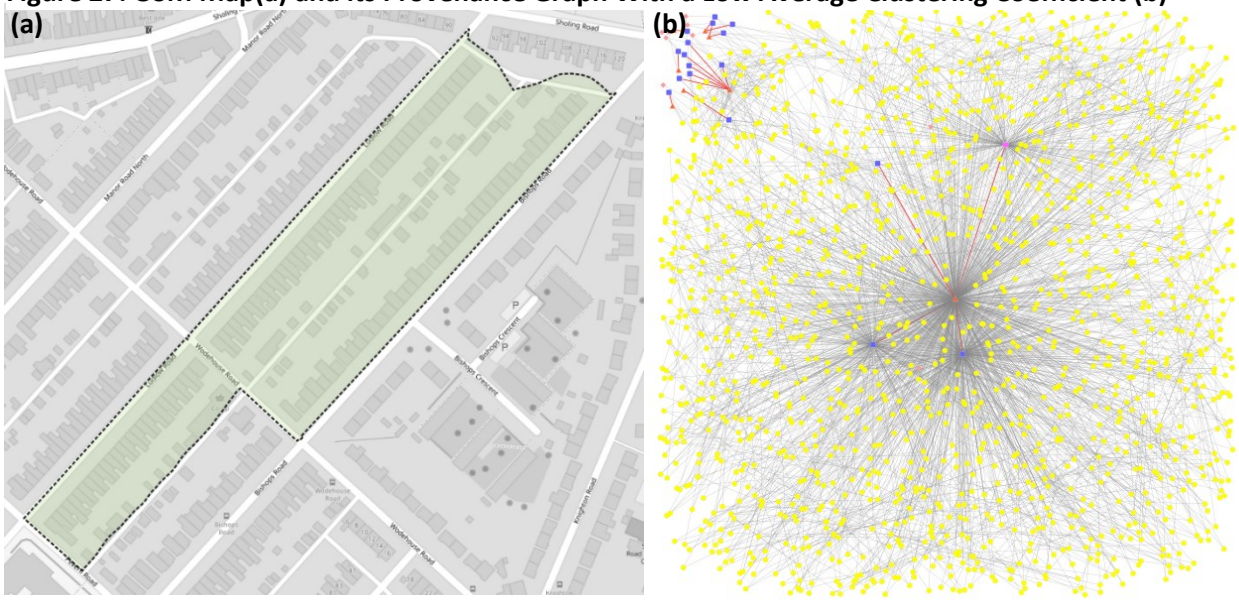


The other notable thing about this graph is the large number of osm:Ways relative to the number of osm:nodes. These osm:Ways are edit versions of building footprints in a suburban residential area. Where a contributor edits these in more than one changeset, triangular structures emerge in the graph, and where these osm:Way versions share osm:Nodes, even more triangles are generated, which gives rise to the high average clustering coefficient. Inspecting the OSM changesets and edit history confirms this. This area was originally mapped by a prolific contributor who has been an active mapper in the surrounding area since 2009. They have returned to the area regularly over this period to enhance their work. Changeset comments such as “improved detail”, “building keys changed to house garage”, and “addresses improved and updated” tell a story of continuous maintenance and enhancement. Examining the bounds of this contributor’s changesets suggests that this is a local contributor who is continuously enhancing the area.

Low Values. The map and provenance graph shown in Figure 27 show a contrasting output area with a low clustering coefficient. Again, the bulk of the editing is done by one dominant editor

but in this graph almost all the edits are creation edits. The three changesets associated with almost all the edits took place over the Christmas period, one just before Christmas and two others on New Year's Eve in the afternoon. The map coverage is detailed and comprehensive and appears to have been produced by an expert contributor, probably enjoying time off work over the holiday period. Most of these features have been created in one go and include postcodes from code point data, car parking, garages, and accurately mapped building footprints. Rather than having been built up gradually as in Figure 26, this area has been comprehensively mapped over a short period with most features having comprehensively mapped in a single changeset. Other editors are unlikely to contribute later as the work is so comprehensive. Although this contributor has created a lot of changesets to the east of central Southampton, they have also been active in other areas of the UK including London, Reading and rural areas to the north of the New Forest National Park. This graph suggests a contributor focusing on an area to create a complete and highly detailed coverage and then moving on, in contrast to Figure 26, where the principal contributor returned several times over a long period gradually building up and maintaining the map coverage, which suggests a local contributor mapping their neighbourhood.

Figure 27: OSM Map(a) and its Provenance Graph With a Low Average Clustering Coefficient (b)



Summary.

- High values can result from regular editing by a single contributor over a long period. It is also likely that this is a feature of areas local to prolific OSM editors.

- Low values can result from intensive editing by a single expert user who moves on once editing is complete

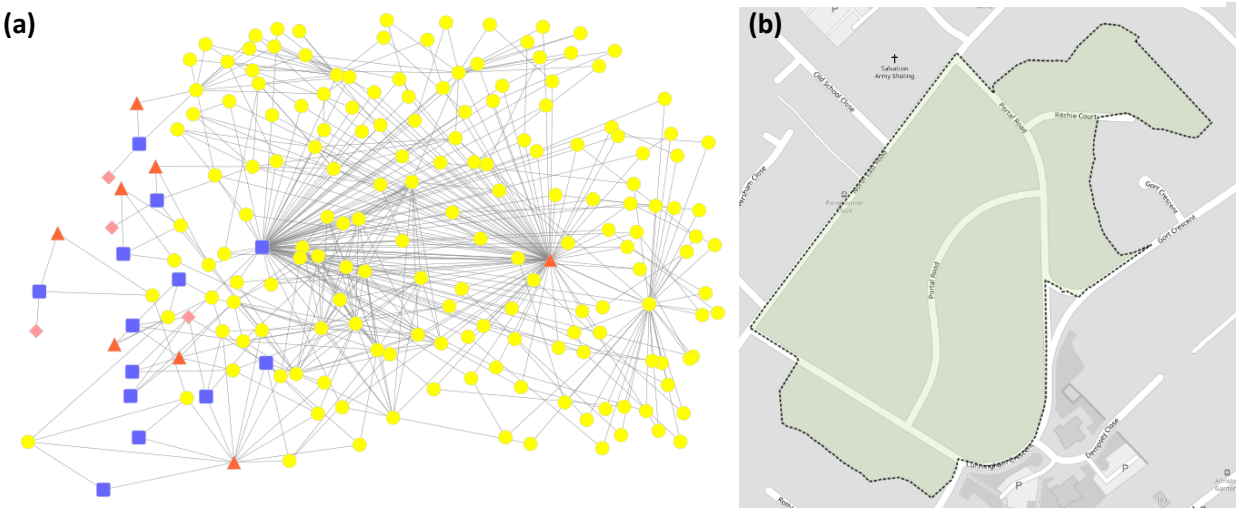
Node specific clustering.

Average clustering coefficient calculations for the three provenance vertex types are also available. The average clustering coefficient for prov:Entities has a very strong linear relationship with the overall average clustering coefficient. This is because prov:Entities are generally the most numerous vertex in in OpenStreetMap provenance graph and so the number of edges connected to prov:Entities and forming triangles with other vertices has the strongest effect on the mean clustering coefficient of all vertices.

Average Clustering Coefficient for Activities.

High Values. Higher values for average activity clustering coefficient can occur in sparsely edited areas. In the provenance graph in Figure 28, most edits were carried out by one contributor in one changeset, which has a degree of 85. Eighty-three of these vertices are connections to prov:Entities and have the potential to form a triangle with the changeset and every other connected entity. Features which have versions edited in another changeset break a potential triangle, lowering the clustering coefficient. Seven other contributors have also edited to a much lesser extent, but most have edited features created in the dominant contributors changeset which as a result, has the lowest clustering coefficient in the graph.

Figure 28: Provenance Graph With High Activity Clustering Coefficient (a) and its OSM Map (b)



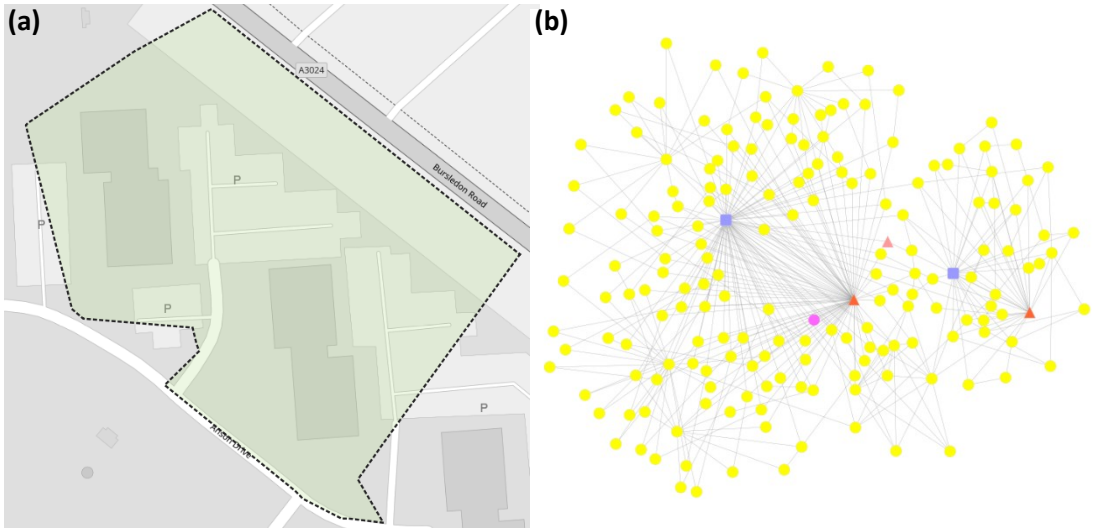
The reason why this graph has such a high average activity clustering coefficient is because of the low degree changesets. Two of these have only one edit in this graph and no link to a software agent, resulting the maximum coefficient value of 1.0 and significantly raising the mean. The others are either associated with software agents, which have a degree of one and thus cannot form

triangles or have two or more associations with prov:Entities which reduce the coefficients to either 0.33 or 0.66. The higher coefficients raise the mean activity clustering coefficient.

These low-degree changesets are another characteristic of a theme we will see quite often in OpenStreetMap output area provenance graphs: that of ***wide area editing***. Wide area editing can be the result of data imports or simply contributors editing a very large area, either because they have a preferred feature which is sparsely distributed or because they are editing in the early days of OpenStreetMap, when the mapping was in an *exploration phase* where mappers were generally aiming for coverage rather than detail. This follows a known elemental pattern of urban development called *exploration and densification* [84] which has also been shown to be common to VGI mapping [85]. These types of edits crop up in all output areas but are magnified in sparsely contributed ones.

Low Values. The map and graph in Figure 29 are from an output area with much lower values. There are some similarities with the previous graph in that much of the work here is replacing content deleted after a license change in 2013. There are only two changesets, both with higher degree sharing several edits. This OSM coverage is more detailed but there are only 3 buildings which are blocks of flats added in the older data replacement changeset and then edited in the newer one. Other graphs with much more detailed content also tend to have a lot of data edited in more than one changeset.

Figure 29: Map and Provenance Graph With a Low Activity Clustering Coefficient



The coefficient value can also be lowered by the presence of software agents (these are absent in early changesets prior to the API change) which cannot form triangles. The average clustering coefficient is also significantly increased by the number of changesets in which only one edit is carried out as this results in a clustering coefficient of one.

Summary.

- high activity clustering coefficients are characteristic of a high proportion of deleted data, and data created prior to the API change in 2008.
- High values can be found in sparsely edited areas where little or no recent editing has occurred. The implication of this is that high values of average Activity clustering coefficient in provenance extracted using output area polygons are likely to be a good indicator of poor data completeness.
- Lower values occur where features are edited in multiple changesets
- Lower values are associated with the presence of software agents in low degree changesets

5.2.4 PROV-DM vertex counts

These are three measurements derived from the number of PROV-DM agents, prov:Entities and activities present in the graph. The measurements used in these studies have all been normalised by the surface area of the output area polygon. In isolation these measurements can vary for a range of reasons, but they are useful because their variations can help to explain other metrics.

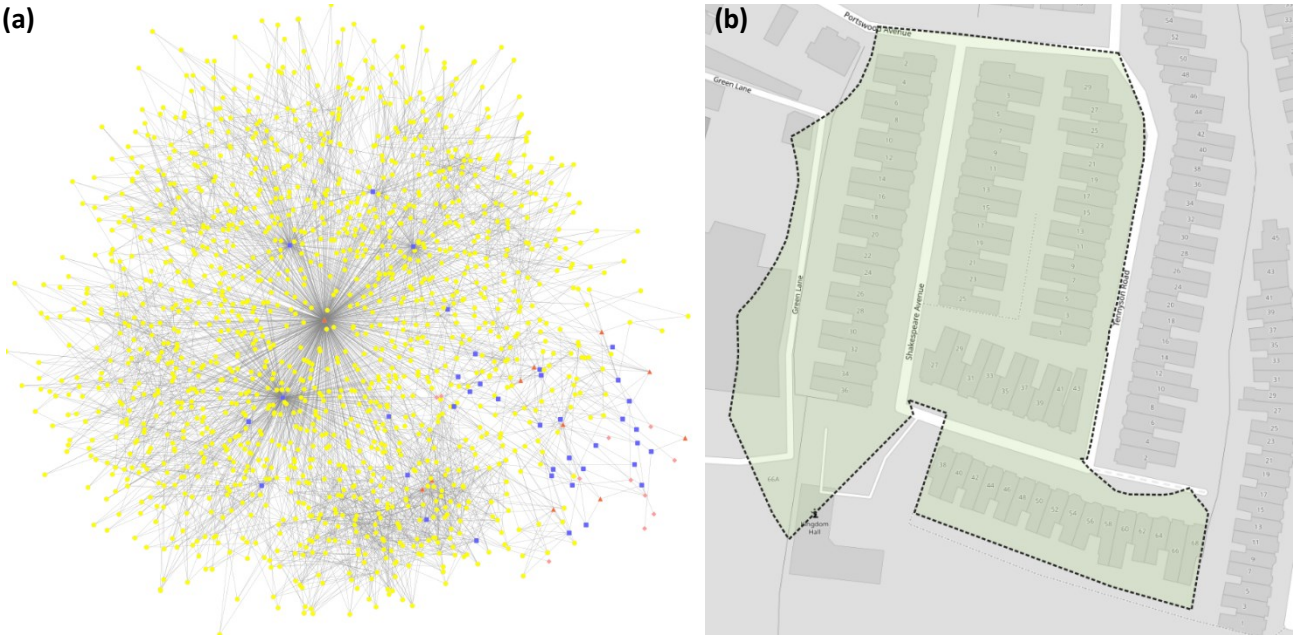
Prov:Entities.

The number of prov:Entities present in a provenance graph can be a function of the type of features present, or the editing intensity as each edit version is counted as an entity in the provenance graph.

High Values. In the graph shown in Figure 30 we can see that much of the editing has been carried out by a single dominant contributor. In this case it was possible to learn more about this contributor, who has posted about their work on the OpenStreetMap blog.

They were a university student and in their blog state that they are particularly interested in mapping building footprints in the residential areas they walked through on the way to university. The building footprints are particularly accurate and refined, with more osm:nodes used for each building than nearby building footprints added by different contributors. These building footprints were edited in several passes with outlines being created in one changeset and then house numbers

Figure 30: Provenance Graph (a) With a High Entity Count and its OSM Map (b)



and postcodes added in subsequent changesets which means there are multiple versions of each building osm:Way. The surrounding street network has also been heavily edited. This area is in part of Southampton with a high student population, quite close to the University. The prov:Entity count seems to be affected to some degree by the type of feature present and quite possibly by the location e.g. a residential area with a lot of potential for building footprints near to a university.

Low values. An output area with a very low entity count, as seen in Figure 31, contains one block of flats in a small output area some distance away. Given that some dwelling types occur more frequently in certain output area classifications, it seems reasonable to suggest that entity count may show some geodemographic variation. Low entity counts can also result from poor completeness as in **Error! Reference source not found.** which has only unnamed side-streets (the p layground polygon visible was added after the provenance was recorded.) Although some studies have shown that poor data completeness occurs in more deprived areas and this one has an OAC classification of "hard-pressed living" there are other drivers for entity count such as the size of the output area, land-use cover and built environment features. This means that such geodemographic links are likely to be hard to find.

Summary.

- High values in areas with a lot of buildings in areas where completeness is high, e.g. residential areas near a university
- Low values can indicate low completeness

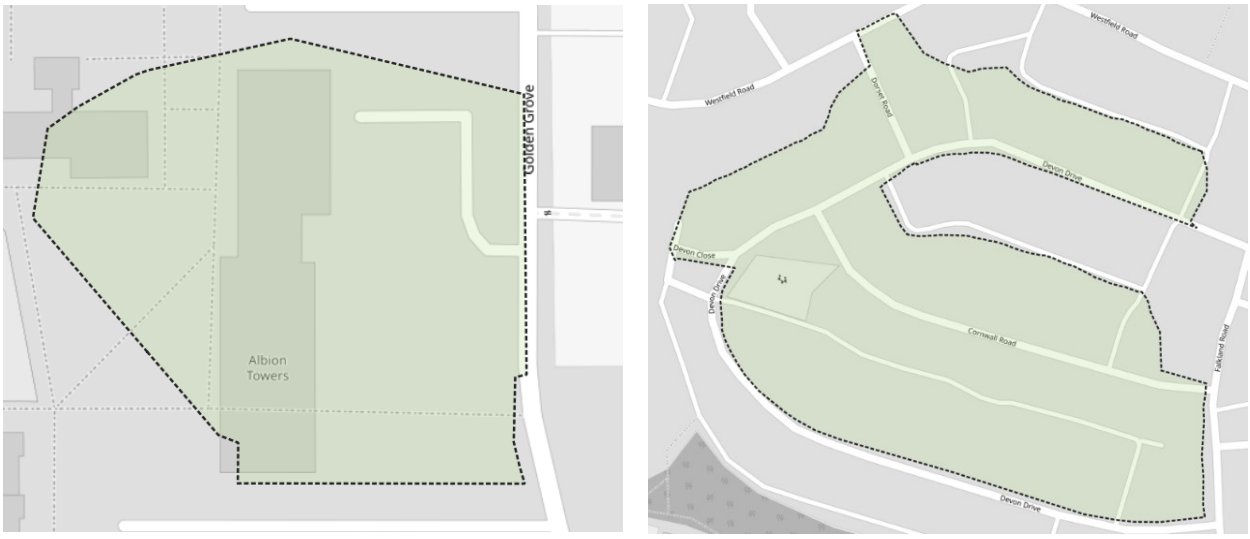
(a) **(b)**

- Low values can result from densely populated output areas with tower blocks which result in a very small polygon with few buildings

Activities and Agents.

These variables positively correlate because each agent will create one or more changesets and it therefore follows that the number of changesets will increase with the number of agents. The prov:Agents counted by the *agents* variable include both prov:SoftwareAgents, i.e. OpenStreetMap editing software, and human contributors, both of which have responsibility in a provenance graph. The levels and variations of *agents* would of course vary had we made the modelling decision to treat these as separate PROV-DM vertex types.

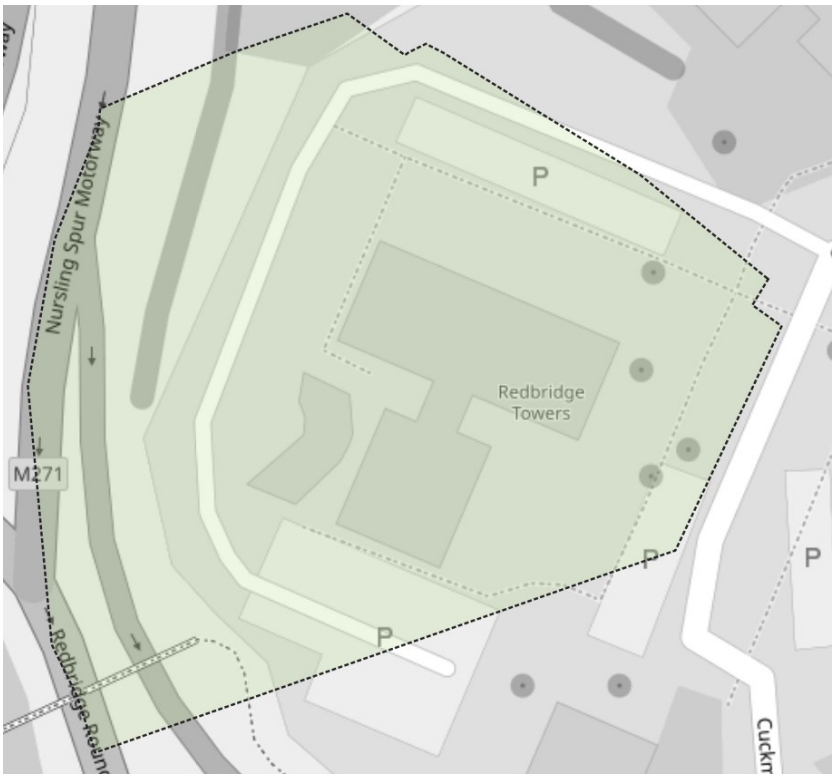
Figure 31: OSM Maps for Output Areas (shaded) Whose Provenance Graphs Have a Low Entity Count



High Values. The map shown in Figure 32 has a provenance graph with a very high prov:Agent count. In this case the output area is small because the only building is a tower block with a high population. This also means the feature count is quite small compared with a mixed residential area. The high agent count is due to the presence of the M271 motorway which is a large intensely edited public feature. Major highways, although rendered quite simply, contain a lot of other invisible information such as direction, lanes, speed limits, public transport routes, etc. This information tends to change regularly over time and frequently updated, often by a different agent on each occasion. For example, a cursory inspection of the changesets for this area reveals speed limit and bus route changes. Other map content with a high agent count tends to involve blocks of flats and public amenities or highways. Central urban areas also often have quite a lot of wide area

edits. Because these output areas are small and this metric is normalised by polygon area, high agent count values result.

Figure 32: OSM Map for Output Area (shaded) Whose Provenance Graph has a High Agent Count



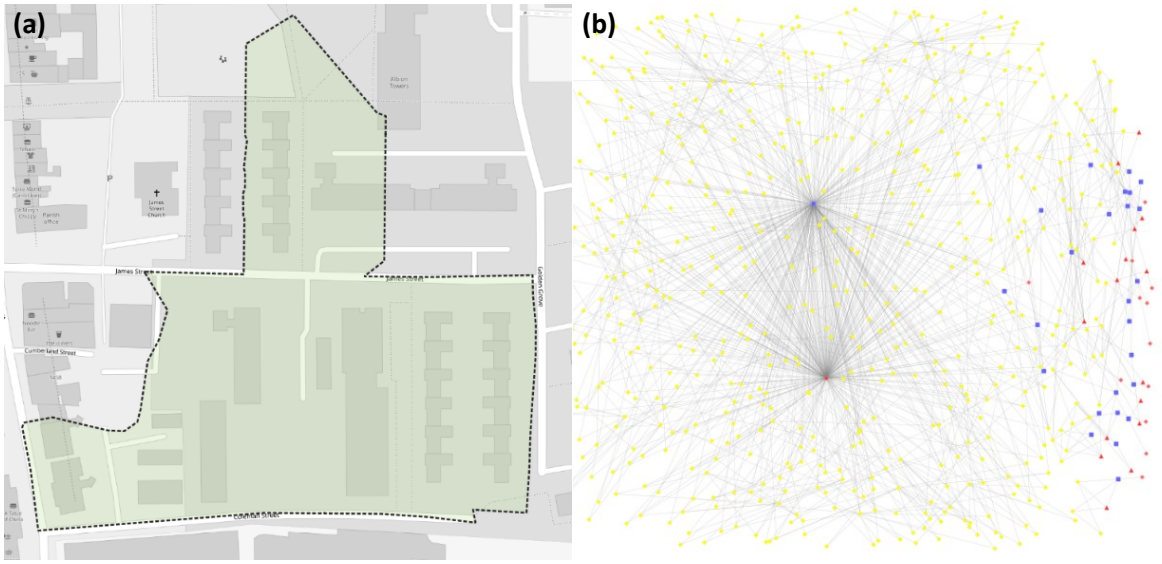
Average Rich Club Coefficient.

The rich club coefficient (RCC) is a measure of the extent to which vertices are connected to other well-connected vertices. It is calculated by evaluating each vertex degree in the graph. For each degree *K* the rich club algorithm (NetworkX) will measure the ratio of actual to possible vertices with degrees greater than *K*. We record the mean rich club coefficient for each output area graph.

High Values. High rich club coefficients typically occur in OSM provenance graphs when a prov:Agent carries out lot of edits in one changeset and when prov:Agents edit features created/edited by other prov:Agents who have also done a lot of editing in the graph. The example in Figure 33 is typical of graphs with a high value RCC. Most of the content is a set of painstakingly drawn building footprint polygons. These appear to be of high quality and drawn by an expert contributor. Inspection of the changeset reveals this is the same individual mentioned earlier (Section 5.2.1), who specialises in building footprints.

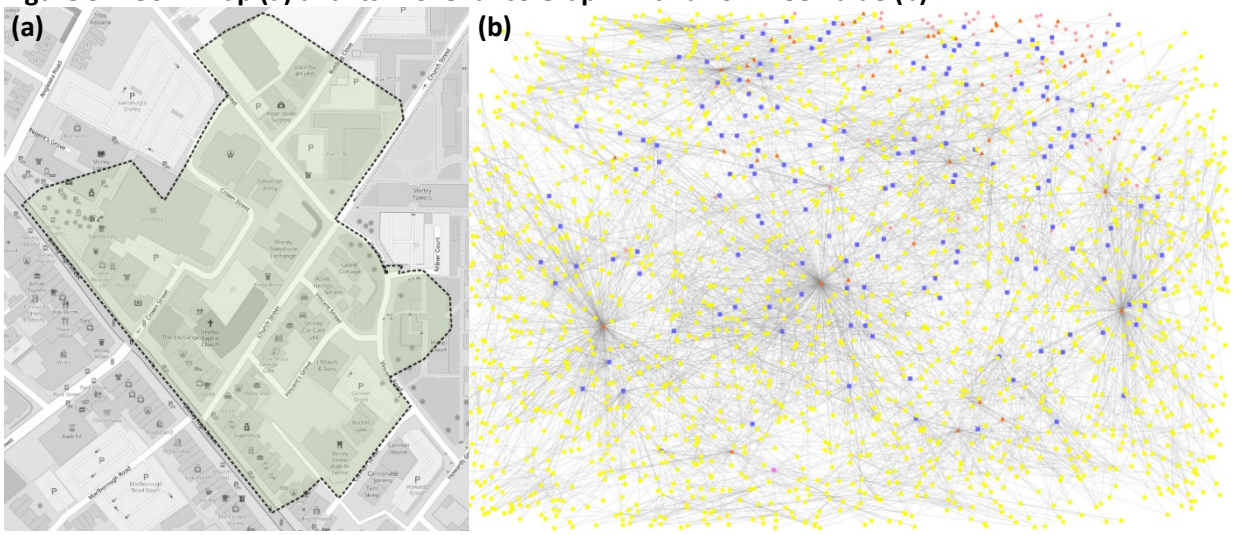
The edits took place in a single intensive editing session resulting in a high degree prov:Agent connected to a high degree prov:Activity. This dominates the graph and the other changesets and agents are all associated with small, wide area edits.

Figure 33: OSM Map (a) and its Provenance Graph With a High Rich Club Coefficient (b)



Low Values. Lower values, on the other hand, have many more changesets and the work is distributed among more agents, resulting in lower degrees. The graphs have much more edit interactions, i.e. agents editing the work of other agents. The OSM coverage with low RCC's will tend to have more variety of features which are more likely to be richly tagged. The graph shown in Figure 34 is typical. This area has numerous building footprints on a busy shopping high street in the Shirley area of Southampton. All buildings are tagged with business names, amenity types etc, and these along with urban highways, footpaths and public transport infrastructure have been regularly tweaked over a long period by several contributors. This pattern is likely to occur in urban centres and commercial districts, where there is a prevalence of features which change regularly over long periods. This is consistent with the positive correlation with maintenance ratio ($\rho(1178) = 0.484$, $P < .001$), a metric which decreases as the proportion of maintenance edits increases, and a negative relationship with interactivity, the average version number $\rho(1178) = -0.493$, $P < .001$. For both metrics this relationship is strongest in supergroup 2 and is reversed in rural areas (supergroup 3).

Figure 34: OSM Map (a) and its Provenance Graph With a Low RCC Value (b)



Summary.

- High values are associated with intensive editing
- Low values with continuous maintenance editing over longer periods, often in areas with a high rate of change such as commercial high streets

5.3 Discussion

In this section we have visually examined the network properties of OpenStreetMap provenance graphs. We have examined causes of variation in our network metrics by comparing and contrasting graphs with high and low values for each metric. As well as visual examination of the graphs we have been able to learn a great deal from inspection of the changesets represented by each prov:Activity. Using the OpenStreetMap website and API we are able to examine the bounds of the changeset, i.e. the area the edits within it took place in, as well as other changesets by the contributor. We are also able to view the contributor's personal page which often contains notes, diary entries etc. Examination of the OpenStreetMap coverage also provides an overview of the type of area from which the provenance was captured.

The first thing which has become apparent from this exercise is that the factors which drive these variations fall into three main categories: feature dynamics, editing dynamics and spatial effects. In most, if not all the graphs we have looked at, more than one of these factors is at work. This complicates the isolation of aspects of contributor behaviour or demographics as driving factors, particularly as geodemographic classification of an area is related to land-use cover and the type of built environment.

5.3.1 Feature Dynamics

Feature dynamics relates to variations in physical factors such as the type of built environment and land use cover. These factors influence the presence of different feature types which in turn affect the editing process and the structure of provenance graphs. Our provenance capture policy generates provenance graphs with edges connecting osm:Ways and their member osm:nodes, as well as between all of the edit versions of bot primitive types. This means, for example, a linear feature represented by an osm:Way with 30 osm:nodes some of which are shared with other heavily edited features, will have a different provenance network signature from an osm:Way which is a building footprint in a residential area, which is a closed polygon with four osm:Nodes. These differences will also become magnified by editing.

In our dataset, one example of a feature dynamic is the effect of numerous linear features such as street/road networks which can have high osm:node counts and are often heavily edited and involve a lot of osm:node reuse. This can result in high average entity degrees. The network metrics we have examined are affected by feature dynamics in varying degrees. For example, values of *activity power law coefficient* tend to be lower in smaller output areas containing tower blocks.

The use of provenance network graphs for understanding and visualising spatial and temporal distributions of the type of features and land use coverage mapped in VGI is a potentially valuable “remotes sensing” application and the provenance capture policy could be optimised to enhance these signals. Moreau et al , in their introduction to PROV, describe views of provenance [163] which could inform such policies and in this case we would be interested in some variant of the data flow view.

5.3.2 Spatial Effects

The MAUP. An important spatial effect is the *modifiable aerial unit problem* (MAUP) [221] in that adjustments to the boundary and polygon size have the potential to produce very different aggregate measurement values. Larger polygons will tend to have smaller variance owing to the smoothing effect of the larger scale and number of data points [226]. Our use of nonarbitrary output area boundaries complicates consideration of the MAUP because the polygon size and shape are themselves driven by geodemographic factors which through the MAUP scale effect, indirectly drive the network metric values.

The geometry we used to extract the provenance data is based on census output area polygons. These have been designed to standardise their population, so their size is a function of it and can vary considerably [287]. The small size of some output areas can affect the nature of the

provenance graph, often because smaller polygons have fewer features, especially in urban areas. This effect also has a demographic dimension in that it relates to population density and housing type, e.g. tower blocks. These tend to be in central urban areas where the small output area polygons also often clip heavily edited urban features such as motorways, introducing an additional feature dynamic that can dominate the graph.

Wide Area Editing. Wide area editing is both a spatial effect and an editing dynamic. It is present over most of the study area and refers to edits carried out in changesets over large areas, often on a regional or national scale. These can be automated processes such as imports or bot activity, or they can be contributors who edit sparsely distributed features. For example, water culverts, which occur infrequently but over wide areas. Because the provenance capture policy only considers the contents of an output area, other edits in the changeset are not recorded, and the result is changesets which appear very small, often with only one edit. This is less noticeable in larger graphs, i.e. in areas with higher completeness, but can result in high power law exponent values, average activity and agent degrees and Activity clustering coefficients in sparsely edited areas.

5.3.3 Editing Dynamics

Editing dynamics are factors related to the editing practices of OpenStreetMap data. On the surface, given the nature of this project, understanding feature dynamics is our most important, and main objective. What this chapter shows however, is that editing dynamics do not happen in isolation. When investigating spatially defined provenance graphs we need to understand spatial and feature dynamics. We need to control for them and/or account for them if we are to use provenance network analytics to gain insights into how OpenStreetMap data is created. These editing practices are frequently linked to the type of features being mapped and much of the variation in network metrics is a result of the interplay between feature dynamics, spatial effects and editing dynamics. For instance, entity power law exponents tend to be higher in sparsely mapped areas containing only simple street outlines, which is a feature dynamic. However, they can be magnified by the presence of heavily edited features which are clipped by the output area boundary which is both an editing dynamic and a spatial effect

Despite the various complications and confounding factors, visual inspection of the provenance graphs in this chapter has revealed several themes related to editing behaviour.

Collaborative Editing. occurs when several mappers edit the same features. This is a phenomenon recognised in the literature and we have seen it occurring in the graphs in this chapter, particularly in residential areas. In our data we have seen prolific editors preferentially editing building footprints, and then later, another prolific editor comes to the area and adds postcodes or splits

building footprints into separate buildings. This can result in particularly high average activity and entity degrees, and low agent power law exponent, rich club, and activity clustering coefficient.

Focused Editing (Expert Editing). Focused editing occurs when a single contributor generates a large amount of content in one changeset. Some focused editing can be in more than one changeset, but this will usually be on consecutive days. This type of editing pattern is common among experienced or expert contributors. It will also result in high quality complete data. It can be indicated by a low average clustering coefficient, high average agent and activity degrees, and high average rich club coefficients. Where the focused editing is recent or extremely thorough, low average entity degrees can occur.

Wide Area Editing. Wide area editing, mentioned earlier as a spatial effect, is also an editing dynamic. This this is because some contributors preferentially map features which are dispersed over a wide area. This may be for professional reasons or purely personal preference.

Local Editing. Local editing occurs where the dominant editor/s live or work in the area being edited. In this situation a single agent will return to the area and make significant edits at regular intervals. Inspection of the dominant agent's changesets reveals a concentration of editing in the local area. This can result in high average clustering coefficients as the same contributor repeatedly edits data. This is in contrast to focused editing where the contributor often completes editing in an area to a high standard before moving on and not returning. Entity power law exponents will also tend to be quite low owing to the highly detailed coverage that is often produced.

Maintenance Editing. Maintenance Editing often occurs with local editing, although it involves more mappers and can result in low rich club coefficients. Edits in multiple changesets also raise the average clustering coefficient. We have seen it in business districts with many tagged buildings which are business premises and subject to periodic change. It is also associated with transport infrastructure.

Sparse Editing. Sparse Editing is a pattern related to low completeness or, less often, where data was redacted following a license change in 2013. It can result in high activity clustering coefficients. Entity power law exponents tend to be higher because of the spiky profile of degree distributions with lower data volume. It is more likely to appear in areas away from city centres and Universities

Specialisation. Specialisation occurs where there is intensive activity by a single contributor concentrating on mapping feature which is usually predominant in the area. This is often building footprints. Where features are more widely and sparsely distributed, this type of editing is not well captured by our provenance modelling because it only shows up as single isolated edits which generally affect areas with low completeness. Where the specialisation is in a locally dominant

feature such as a building footprint, editing is often highly focused, resulting in a high activity power law exponent.

Creation Editing. Creation editing occurs where the current version of a feature has been produced by the same mapper who created it. Is often carried out by highly focused, expert contributors who edit data in an area exhaustively and to a high standard, such that there is little scope for further editing. This often results in average high rich club coefficient, average clustering coefficients and activity degrees and low average entity degrees.

5.4 Conclusions

In this chapter we have addressed *research question 2* by identifying insights which can be gained from considering graph theoretic network properties of provenance graphs. We have addressed the enigmatic nature of these mathematical constructs by inspecting provenance graphs in detail, comparing and contrasting those with high and low values for several of the metrics to understand the factors which drive their variation.

We have uncovered several themes which seem to be responsible: spatial effects, editing dynamics and feature dynamics all of which leave a signature in the properties of provenance graphs. We have also seen how more specific insights can be derived such as the identification of local editing and expert intensive editing. As well as uncovering specific editing practices we can also gain potentially insights about the characteristics of an area such as its rate of change, which leaves a provenance signature which can indicate the presence of commercial districts and business activity in urban areas.

We have also seen how these different dynamics interact with each other, which potentially complicates and confounds the task of isolating individual factors. The provenance model used to reconstruct provenance graphs from the edit history was not conceived with any of this in mind and was merely designed to capture as much provenance information as possible. We have seen that the dominance of certain features, particularly building footprints, but also linear features, leaves signatures in provenance graphs as does the practice of capturing the provenance of member osm:nodes for osm:Ways. This opens up potential for controlling these and other factors in order to focus investigations on specific aspects of OSM data creation. It provides a basis to address research question one by underpinning new provenance capture strategies. Our interpretation of these graphs has also been aided by consideration of some of the concrete maturity metrics and it is clear that these measurement strategies complement one another, providing additional interpretation strategies.

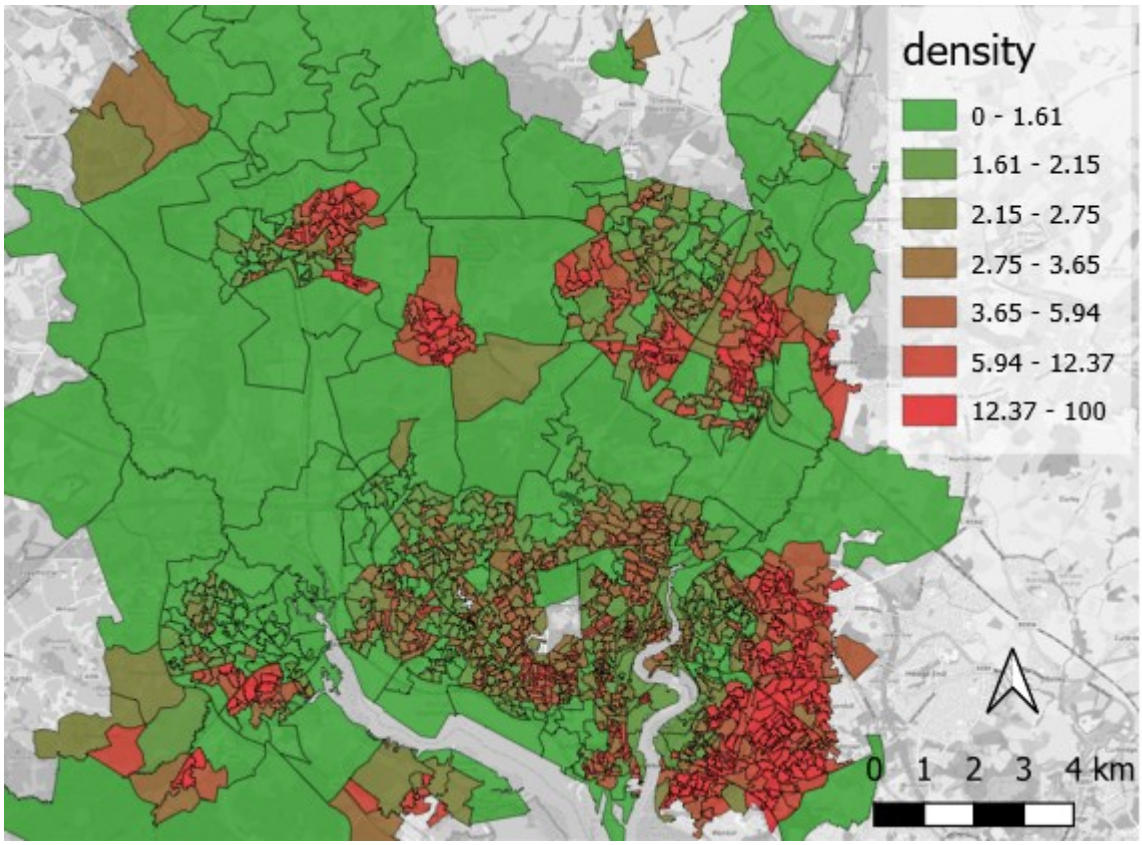
This investigation is by no means exhaustive and many of the effects we have identified here may not generalise to other parts of the world. This not because the general principles do not hold, but because there are likely to be other themes, we do not have instances of here. We also have not investigated all the metrics and further work is likely to uncover more effects. Although some of the findings are reinforced by dataset wide correlations with maturity metrics, it is likely that many of the contributor dynamics will vary with different feature types and land uses. Building footprints have played a pivotal role in many of the graphs we have examined as has data completeness. Better understanding of the provenance for these contexts and feature types will be needed to provide more systematic and generalisable findings. We have however, uncovered useful insights both into OSM editing practices and contexts as well as potential strategies for refining techniques based on alternative provenance capture strategies.

Chapter 6 VGI Provenance as a Geospatial Variable

In this chapter we continue our analysis by investigating what relationship OpenStreetMap has with the physical area being mapped. We examine spatial properties using thematic maps, and physical properties using measurements of the natural and built environment. Visual inspection of the thematic maps reveals interesting and distinctive spatial patterns. The example values shown on the maps in Figure 35 and Figure 36 are clearly clustered in a manner which reflects some deterministic, spatially variable phenomena.

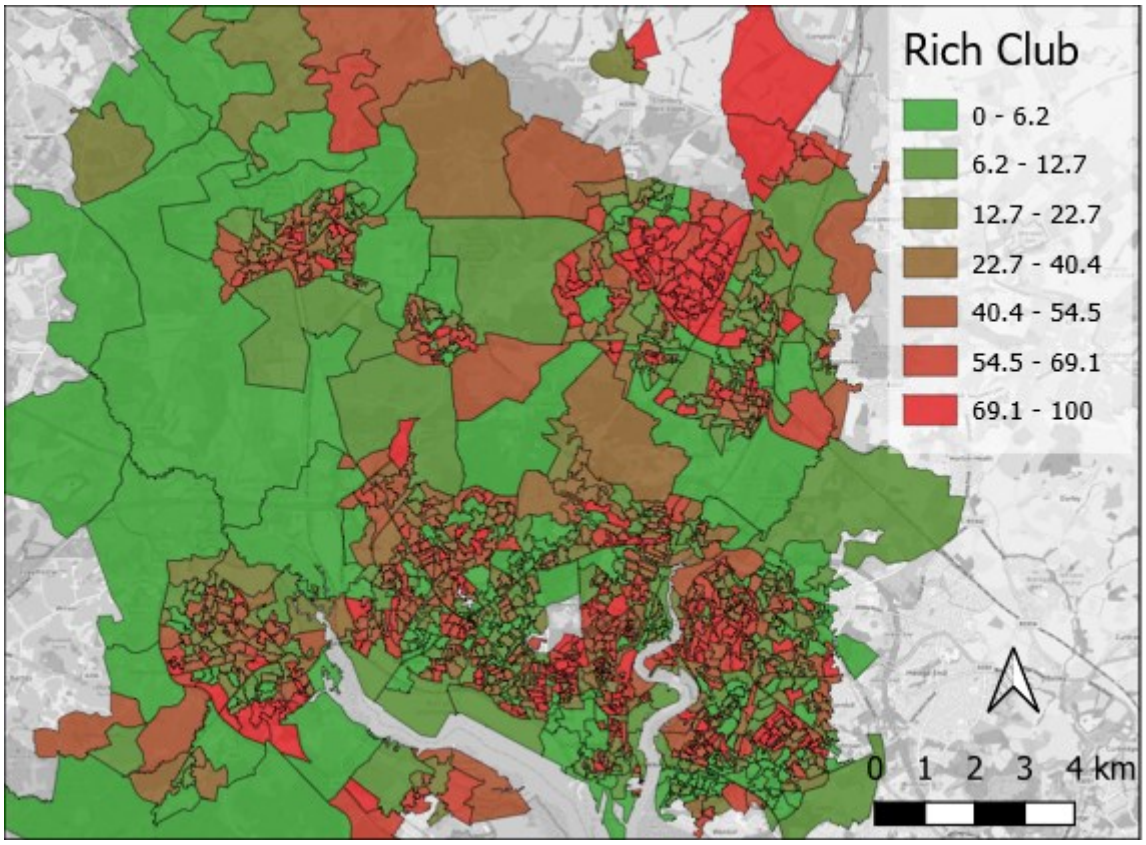
Differences between the OAC supergroups (shown in Figure 37) however, are not readily visually discernible and the spatial clustering we can see does not appear to correspond with the OAC supergroups. To understand these spatial patterns and what drives them, we need to identify any relationships which may exist between the provenance variables and the environment being mapped. To do this we extract data from the Ordnance Survey's MasterMap Topography Layer, from which we derive information about the natural and built environment. We consider this along with information from the 2011 census output area classification pen portraits, which provide demographic information, including descriptions of the typical built environment for the classification groupings.

Figure 35: Graph Density by Output Area



A typical example of these patterns can be seen Figure 35, which shows the *Density* (ratio of vertices to edges) of provenance graphs by output area. This map shows higher values (Reds) in more central urban areas and lower values (greens) in suburban/rural areas. There is a cluster of particularly high values on the south-east side of the map, and we know this was a sparsely mapped area when the provenance graphs were captured. There are also clusters of low values in urban areas on the west of the map. Figure 36 shows a similar map of *Rich Club Coefficient* values, showing higher values concentrated in urban areas and a distinctively zone with higher values in the northeast of the map. The OAs with high values tend to be contiguous, which may reflect specific mapping behaviours.

Figure 36: Average Rich Club Coefficient by Output Area



Most of the provenance variables exhibit one or more of these patterns in varying degrees. Possible drivers can be related into the following themes:

- Map maturity/completeness.
- The human/built environment.
- Geodemographic/socio-economic variation.
- Contributor behaviour/dynamics.

The demographic metrics used to classify output areas to create the 2011 census output area classifications [231] also fall within some of these categories. If these attributes are drivers for the spatial patterns we see in thematic maps of provenance variables, then we would expect to see the small and significant differences we find between output area groupings in Chapter 7, Section 7.7. It also is likely that the characteristics of individual contributors have a role to play. We know from other research [32], [68], [88] that the demographic characteristics of an area can affect contributor mapping behaviours. Users often have preferred 'pet' features [90] and land use types [86], whose incidence varies geodemographically. They also tend to be attracted to well-known and more familiar locations which has also been shown to affect contribution patterns [29], [88], [89]. The topographic structure of the features being mapped will also have an effect, i.e. the procedures for constructing mapped features for terraced housing differ from those involved in the mapping of tower blocks or woodland areas.

Several studies have examined the use of human and built environment attributes to predict levels of different types of human activity, so it is not unreasonable to suppose that they may also have an impact on the way VGI is generated. Examples include travel demand, which has been linked to land use and geodemographic characteristics [288]; mental health, which has associations with green space availability; body mass index which is related to a range of geodemographic and environment variables [289] and levels of walking and cycling activity among adults, linked to a range of built environment variables such as address density, proximity of green space and road networks [290].

Parameters of the human and built environment have also been used as data quality predictors in OSM. E.g. Dorn [98] found differences in building completeness between urban and rural areas, and in specific land use types, with forested areas having higher completeness and correctness. More generally they found accuracy and completeness with proximity to densely populated areas. Zhou [291] also found relationships between the building density in an area and data completeness in OSM. Interestingly their findings were restricted to urban areas. Arsanjani [29] also found relationships between built environment data and levels of contribution, although they related these to the demographic characteristics of the contributors, who tended to be urban residents.

In the following sections we describe the measurement of physical environment factors using the Ordnance Survey's MasterMap Topography Layer. We use these to explore relationships between OpenStreetMap provenance graphs and the physical and built environment by calculating correlation coefficients between the MasterMap and provenance variables. As well as examining the

correlation coefficients we look at how these relationships vary between demographic groupings and what the source of those variations might be. We also take a detailed look at some of the spatial patterns in provenance measurements. We interpret them by considering relationships with the physical environment alongside the OpenStreetMap coverage to uncover what they reveal about OpenStreetMap contribution patterns.

6.1 Variables of the Human and Natural Environment: The Ordnance Survey MasterMap Topography Layer

The Ordnance Survey's MasterMap is their flagship geographic dataset which aims to record every feature in the United Kingdom using state-of-the-art survey and remote sensing methods. It is regarded as an authoritative "ground truth" data set, which has long been a standard for studies assessing UK OSM data quality by comparison with reference data . The topography layer contains a vector representation of over 450 million features, each referenced by a unique TOID (**TO**pographic **ID**entifier) [292]. In addition to the geometry of each feature, the layer contains metadata which divides features into themes and provides them with a set of attributes. Other data can also be added using the TOID.

For our purposes, some additional geometric attributes are useful, namely calculated surface areas for each feature, which along with separately available building height attribute data, allow us to calculate building volumes and surface areas for roads and other man-made and natural surfaces. Using QGIS, we can merge the topography layer with the output area geometry and calculate values for variables which measure built and natural environment attributes for each output area.

For each output area in our dataset, we calculate:

- *Building volume*: the total volume of buildings within the output area.
- *Average building height*: the mean building height of buildings within the output area.
- *Maximum height:* the height of the tallest building within the output area.
- *Road area*: the surface area of all roads which cross the output area. The area calculated for this variable will include portions of road not within the output area and in some cases some considerable distance from the boundaries of output area. This value remains interesting because it is potentially a measure of how well served an output area is by the road network.

- *Road count*: the number of road features which cross the output area.
- *Man-made surface area*: for this, two values can be computed; surface area including and not including buildings.
- *Natural surfaces area*: the area of surfaces which are not man-made examples of which include agricultural land, coniferous woodland, standing water.
- *Natural to man-made surfaces ratio*: the ratio of the surface area of natural to man-made surfaces.
- *Address to building ratio*: calculated using Ordnance Survey UPRN data, this variable measures the number of addressable locations associated with each building. It provides an indication of how UPRN how many blocks of flats there are in an area.

6.2 Correlations

We have assessed univariate correlations between the provenance derived variables and human environment variables derived from the Ordnance Survey's MasterMap topography layer. We preferred Spearman's Rho because of the skewed distributions of many of our variables and our cautious approach to outlier removal. To provide added robustness we also used a 1000 sample BCa bootstrapping procedure to calculate 95% confidence intervals [293].

The statistics reported in Table 7 are significant following bootstrapping, i.e. the upper and lower confidence bounds did not intersect zero in each case. They are also 99% significant ($P < .001$). Table 7 shows Spearman's rho values calculated for the entire dataset which show at least moderate relationships between human environment variables and the provenance variables. Absolute values below 0.300 have been omitted. Although modest, these correlations are statistically significant and unlikely to be the result of random chance.

The strongest relationship shown in Table 7 is between the number of agents in a provenance graph and the area of man-made surfaces (including buildings) in its output area: $\boldsymbol{\rho}(1178) = 0.606$, $\boldsymbol{P} < 0.001$. There are also strong relationships with building volume and number of addresses per building: $\boldsymbol{\rho}(1178) = 0.602$, $\boldsymbol{P} < 0.001$ and slightly weaker relationships with building area: and average building height, but no relationship with building count. This suggests higher levels of agents in areas with large residential buildings. The mild negative relationship with levels of natural surfaces would also support this observation.

Table 7: Spearman's ρ Correlation, All OAC Groups

spearman's ρ	building count	mm sf area no buildings	natural surfs	build area	building volume	mm surfs	road count	road area	addr / building	avg mx height
density	0.304	-0.471					0.309			
activities				0.457	0.539	0.606	0.42	0.567	0.599	0.513
agents			-0.325	0.527	0.602	0.688	0.473	0.626	0.652	0.538
entities	0.397		-0.374	0.466	0.455	0.439	0.338	0.455	0.344	0.322
num. editors/cell	0.411	-0.481	-0.332	0.394	0.396	0.451	0.497	0.457	0.403	
avg entity degree	-0.378	0.32								
agent power law		-0.315								
activity power law	0.376	-0.424								
entity-entity MFD	-0.365	0.451	0.317							
transient edit ratio		0.422								
quattrone maturity		0.372								
transitivity	0.361									

* Absolute relationships > 0.3 and 99% significant shown

6.2.1 OAC Supergroup Correlations

The OAC classification is obtained by clustering census output areas using their demographic characteristics, which means that their demographic homogeneity is greater within groups, and so one would expect to see stronger correlations if these factors are driving variation in the provenance variables. This seems to be the case, although the smaller group sample sizes will also inflate the value of the correlation coefficients. OAC supergroups one and three have much smaller sample sizes of 31 and 23 respectively (see Table 8), and these have much stronger correlation effects. Sample size is known to have a negative relationship with correlation coefficients that is thought to stabilise with samples greater than 150 [294]. Although this may vary depending on the phenomenon under investigation, it is likely the other OAC groups will experience small or minimal effects (Table 8). Most of the correlations are statistically significant, and robustness of the Spearman's rho statistic and bootstrapping procedure means that they are unlikely to be spurious, but for supergroups 1 and 3, the much higher correlation values cannot be solely attributed to demographic variation.

Table 8: OAC Group Sample Sizes

OAC	Sample size	%
1	31	2.63
2	141	11.97
3	23	1.95
4	125	10.61
5	356	30.22
6	212	18.00
7	138	11.71
8	152	12.90

When the correlations are assessed against Cohens' effect size criteria [295], several of the variables show a strong effect size ($\rho > 0.5$). Some examples are reported in Table 9 and Table 10. These variations in correlation effect size are typified by the correlations between the number of prov:Agents and building volume. The correlation across all groups is strong, but even stronger in the Rural Residents supergroup 1 , which has a sample size of 31 compared with 1178 for "all groups".

Table 9: Spearman's ρ Abstract Metrics

OAC group	agents: building volume	transitivity: building count	activities: addresses per building
1: Rural Residents	ρ(31) = 0.789, P < .001	ρ(31) = 0.807, P < .001	ρ(31) = 0.413, P = .021 *
2: Cosmopolitans	ρ(141) = 0 .489, P < .001	ρ(141) = 0.286, P = .001	ρ(141) = 0.312, P < .001
3: Ethnicity Central	ρ(23) = 0 .266, P = .220	ρ(23) = -0.217, P = .319	ρ(23) = -0.325, P = .130
4: Multicultural Metropolitans	ρ(125) = 0 .-0.002, P = .980	ρ(125) = 0.235, P = .008	ρ(125) = -0.1, P = .0267
5: Urbanites	ρ(356) = 0.404, P < .001	ρ(356) = 0.238, P < .001	ρ(356) = 0.340, P < .001
6: Suburbanites	ρ(212) = 0 .437, P < .001	ρ(212) = 0.529, P < .001	ρ(212) = 0.227, P < .001
7: Constrained City Dwellers	ρ(138) = 0 .259, P = .002	ρ(138) = 0 .176, P = .039 *	ρ(138) = 0 .352, P < .001
8: Hard Pressed Living	ρ(152) = 0.196, P = .016 *	ρ(152) = 0.417, P < .001	ρ(152) = 0.052, P = .523
All groups	ρ(1178) = 0.602, P < .001	ρ(1178) = 0.361, P < .001	ρ(1178) = 0.258, P = .001

Table 10: Spearman's ρ Concrete Metrics

OAC group	Transient edit ratio: Building count	Maintenance edit ratio: Building count	Editors per cell: Addresses per Building
1: Rural Residents	ρ(31) = -.580, P = .001	~~ρ(31) = -0.086, P = .646~~	~~ρ(31) = -0.327, P = .073 **~~
2: Cosmopolitans	ρ(141) = -0.106, P < .001	ρ(141) = 0.446, P < .001	ρ(141) = 0.474, P < .001
3: Ethnicity Central	~~ρ(23) = -0.038, P = .865~~	-~~ρ(23) = -0.334, P = .119~~	ρ(23) = 0.738, P < .001
4: Multicultural Metropolitans	ρ(125) = -.335, P < .001	ρ(125) = 0.363, P < .001	ρ(125) = 0.337, P <.001
5: Urbanites	ρ(356) = -.342, P < .001	ρ(356) = 0.184, P < .001	ρ(356) = 0.130, P < .001
6: Suburbanites	ρ(212) = -.401, P < .001	~~ρ(212) = -0.108, P = .117~~	~~ρ(212) = 0.005, P = .939~~
7: Constrained City Dwellers	ρ(138) = -0.090, P = .297	ρ(138) = 0 .338, P < .001	ρ(138) = 0 .438, P < .001
8: Hard Pressed Living	ρ(152) = -.217, P = .007 *	ρ(152) = 0.116, P = .156	ρ(152) = 0.143, P = .080
All groups	ρ(1178) = -.284, P < .001	ρ(1178) = 0.210, P < .001	ρ(1178) = 0.403, P = .001

* BCa CI excludes zero – reject H_0
** BCa CI intercepts zero – accept H_0

Sample size alone does not explain the variation in effect size between the supergroups. Some of these effects are likely to be related to features of the built and natural environment present in the output areas, and which are themselves characteristics of the output area supergroups. For instance, there are several correlations in supergroup 5 (n= 356), which are weaker than the smaller groups, e.g. the agents/Building volume correlation in the first column of Table 9 and maintenance edit ratio/Building count in the second column of Table 10. Although it is likely that the correlation coefficient value is inflated for these smaller supergroups, these correlations are statistically significant using both conventional confidence intervals and BCA bootstrapping, which means they are unlikely to be entirely spurious.

Abstract and Concrete Metrics. There is a marked difference in the strength and number of significant correlations between the abstract network graph metrics and the concrete maturity metrics shown in Table 9 and Table 10. The abstract metrics are more likely to be influenced by the structure of the map features, such as the number of member osm:nodes and number of shared member osm:nodes for each osm:Way, which in turn are dictated by the features being mapped. Variations in the physical environment such as housing/building types and land use are factors used in the clustering from which the classification was derived. This means they will contribute to variations in provenance graphs between OAC supergroups and will also be responsible for some of these correlation effects. There is less correlation with the concrete maturity variables. The most interesting are transient edit ratio, maintenance edit ratio and editors per cell, which have relationships with building count and addresses for building. These maturity variables are much more strongly influenced by individual contributor behaviour.

Abstract Metrics. We can see from Table 9 that the number of prov:Agents is positively correlated with building volume. i.e. the number of distinct individuals and distinct software used in editing the map increases with the volume of buildings present. We know that OSM users to preferentially map building features and are drawn to areas they find more interesting, so one might expect editing intensity to increase with building volume in these areas. This could explain why the correlation is weaker in the more deprived supergroups 7 and 8 which are characterised by higher unemployment rates and lower levels of educational attainment which would tend to make buildings less attractive to OSM contributors.

Transitivity is a calculation of the ratio of actual triangles to possible triangles (triples) in the network. It has much lower values in rural and non-built-up areas. One of the strongest predictors for this variable is the ratio of the number of osm:nodes to osm:Ways with higher values tending to have fewer nodes. Osm:Ways which share osm:Nodes create more triangles which increases transitivity values. This is often the case in urban areas with lower building completeness which are comprised of simple street outlines e.g. in Figure 40. Areas with a high actual building count, particularly those characterised by terraced housing, usually still have these street outlines. Where building completeness is low, their effects will dominate, and we would expect a tendency toward higher transitivity values.

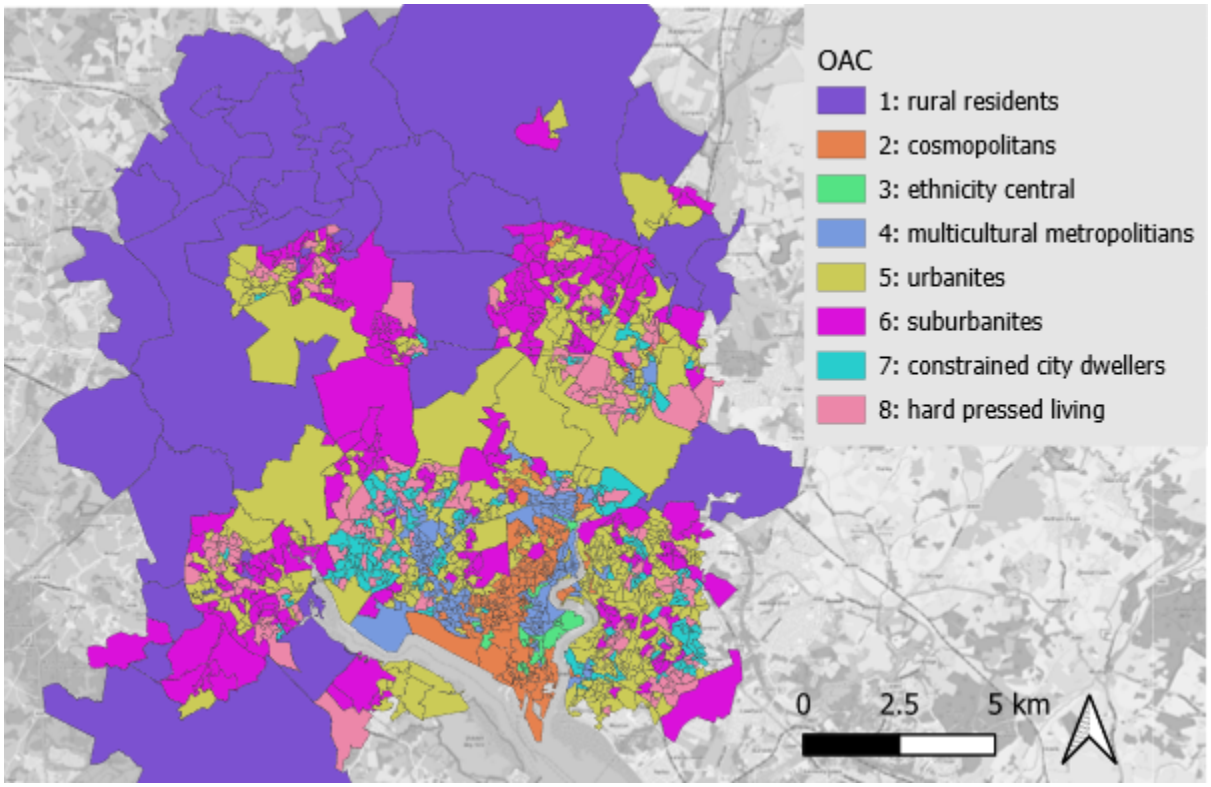
Over the dataset as a whole, transitivity has a moderate correlation with building count (see Table 7). For the supergroup 2 (cosmopolitans) we would expect generally higher completeness levels because of generally higher educational attainment in this supergroup, and here we see a weak correlation. In supergroups 5 (Urbanites) and 8 (Hard-Pressed Living) where we might expect lower building completeness, the correlation effect is greater. Supergroup 7 is an area of below average education qualification levels [237], where one might expect lower map completeness and a strong correlation with building count, but here, the correlation is much weaker. As a whole this group are more likely to live in flats. In the study area this supergroup is mostly composed of group 7a (challenged diversity), who are more likely to live in terraced properties than the supergroup. They also have higher proportion of younger people working in the information and communication industries, which may explain this anomaly.

Concrete Metrics. *Transient edit ratio* is the proportion of edits to tags reverted to their previous state within one month. It has a week negative relationship with building count across the entire dataset. This is weakest in supergroup 2 (cosmopolitans) which is characterised by the higher levels of educational attainment. In Southampton, these supergroup 2 areas tend to be quite close to a university campus and are dominated by groups 2a and 2b, which both have a high student population. This building count in supergroup 2 also has a positive relationship with maintenance edit ratio which, represents the proportion of edits to features in an output area resulting in a version number greater than 1. This means that as the number of mapped buildings increases there is more “tweaking” and re-editing of building features to enhance data, but a slower increase in the correction of problematic edits than one might see across the study area as a whole. It would also confirm the idea that many OSM contributors are students and that contributors map areas local to them more frequently and confidently. Looking at supergroup 5 (Urbanites) this pattern is reversed; with an increasing mapped buildings we see a slower increase in maintenance editing and a stronger decline in the proportion of transient edits to tags.

6.3 Thematic maps

Because of the complex nature of the geometries involved, relationships between the spatial clusters of provenance variables and OAC groupings are quite hard to assess visually. The map in Figure 37 shows the output area boundaries within the study area and their classifications.

Figure 37: UK 2011 Census Output Area Classifications – Southampton Area, UK



Output areas are classified into one of 8 OAC supergroups, and we can make broad statements about the probable characteristics of the human built environment within that output area, based on its classification. This is with the proviso that there will always be a degree of variability within each output area. Based on the ONS Pen Portraits document [237], we summarise the characteristics for each 2011 OAC supergroup in Table 11. We can also derive more granular information by considering the child groups and subgroups. It should be noted that these are typical characteristics of Output Areas, and in reality, there is some degree of heterogeneity.

Table 11: OAC Supergroup Characteristics Based on ONS Pen Portraits Document [237]

2011 OAC supergroup	Housing	Population density	Transport	Employment/education
1: rural residents	Larger detached properties above-average communal establishments	low	motor vehicles	agriculture, forestry, fishing
2: cosmopolitans	Flats and communal establishments, privately rented	High	Public transport, cycling	Full-time students, accommodation, information, communication, and financial industries
3: ethnicity Central	Flats	High	Public transport	Higher unemployment, accommodation, information communication financial and administrative industries
4: multicultural metropolitans	Terraced housing	High	Public transport, households less likely to have multiple vehicles	Transport and administrative related industries, families with children at school or college
5: urbanites	Flats and terraced housing	Dense in southern England, less so elsewhere	…	Lower than average unemployment, information communication, financial, public administration, and education sectors
6: suburbanites	semi-detached or detached properties	medium	Private transport	Below average unemployment information and communication financial public administration and education sectors
7: constrained City dwellers	flats, social housing	dense, higher levels of overcrowding	…	Above-average unemployment underrepresentation in information and communication, and education sector. lower qualification levels than nationally
8: hard-pressed living	semi-detached or terraced properties, social housing	medium	…	Mining, manufacturing, energy, wholesale and retail, transport, higher unemployment rates, smaller proportion of people with higher level qualifications

6.3.1 Visual Clusters

Visual inspection of thematic maps derived from the provenance variables reveals clusters of high and low measurement values which are suggestive of some degree of determinism and spatial

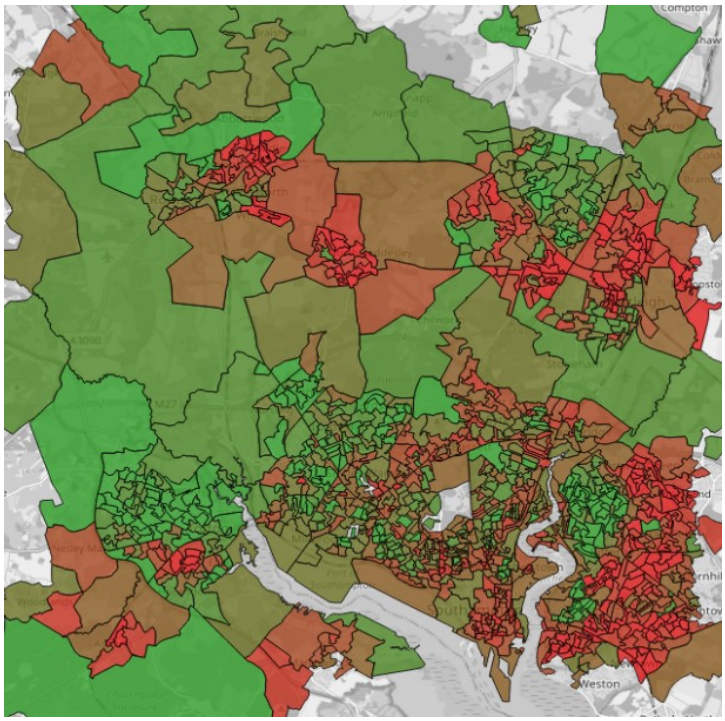
variability. In this section we examine three of the most striking patterns, described in Table 12, alongside measurements of the physical environment and examinations of the map content using the OSM querying tool and visual inspection

Table 12: Spatial Patterns Assessed by Visual Map Inspection

type	description
Pattern 1	Differentiation between high and low values (reds and greens) in two zones in the south-east of the study area, shown in Figure 39
Pattern 2	Low values in an inverted triangular zone in the north-east of the study area, as seen in Figure 35
Pattern 3	A tendency for higher values centred within clusters of populated urban areas, lower values at the peripheries
Pattern 4	A distinction between urban and rural areas

The map in Figure 38 also shows two of the most distinctive patterns which occur for many of the variables:

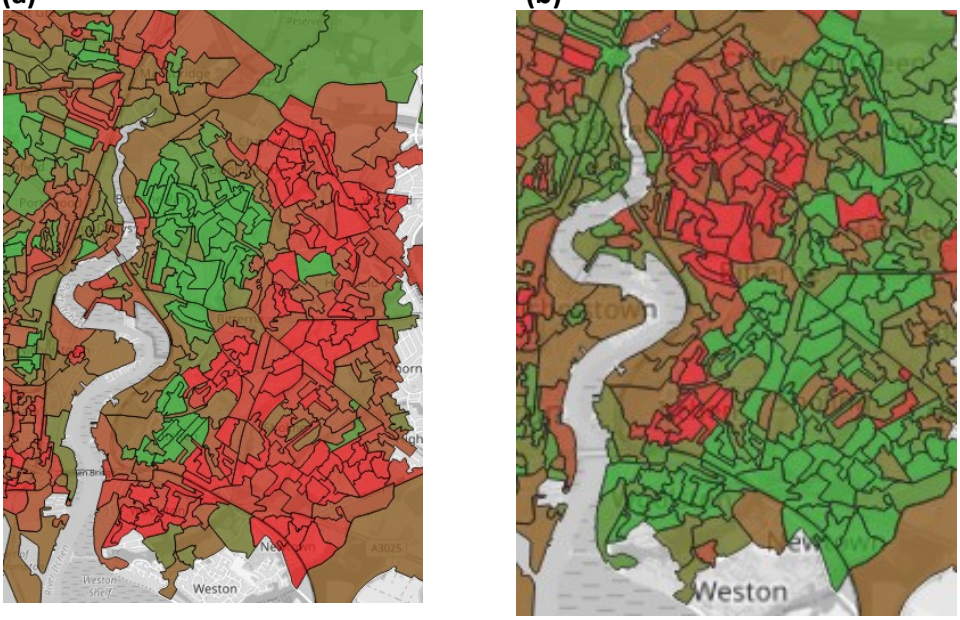
Figure 38: Prov:Agents Count by Output Area



Pattern 1. Pattern 1 occurs as two contrasting zones to the east of the River Itchen in Southampton. The north-west of this zone is differentiated from the south and east, most commonly with lower measurement values indicated by green shading. For some variables this pattern occurs in reverse, as indicated in Table 13 and shown in Figure 39

Figure 39: Pattern 1: High Density Value Clusters in the South-East of the Study Area (Left, a), and Reversed Pattern 1: Low Prov:Entity Count Values in the South-East of the Study Area (Right, b)

(a) **(b)**

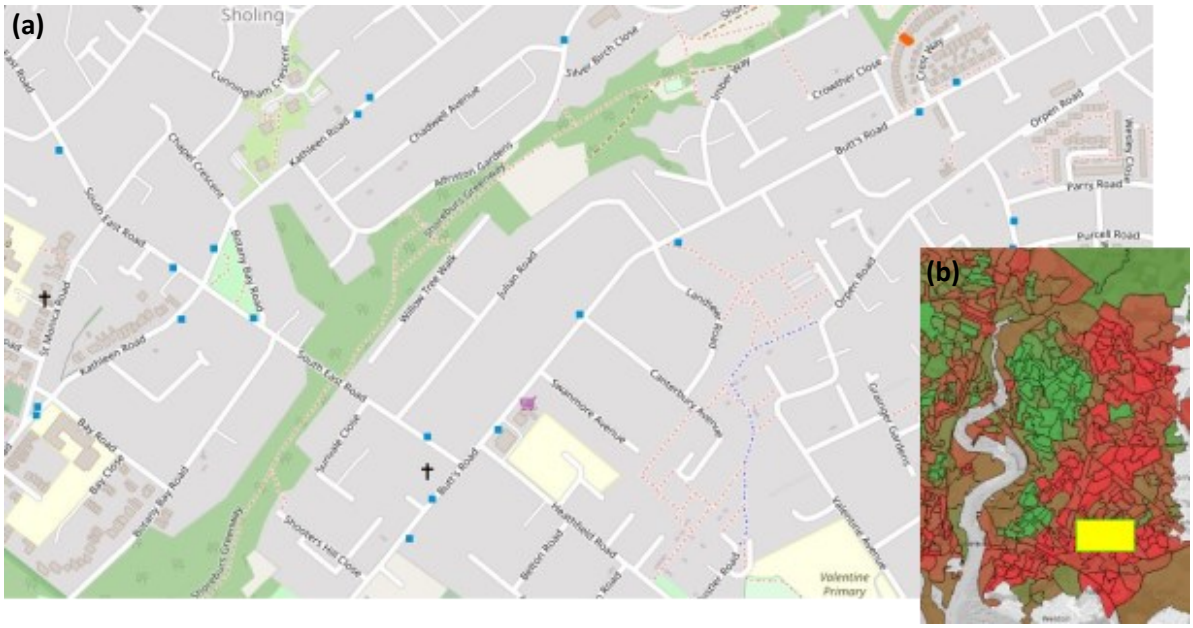


Demographically, the area is almost entirely classified in OAC supergroups 5, 6, 7 and 8, with supergroup 5 being prevalent, comprised of groups 5a and 5b. The green region shown in the north-west of Figure 39 (left image) is primarily from the urbanites (supergroup 5) and belongs to subgroups 5a1, 5a2, 5b1 and 5b3. These child OAC groups provide more granular information and give more indication of specific building types. Terraced housing is likely to be dominant here, but a significant number of these child groups also indicate detached housing and flats. The east side of the area has a wider diversity of OAC classifications, with more output areas belonging to group 6 (suburbanites), group 7 (constrained city dwellers) and group 8 (hard-pressed living). Groups7 and 8 tend to have higher deprivation indices, but looking at the subgroups suggests a similar range of building types [237]. Examination of the Ordnance Survey derived physical environment variables also does not show the clustering one might expect if these patterns were driven by human environment factors such as building or surface type.

Visual inspection of this region's OSM map reveals that the eastern half appears to have been less completely mapped than the northwest (see Figure 40). Our maturity metrics also suggest this. E.g. *Average days since last edit* is higher on the eastern cluster and the prov:Entity count

(*entities*) much lower. Further examination with the OSM query tool shows that much of the map in this area was edited at least 3 years ago and often 10 years ago. The low map completeness also explains why many of the Ordnance Survey derived built environment variables do not reflect the same clustering patterns, because the features in the Ordnance Survey dataset have not been mapped in OSM.

Figure 40: OSM Map Content (a) on the East Side of the Pattern 1 Zone (b) Map Area in (a) Shown by Yellow Box in (b)



In contrast, the north-western region shows signs of having been recently edited, with higher values for *New Edits* and *Maintenance Edits*. This area is also generally uniform and contiguous, which is suggestive of editing activity characteristics, i.e. an individual focusing on, and completing work in a region and then continuing to map adjacent regions. The high activity and agent degrees would also tend to confirm this. In Chapter 5, Section 5.2.2, we identified high values of these metrics as being indicative of concentrated activity by a single user. We also found high activity degree values to signify expert editing, i.e. intensive mapping with no subsequent edits needed, owing to the high quality and completeness of the work.

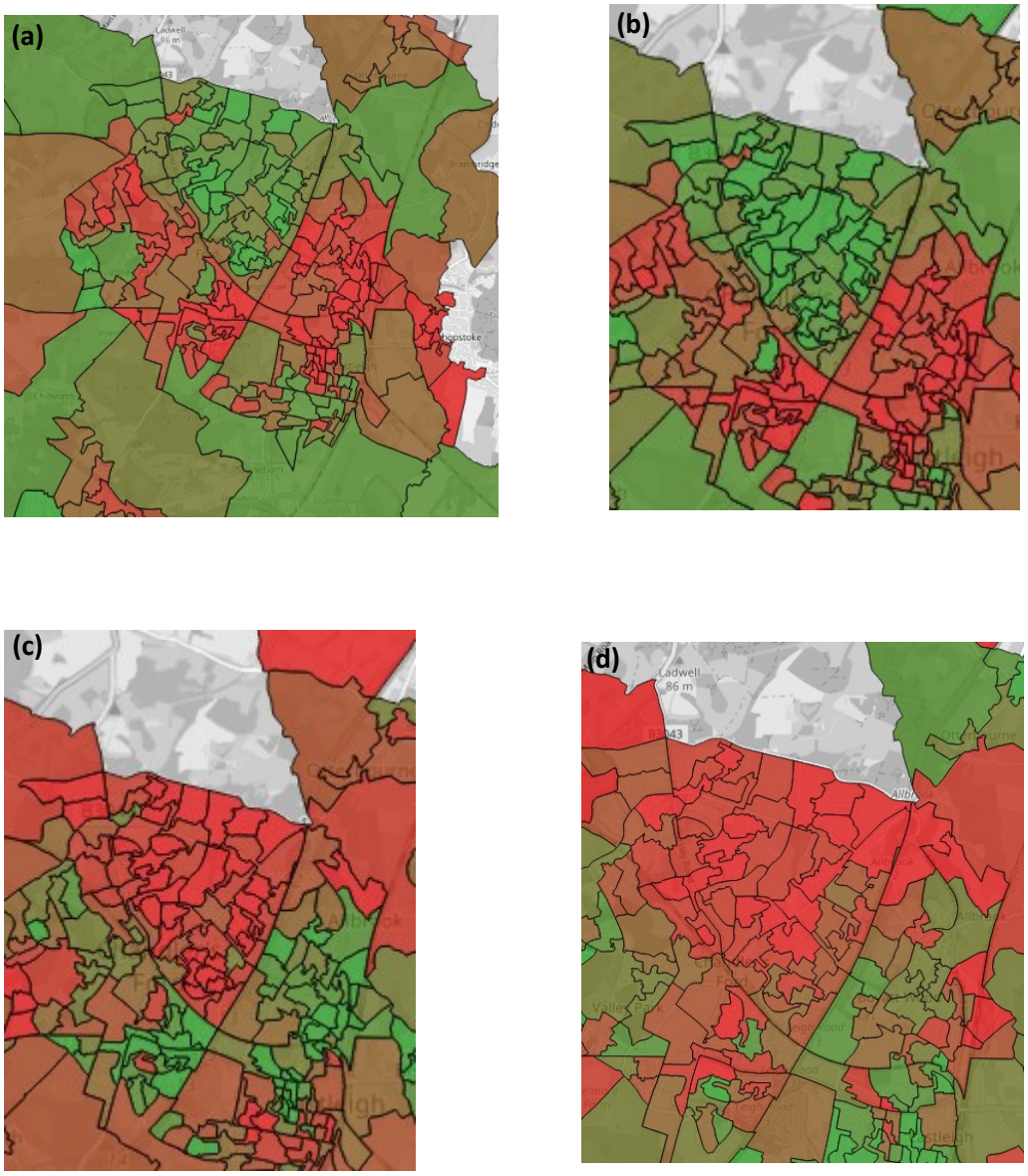
Examination of the area using OSM's query tool shows that almost all of the data in this area was indeed created by a single user. Frequent changeset annotations such as "added house and building footprints, postcodes inferred from code point open centroids…" also suggest that this individual has geographic expertise. Most of the data queried resulted from changesets created two or three months before the download of the history data. Interestingly, further east is an anomalous output area with much higher values than its surroundings and more consistent with the north-

western zone. Use of the OSM query tool shows that this is an output area containing a primary school which has also been the subject of intensive editing by another single user. This individual's changesets are all centred on the area, so they are presumably local.

Figure 41: OSM Map Content (a) in the North West of the Pattern 1 Zone (a). Map Area in (a) Shown by Yellow Box in (b)

Pattern 2. This zone is mostly made up of output areas from OAC supergroups 5 and 6, urbanites and suburbanites respectively. Most of the child groups and subgroups are also represented, suggesting high probabilities of a variety of housing types including flats, communal establishments, and detached, semi-detached and terraced housing. As with pattern 1, there seems to be no visual discernible relationship with the Ordnance Survey variables except for *Natural Surfaces*, which unsurprisingly, are at lower levels than the surrounding rural areas. The east side of the triangle is delineated by a motorway which results in higher levels of natural surface owing to verges and strips of woodland on each side. The west side of the triangle is delineated by railway and both this and the motorway are used by the output area zoning algorithm to clip the output areas. This accentuates the pattern zone and although there is clearly non-random clustering occurring here, it is less distinct than that described in Pattern 1.

Figure 42: Pattern 2 Clusters, North-East of the Study Area; (a) Density; (b) Revert Count; (c) Average Activity Degree; (d) Average Clustering Coefficient for Entitles



Visual inspection of the map reveals differences in map completeness either side of the delineating railway/motorway. Inside the pattern 2 zone there are extensive building footprints, whereas on the other side of the railway/motorway these are mostly absent, despite the OS data showing little difference in actual building volume and the similarities in the area types indicated by the OAC groupings. This suggests that map features may have been affecting the behaviour of map editors, who seem to be using these linear features as containers within which they conduct mapping activity. Figure 43 shows examples of this over the study area, including in the map content for the pattern 2 zone.

Figure 43: Examples of Linear Feature Delineation of Map Completeness: Chandlers Ford, Southampton (a and b) Showing a Railway Boundary (a) and Motorway (b); Sholing, Southampton (c) Shows a Street Boundary (Bursledon Rd)

Visual inspection, and examination of features and changesets in the area using the OSM query tool reveals a similar and largely complete set of building footprints to those in the pattern 1 zone. However, this data is much older, with changesets ranging from 3 to 10 years old and several versions of most features. A typical building footprint in this zone has 7 or 8 versions, with components such as building outline, house number, drives and service roads added in different changesets often by two different users over a few years. This contrasts with the pattern 1 region, where the content is mostly created in one changeset, and the whole region mapped to a high level of completeness in the space of a few weeks.

Table 13 shows some discrepancies between the occurrence of pattern 1 and pattern 2 zones among the provenance variables. These suggest that map completeness is not the only factor here, and that editing dynamics are affecting some of the variables differently. Collaborative editing, currency, user expertise and editing intensity are all having an effect. It is also likely that these effects are much stronger in more recent data, and where the measurement values seem to be contiguous.

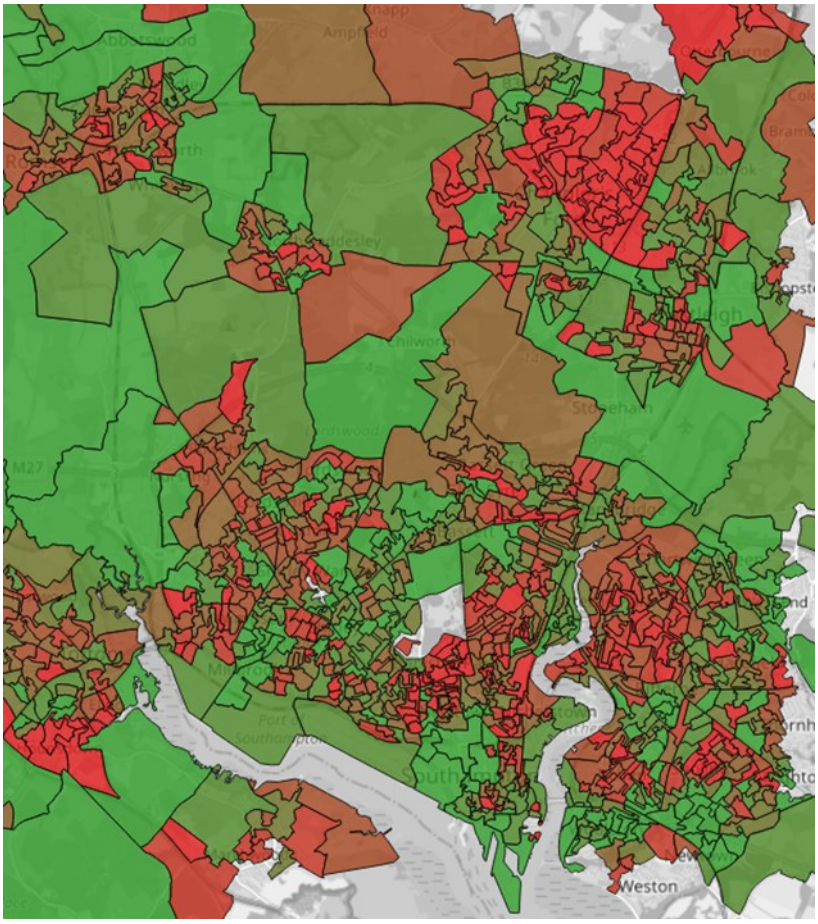
The Activity Clustering Coefficient values illustrate this. The north-west of the pattern 1 region is characterised by low values. In the pattern 2 region, these values are generally higher and less uniform. This supports the finding from Chapter 5 that low values of this variable are characteristic of intensive editing by an expert user, and that higher values indicate more collaborative editing over a longer period. The intrinsically more informative maturity measurements tend to support the characterisations derived from the network metrics. Quattrone maturity has higher values, showing that the north-west pattern 1 and pattern 2 areas have higher completeness. Revert rate is much lower for the north-west pattern 1 zone, showing that it has undergone little editing once created, whereas higher values in the pattern to zone suggest older data and more collaboration.

Table 13: Variables Showing Patterns 1 and 2

	Pattern 1 high/low value clusters in the south-east of the study area (Figure 39)	**pattern 2 high/low value clusters in the north-east of the study area (Figure 42)**
Activities	yes	yes
Entities	yes	yes
Agents	yes	yes
Average clustering	yes	yes
Average agent degree	yes	yes
Average Activity degree	yes	yes
Density	yes	yes
Edges	inverted	inverted
Nodes	inverted	inverted
Rich Club Coefficient	inverted	inverted
Transitivity	yes	inverted
Avg degree centrality	yes	yes
Density	yes	yes
Average clustering	yes	inverted
Maintenance ratio	yes	yes
Number of editors per	yes	yes
Average creators per	yes	yes
Quattrone maturity	inverted	inverted
Transient ratio	inverted	inverted
Interactivity	yes	no
Edits per cell	yes	no
Days since last edit	yes	inverted
Interactivity	yes	inverted
Life-cycle edits	yes	inverted
Maintenance ratio	inverted	inverted
New edits currency	inverted	yes

Pattern 3. Pattern 3 clustering refers to high or low values in the smaller more densely populated output areas with a tendency for the highest values to occur in more central areas with lower values on the peripheries. Examples of pattern 3 clustering can be seen in Figure 44.

Figure 44: Pattern 3 Clustering of Average Rich Club Coefficient



Pattern 4. Pattern 4 clustering is a distinction between urban and rural areas which can be seen after normalisation by polygon surface area. This pattern is distinct from the others in that it consistent with the Ordnance Survey MasterMap data and appears to be more strongly linked to environmental factors, although as with the other patterns, data completeness affects provenance variable values.

The census output area zoning algorithm optimises their polygons by population size, such that the population of each falls within a specified range: between 100 individuals, 40 households and 625 individuals 250 households. This means that larger output area polygons tend to be less densely populated, so those output areas which fall into the rural classification, OAC supergroup 1, are much larger than some of the more urban output areas, although those in the suburbanite group can also be quite large. Figure 45 shows output areas in OAC supergroup one: rural residents.

Figure 46: OS Topography Layer - Manmade Surfaces

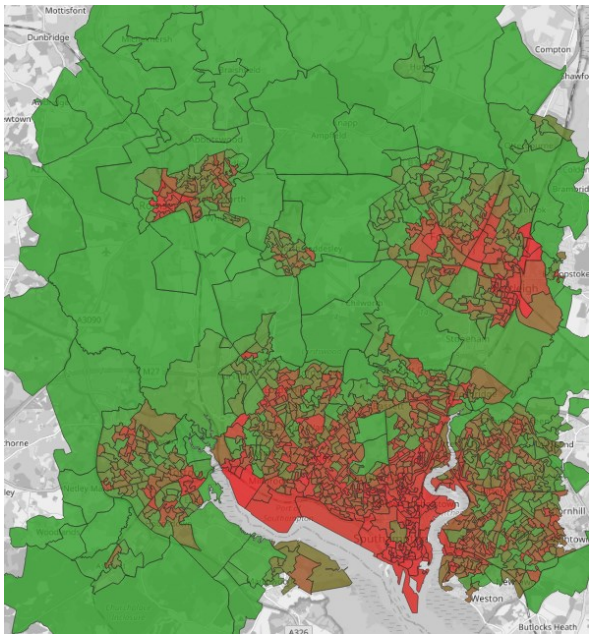
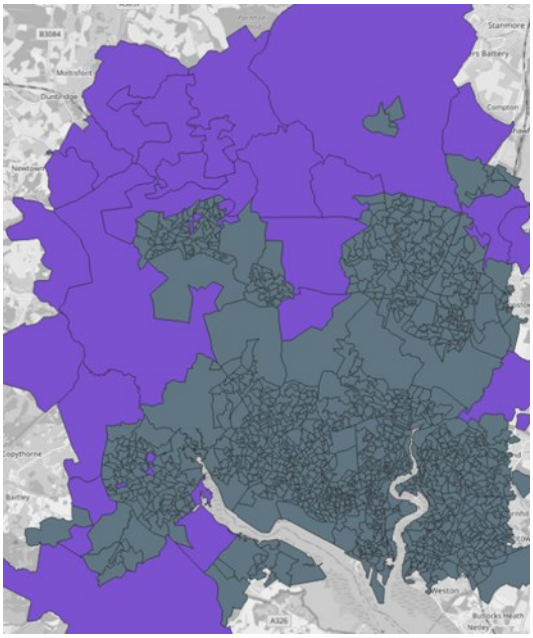


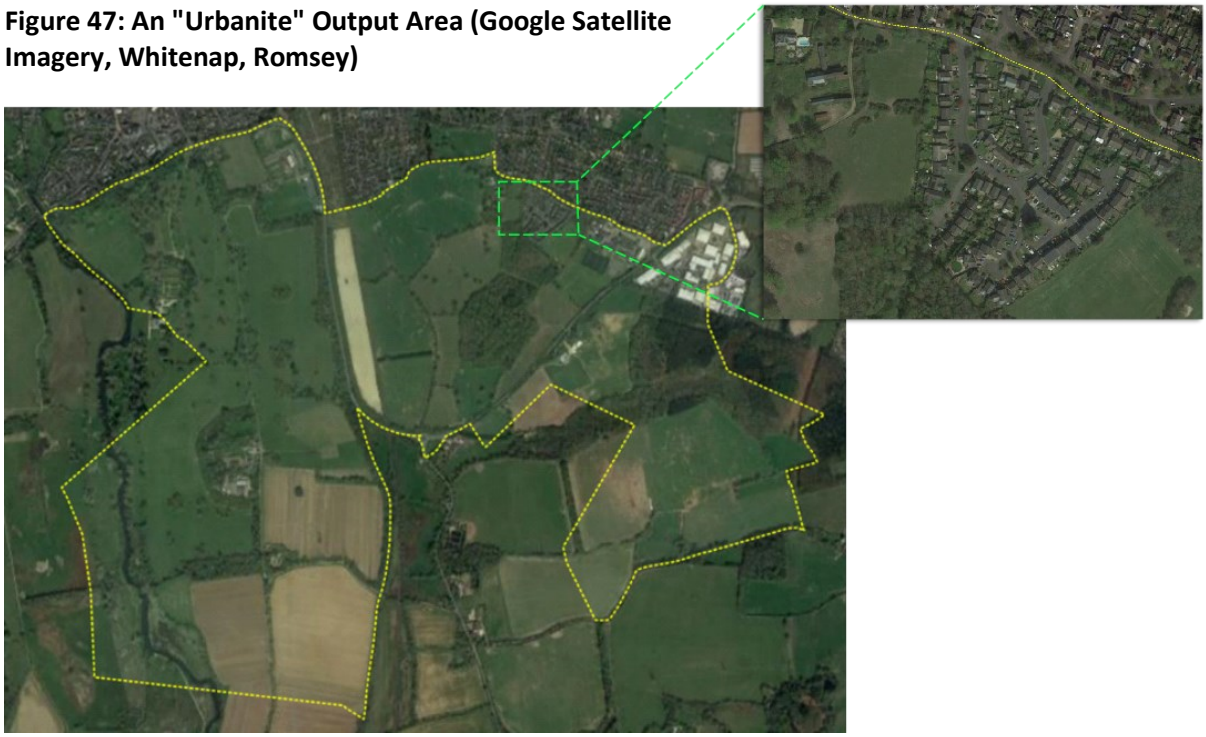
Figure 45: OAC Supergroups - The Urban/Rural Divide



We can see from many of the Ordnance Survey variables that there are marked differences between these output areas and the other supergroups. Given the obvious and profound differences in land use between urban and rural areas this is unsurprising However there are also some discrepancies and closer inspection of the OSM map and examination of the provenance variables reveals some interesting facets.

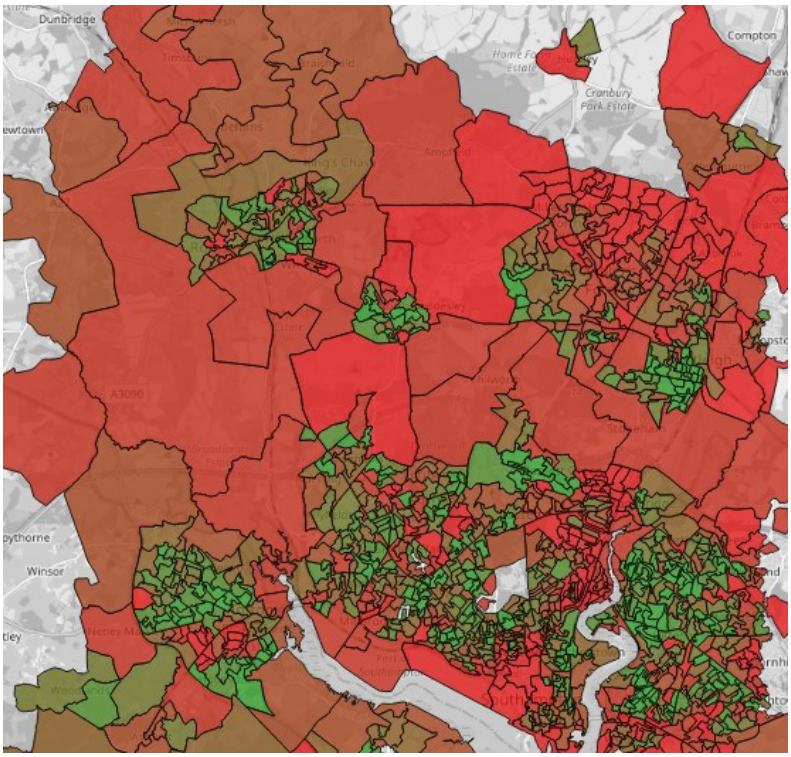
To the north of the main urban part of Southampton are several seemingly anomalous output areas which do not belong to the Rural Residents group but have similar man-made surface levels (see Figure 46). This is likely to be due to the output area zoning algorithm, which creates the areas using population characteristics [287], but must include all land areas. An output area can therefore be environmentally rural, but demographically urban. These output areas are all in the group 5 ‘urbanites’ class, but the classification seems to be due to the parts of the area which are inhabited. Most of these areas appear rural when viewed in satellite imagery even though they are demographically urbanite according to the OAC. The example shown in Figure 47 is clearly not a predominantly urban surface and seems to have obtained its classification from one small region.

Figure 47: An "Urbanite" Output Area (Google Satellite Imagery, Whitenap, Romsey)



Our man-made surfaces measurement seems to be providing a stronger indication of environment type than the OAC, and the urban rural divide is well reflected in these topography layer variables. Building height measurements do seem to pick out some of the output areas which are not demographically rural, but also have low values in some of the

Figure 48: Avg. Entity Degree



smaller more densely populated city zones and it is likely that MAUP effects are at play here with aggregated measurements more likely to be skewed by extreme values in larger areas.

6.4 Conclusions

In this chapter we have investigated our provenance variables using thematic maps of the study area. Visual inspection of these clearly reveals spatial patterns driven by some non-random, deterministic spatial phenomena. We have investigated by zeroing in on specific regions of the study area where some distinctive clustering of either high or low values is visually obvious. We compared these clustered measurements with "ground truth" physical environment data from the Ordnance Survey, and demographic data from the UK 2011 census output area classification. Using this information, and insights into the nature of our measurements derived from examining the network graphs in Chapter 5, we have described and explained several of the more distinct clusters found in thematic maps of the study area.

The provenance variables correlate in varying degrees with measurable aspects of the human, physical and built environment, derived from the Ordnance Survey MasterMap topography layer. These correlations exist when variables are assessed across the entire dataset and can also be seen within OAC supergroups where the effect strength varies between classifications. In 2 of the groups these effect sizes are likely to have been inflated by the smaller sample size but in many cases are statistically significant effects. Examination of thematic maps, MasterMap derived built environment variables, OAC supergroups and child groups, and their pen portraits, along with OSM coverage provides an indication of potential drivers of these correlation effects. The physical and built environment is a source of variation, but the effects of individual contributor behaviours are also reflected in the measurable structure of provenance graphs. Map completeness is also an important source of variation in the provenance variables in built up areas.

We have also identified visually discernible clusters of high and low values for our provenance measurements which are evident on choropleth maps of the study area aggregated by census output area. These patterns are not consistent with variation in the physical and built environment or levels of map completeness and examination of the OSM coverage and changeset data provides evidence that individual contributor dynamics plays a role in the occurrence of these patterns. For example, both of the pattern 2 zones have similar coverage in that they are well mapped residential areas with a more or less complete set of building footprints. However, they are clearly differentiated on thematic maps of provenance variables. They differ in that the pattern 1 area was much more recently and intensively mapped, whereas the pattern 2 area has content that was created less intensively over a longer period by more than one user.

The interaction between map completeness, contributor dynamics and the physical environment as drivers for the provenance variables means that identifying precise causes of variation in provenance measurements is likely to be problematic. The provenance graphs for each output area capture detailed provenance for every feature and so carry a range of “signals”, e.g. structural feature characteristics, individual contributor characteristics and physical environment characteristics. This means that the characteristics of an output area provenance graph will be affected by its roads, buildings, and natural and man-made surfaces and how individual contributors interact with these. It is highly likely that capturing more specific provenance graphs e.g. focusing on building footprints or street networks will provide more detailed and specific insights into the way OSM is created.

Chapter 7 Metric Analysis

7.1 Introduction

In this chapter we perform three sets of statistical analyses which cement the insights we have gathered from close inspection of the provenance graphs and consideration of the spatial and physical context of the provenance in the previous two chapters.

In the first section we look at concrete maturity metrics to assess the extent to which they represent data maturity, and what relationship they have with proxy estimates of data quality. This assessment concrete provenance measurement addresses research question one by evaluating this approach, which is distinct from network analytics. The results also provide some insights into the relationship between contribution patterns and data quality proxies (research question two).

In the second section we perform an exploratory factor analysis which uncovers latent variables which cannot otherwise be directly measured, and which provide insights into the drivers of variance among provenance graphs in terms of specific contribution patterns and how these interact with the physical and built environment (research question two).

In the third experiment we investigate potential demographic drivers of variation by using MANOVA to investigate whether the characteristics of provenance graphs vary according to UK census output area classification supergroups. Using discriminant function analysis as a follow-up procedure we investigate the details of the distinctions found between these supergroups, identifying those which allow us to predict group membership significantly better than by random chance. This demonstrates insights about map contribution behaviour and how it is affected by the demographic characteristics of the map area (research question two).

7.2 Investigating Data Maturity in OSM

In Chapter 3 we describe two approaches to the measurement of provenance graphs: abstract measurements, derived using graph theory and network properties of provenance graphs, and concrete measurements, which assess more tangible aspects of provenance. Our concrete measurement strategy is based on a concept of data maturity which uses research into user generated content as a theoretical framework (see Chapter 3, Section 3.3.4). Research question 1 asks how useful insights can be gained from different approaches to the measurement of provenance graphs. To address this, we need to know whether the concept of data maturity as we define it, bears any relation to other assessments of data quality/map maturity. If this is the case, it

shows how research into user generated content can be used as a theoretical basis to design concrete metrics for targeted analytics.

In this section, we provide an experimental evaluation of our maturity metric. We derive two summary measures of data quality/maturity using metrics based on comparison with satellite imagery and using the output of an automated error detection engine. We assess relationships with maturity metrics using Spearman's correlation coefficients.

7.2.1 Measurements Implementation

The output area geometry is the output of a zoning algorithm which uses demographic variables to generate polygons [221], [225]. It can result in erratic shapes with wide variations in size. In areas with high population density, polygons can be too small for practical visual assessment, and in some rural areas they can be too large. For this investigation we used the same techniques described in Chapter 4 but used hexagonal grid cells as the extraction geometry, rather than output areas.

Some of the maturity measurements we use are affected by the number of OSM primitives. Measurements such as edit count are clearly related to the number of primitives within a cell, i.e. 400 edits in a cell containing 30 primitives is clearly a different value to 400 edits in a cell containing 10,000. For these arbitrary hexagons there is also wider variation than for output areas. The affected measurements are edit count, edit count, new edits count and transient edit count. To make these more meaningful we weight their values according to the number of OSM primitives in the cell.

7.2.2 Assessing Maturity Metrics Using Proxies for Data Quality

In this section, we assess the extent to which our concrete maturity metrics reflect real world map maturity as defined in Chapter 3, Section 3.3.4 in OSM data, by assessing the relationship between maturity measurements and two other proxy indicators of OSM map data quality: visual survey results and the output of an automated error detection engine.

Visual Review. To conduct the visual review, we chose 30 grid cells, 15 in rural/non-built-up areas and 15 in urban, built-up areas. These were loaded into QGIS along with a layer of satellite imagery. Urban/rural areas were identified using visual inspection of the satellite imagery layer. The OpenStreetMap data was then compared to the satellite imagery using a set of predefined criteria. Each hexagon was given a 4-point quality score in each of 3 categories. Slightly different evaluation criteria were used for urban and rural areas as shown in Table 2. For both areas the criteria were based mostly on data completeness, but also considered semantic and positional accuracy.

***Urban*.**

buildings	**Score**
accurate polygons with name/number no significant inaccuracy	4
polygons, possibly some inaccuracy	3
mostly represented as nodes, or few polygons, several inaccuracies	2
not delineated (or few nodes+.5)	1

roads

all named streets / footpaths and tracks, no noticeable omissions	4
1 or 2 unnamed streets/ unmapped footpaths and tracks. minor inconsistency inaccuracy	3
3 -5 unnamed, or inconsistently mapped, e.g. some alleys mapped as streets others omitted	2
unmapped major roads, more than 5 unnamed streets	1

green areas

individual trees, accurate hi-res green areas	4
small blocks of woodland, some green areas, occasional inaccuracy	3
large woods only, large areas of unmapped green, inaccurately mapped areas	2
most green areas unmapped	1

***Rural*.**

Buildings

polygons with name/number, accurate complete structures	4
accurate polygons, occasional missing structures	3
merged buildings, minor omissions, some inaccuracy, many missing structures	2
significant missing buildings	1

Land

land use boundaries, small woods, named woods, small detail e.g. Watercourses, gates	4
larger woods only, many unnamed, some boundaries missing	3
woods only, no names. Many significant boundaries missing	2
no delineation	1

Roads

all roads, tracks, streets are accurate, named and appear complete	4
accurate Roads tracks streets, 1, or 2 missing, many unnamed	3
a missing road, many missing tracks, substantial misrepresentation e.g. track as metalled road	2
substantial part of network absent	1

In the study area there are five distinct regions: central Southampton, Eastleigh/Chandlers Ford, Hedge End, Romsey area, and Totton/Hythe. 6 cells were selected randomly from each to give as broader coverage of the study area as possible. For the rural land coverage, buildings and structures were part of the assessment criteria and so cells with none of these were ignored. Each selected grid cell was given a total score providing subjective summary measure of OSM map quality.

We then calculate correlation coefficients with the maturity variables to see if there is any relationship. Because several of the maturity variables do not have a normal distribution, we opted for the Spearman's rho coefficient. This statistic is suitable for sample sizes of 30 and above, but because we only have the minimum sample size, we opted to also perform 1000 sample bootstrapping to provide enhanced significance scores. Average creators per feature and maintenance ratio both had moderate correlations at 95% significance and average edits per feature at 99%. For revert count the P value was 0.055 indicating that, although borderline, this correlation was not significant, however the BCa 95% confidence interval did not intersect zero, which means this correlation was significant following bootstrapping. Therefore we can reject the null hypothesis for the variables shown in Table 14, and conclude that these variables correlate significantly with our quality score.

Table 14: Survey Correlations (Spearman's ρ) Between Maturity Metrics and the Survey-Based Quality Measure

avgCreatorsPerFeature	Correlation Coefficient		-.418*
	Sig. (2-tailed)		0.022
	Bootstrap BCa 95% Confidence Interval	Lower	-0.707
		Upper	-0.053
avgEditsPerFeature	Correlation Coefficient		-.505**
	Sig. (2-tailed)		0.004
	Bootstrap BCa 95% Confidence Interval	Lower	-0.746
		Upper	-0.108
Maintenance Ratio	Correlation Coefficient		.390*
	Sig. (2-tailed)		0.033
	Bootstrap BCa 95% Confidence Interval	Lower	0.066
		Upper	0.675
RevertCount	Correlation Coefficient		0.354
	Sig. (2-tailed)		0.055
	Bootstrap BCa 95% Confidence Interval	Lower	0.001
		Upper	0.656

7.2.3 Summary

The negative correlations between the survey scores and average edits in editors per feature seems to be counterintuitive. These maturity measurements were conceived as metrics to gauge the extent to which OSM data is affected by Linus's law. This is the idea that "many eyes make bugs shallow" i.e. the probability of high-quality data increases with the number of humans which interact with it. This result is inconsistent with that idea. The effect may be due to the limitations of our experiment, which is carried out on a small sample, using a crude indicator of data quality. The high values per feature could be due to a paucity of features, the presence of a few heavily edited features, or the specific quality dimensions being assessed. Linus's law is after all, a predictor of the presence of "bugs" rather than levels of specific geographic data quality dimensions.

In Chapter 5, we found that intensive editing was a pattern which left signatures in graph network metrics. These occur where a single user edits an area to completion and to a high standard, moving on once all editing in an area is complete. This seems to be the hallmark of an expert and is consistent with our negative correlation. These areas will often only have one or 2 editors because once they have contributed to an area, there is little scope for further editing, and this would explain the negative correlation with edit count. The intensively edited graphs we examined in Chapter 5 generally had the work completed in one or 2 edit sessions, which would also explain the negative correlation with edits per feature.

Maintenance ratio decreases as the proportion of maintenance edits increases, so the positive correlation indicates that the survey score decreases with increasing proportions of maintenance edits. This and the positive correlation with revert count, a measure of the number of tags reverted to their previous value within one month is also likely to be a hallmark of intensive expert editing. The completeness and quality of these type of contributions leaves little scope for further alteration. In Wikipedia, reversion rates are regarded as a characteristic of mature content, where editing is seen as undesirable, and edits are more frequently reverted back to a previous state. Our results provide no indication as to whether this is a factor in or study. In Wikipedia, important articles are often "watched" by members of the editing community , who can be quickly alerted to changes. No such facility exists in OSM, so it is likely that edit reversion rates have a different significance.

The small sample size and skewed distributions of this data mean we cannot rule out the existence of other systematic variations in the values of our maturity metrics. There is evidence for four of the maturity metrics, shown in Table 14 having some direct relationship with visually assessed data quality. The metrics all assess levels of editing activity, and two of them were

conceived using assumptions derived from Linus's law. However we found that Linus's law does not apply as expected in our results and that there is a negative relationship between the number of contributors and our survey score, which is consistent with the characteristics of expert/intensive editing which we identified in Chapter 5.

7.3 OSMOSE

The **O**pen**S**treet**M**ap **O**ver**S**ight **E**ngine (OSMOSE) is a web application which detects 233 error types in the OSM map. Typical errors flagged include tag misspellings, orphaned nodes, and duplicate objects. The error data is obtained from the OSMOSE API, which we convert into RDF, encoding positional information as WKT coordinates. This allows us to count the errors in a grid cell using GeoSPARQL. We hypothesise that the number of errors in a grid cell provides an indication of its quality, and that mature data will have fewer errors.

Approximately two thirds of the cells had no errors, but the number of primitives is highly variable, and another assumption we make is that the more primitives a cell has the higher the probability of an OSMOSE error existing, so a cell with a large number of OSM primitives that has no errors should have a much lower rating than a cell containing few primitives and no errors. Our error rating is therefore derived by $X = \frac{(e+1)}{Os}$

…where **e** is the OSMOSE error count, and **Os**, the primitive count. This gives us an odds type value which reflects the error incidence vs the probability of their occurrence.

We then calculated correlation coefficients for each of the maturity variables. Because many of the skewed distribution of many of the variables and the presence of extreme values, we opted for the nonparametric Spearman's rho correlation coefficient as a more robust procedure than the alternative Pearson's R. The results are shown in Table 15. All the correlations were significant at the 99% level. The strongest relationship was with editors per cell where we found a moderate correlation with OSMOSE errors: $\rho(453) = 0.539$, $P < 0.001$, allowing us to reject the null hypothesis that there is no relationship between the two variables.

Table 15: OSMOSE Correlations

	Spearman's ρ	p-value (2-tailed)
Editors/Cell	0.440	< 0.001
Avg. Creators/Feature	0.316	< 0.001
Avg. Edits/Feature	0.349	< 0.001
Days since Last Edit	-0.163	< 0.001
Edits/Cell	-0.539	< 0.001
Maintenance Ratio	-0.236	< 0.001
Revert Count	-0.352	< 0.001
Transient Ratio	-0.511	< 0.001

7.3.1 Summary

Table 15 shows a range of weak and moderate correlations with the OSMOSE error score. Revert count, transient edit ratio, and maintenance ratio are all variables which relate to the editing of existing data. Along with edits per cell, they are also indicators of edit intensity. The editors per cell, and creators and edits per feature variables seem to be consistent with the findings in the previous section, i.e. they are positively correlated with the error rate, which is not consistent Linus's law.

Transient ratio represents the proportion of edits to tags which are reverted within one month, and revert count is a simple count of those returned to their previous state in any subsequent edit. They are a measure of volatility in the data. These measures do not indicate how recently the tag reversions took place, but the negative relationship with the OSMOSE score suggests that the error rate is lower if data shows signs of having been volatile at some stage. This contradicts the findings in the previous section and suggests the automated detection of specific errors carried out by OSMOSE provides a different measure to the more general assessment from visual survey.

7.4 Conclusions

Both investigations in this section shed light on research question one by uncovering significant correlations between concrete provenance maturity metrics and assessments of OpenStreetMap data quality, which suggest that some of these concrete maturity metrics can be useful as an automated means of predicting aspects of data quality. The maturity metrics were partly derived from research into Wikipedia, based on the assumption that these user generated content platforms may have similar characteristics. These results suggest that that is not necessarily the case, and that although notions of maturity in Wikipedia bears some relation to those of OpenStreetMap, the drivers for variation are not always the same.

One unexpected result is the apparent contradiction of Linus's law. This rule has been found to apply in OpenStreetMap for positional accuracy [44], but in our more generalised Survey assessment, the number of editors, and human interactions with the data seems to be negatively correlated with estimates of its quality for both OSMOSE and the visual survey. This result is consistent with our assessment of provenance network metrics, particularly the theme of expert editing, where data is comprehensively edited to a high standard by a single expert user.

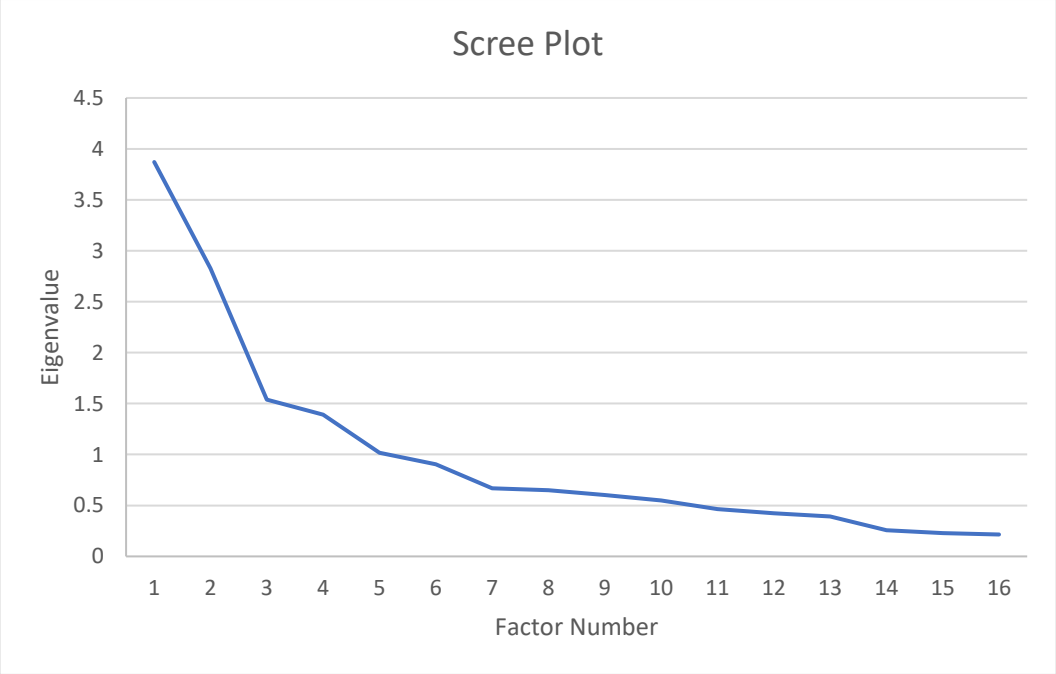
7.5 Factor Analysis: Identifying Latent Variables

In this section, we address research question two by examining our provenance metric data to uncover latent variables, i.e., insights which can be derived from provenance measurements but cannot themselves be directly measured or easily observed. We do this using exploratory factor analysis as described in Chapter 3. Please see that chapter for details of the procedure and assumptions testing.

7.5.1 Assumptions tests

The initial factor analysis produced an R-matrix determinant of 3.785×1012 which is less than 0.00001 indicating that multicollinearity was present. Several regression procedures were run to generate variance inflation factor (VIF) values. Average entity clustering coefficient, average clustering coefficient, average editors per feature, agents, entities, activities, and average activity degree had VIF values greater than 10 and were removed from the analysis. This reduced the R-matrix determinant to an acceptable 0.00002574.

Figure 49: Scree Plot



Communality values for life-cycle edits and entity power law coefficient were below 0.3, and these variables had no high factor loadings and so were also excluded. This factor procedure produced a KMO score of 0.80 and a significant result for Bartlett's test of sphericity.

Examination of the scree plot in Figure 49 shows an inflection point at the fifth factor, suggesting a 4-factor solution. Factor 5 of the solution has an eigenvalue of 1 which also agrees with Kaiser's criterion.

Table 16: Factor Analysis Results for All Data

	Factor			
	1	**2**	**3**	**4**
transient ratio	0.759			
quattrone maturity	0.722			
avg. agent degree	0.683			
density	-0.596			
avg. clustering activities	-0.548			
avg. clustering agents	-0.423			
avg. days since last edit		-0.730		
maintenance ratio		0.689		
avg. entity degree		-0.633		
new edits		0.540		
agent power law			-0.553	
activity power law			-0.553	
data age			0.444	
entity-entity MFD			0.439	
assortativity				0.780
avg. rich club coefficient				-0.551

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization
Rotation Converged In 6 Iterations.

Running the factor analysis with the number of factors extracted set to 4 produces a model which explains 46.9% variance. The resulting factors all had absolute correlations below 0.3, so an orthogonal rotation was selected for a solution with uncorrelated factors. There were still cross loadings, i.e. variables loading strongly on more than one factor. When loadings are similar or there is no strong loading on any factor, variables were eliminated, and the factor analysis procedure was repeated until Thurstone's simple structure was achieved. Transitivity, interactivity, number of

editors per cell, average edits per feature were removed, resulting in the factor matrix shown in Table 16. This solution has a simple structure suggesting four latent variables. The solution has a slightly smaller, but still acceptable KMO score of 0.752, a significant Bartlett's test of sphericity. It accounts for 46.276% of the variance in our data. The correlation matrix determinant is now 0.002 indicating no issues with multicollinearity.

7.5.2 The Factors

Examination of the variable loadings provides insights into the natures of these latent factors. The presence of some of the concrete maturity metrics aids this interpretation as do findings from Chapter 5.

Factor 1. Factor 1 has a strong positive loading for transient edit ratio, i.e. the proportion of edits reverted within one month. The strong positive loading for Quattrone maturity indicates that high levels of this factor occur in areas with a high proportion of mapped features to population indicating well mapped urban areas. Higher values of average agent degree can occur where there is intensive editing by a small number of users as we saw in Chapter 5. The negative loadings for average clustering coefficients for activities and agents can indicate editing in multiple changesets. These clustering coefficient values can also be reduced by the presence of several software agents. The negative loading of density indicates that higher values for this factor are associated with graphs containing fewer edges in relation to the number of nodes. Software agents will reduce the density, especially if editing is spread out over a longer period because this will result in separate software agent vertices for each software version, and each is more likely only to be connected to a single changeset. Density is also moderately positively correlated with the number of editors per cell and interactivity. To summarise, this factor is likely to have high values where there is a combination of high levels of editing intensity by more than one contributor, distributed over longer periods in well mapped areas i.e. a history of sustained collaborative editing.

Factor 2. Factor 2 appears to have a more explicitly temporal dimension. The negative loading of average days since last edit, and positive loading of new edit counts indicates that this factor is likely to have higher values in more recently edited content. Maintenance ratio has lower values with larger numbers of maintenance edits, so the positive loading means that high values of this factor increase the likelihood of data being at its first version. We also noted in Chapter 5 that average entity degrees can be particularly low in larger graphs with recent editing, which would be consistent with the negative loading on this factor. To summarise, high values of this factor are suggestive of recent or ongoing intensive editing.

Factor 3. Factor 3 also has a time related metric loading. The positive loading of data age indicates this factor is likely to have high values in areas edited very early in OpenStreetMap history, and which contain data edited a long time ago. This provides no indication of currency because it relates to the date of data in the cell was first edited rather than the most recent edit. Positive loadings of entity-to-entity maximum finite distance (MFD), a variable that usually measures the length of version chains, suggests a longer history of active editing. We have noted in Chapter 5 that low values of agent power law exponent can be caused by gradual editing over a longer period with larger numbers of contributors making significant edits in larger graphs.

Factor 4. Factor 4 accounts for the smallest amount of variance and has only 2 variables loading on it both of which are related to assorted mixing in Chapter 5 we noted that low rich club coefficients can occur in areas with a variety of richly tagged features which change rapidly over time, typical of busy high streets and city centre areas. Assortativity measures a preference for vertices to connect to other vertices of similar degree unlike rich club coefficient which measures the preference for vertices of higher degree, so the negative loading here is unsurprising.

7.5.3 Summary

The 4 factors we have identified show some themes which are consistent with those which have emerged from our examination of the network graphs in Chapter 5. We can detect sustained collaborative editing, volatile data, edit intensities and contribution inequality as drivers for some of these themes as well as for the variables in Chapter 5. We also see the differences in recently and intensively mapped areas and content created over a longer period by more than one user that we did in the zones we identified in Chapter 6 . All of these emergent themes are affected by characteristics of the physical environment, i.e. the things being mapped. For example, in the factors we have identified, the strong loadings for clustering coefficients and entity degrees can be affected by the proportion of osm:Ways in the data. We also know both from the literature [46], [90], and from Chapter 5 that some OSM contributors have a tendency to focus on single features which are often building footprints. In view of this it is likely that the factors we have uncovered may be specific to land coverage types and features. E.g. the factors characteristics may vary between urban and rural areas. We investigate this further in the following section.

7.6 Factor Analysis in Urban Areas

An advantage of using output areas as a geometry for capturing provenance data is that we can use output area classifications as an approximate indicator of environment types. Output areas 1 and 6 correspond to the OAC groups “rural residents” and “suburbanites” respectively. In our study

area, many of the output areas in the Suburbanites supergroup appear rural when viewed in satellite imagery but are not classified as such because of the existence of a small settlement, or part of a larger one, within the area containing a population which is demographically suburbanite. The other supergroups are almost all entirely urban in terms of their built environment, i.e. dominated by building footprints and separated by other demographic characteristics [237]. These groups, consisting of all output areas in OAC supergroups 2, 3, 4, 5, 7 and 8 provide a sample size of 935 which is adequate for factor analysis. To confirm that there is a significant difference between these OAC based groupings we conducted a t-test which confirms that the building count is significantly higher in our "urban" subgroup than in the combined rural/suburbanite grouping: $t(1176)= -9.9954$, $P <.001$ and thus this subgroup has a higher proportion of building footprints than the whole study area. The first run of the factor analysis procedure produced results with similar multicollinearity issues and using the same VIF method we removed average clustering coefficient, entities, agents, average activity degree, editors per cell and, interactivity, which brought the matrix determinant down to an acceptable value of 0.0000156. The KMO score was an acceptable 0.775 and Bartlett's test of sphericity was significant.

Examination of the scree plot (not shown) suggest either a 4-factor or a 7-factor solution. Each extra factor in the 7-factor solution only explains a small amount of variance, so to avoid introducing unnecessary complexity to the model, a 4-factor solution was chosen. This solution had no factors with a correlation coefficient above 0.3 and so we reran this procedure using an orthogonal rotation. We removed variables with very low communality: life cycle edits, reverse count, entity power law exponent and power law exponent. The rotated factor matrix revealed cross loadings for data age, average creators per feature, average edits per feature and a very weak loading for transitivity so these were also removed. This resulted in the matrix shown in Table 17 which has a simple structure. This solution had an acceptable KMO score of 0.778. Bartlett's test of sphericity was significant, and the matrix determinant was 0.001 indicating no problems with multi co-linearity.

Table 17: Factor Analysis Results for Urban Data

	Factor			
	1	**2**	**3**	**4**
Activities	0.842			
Transient Ratio	0.729			
Quattrone Maturity	0.722			
Density	-0.625			
Entity-Entity MFD	0.606			
Agent Power Law Exp	-0.420		0.414	
Avg Entity Degree		0.714		
Avg Days Since Last Edit		0.692		
Maintenance Ratio		-0.661		
Avg Clustering Entities		0.652		
New Edits		-0.515		
Avg. Agent Degree			0.745	
Avg. Clustering Activities			-0.580	
Avg. Clustering Agents			-0.460	
Activity Power Law Exp			0.421	
Assortativity				0.741
Avg. Rich Club Coefficient				-0.544

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization

Rotation Converged In 6 Iterations.

Factor 1. Factor 1 has positive loadings for activity count, Quattrone maturity and transient edit ratio, which suggest that this factor measures edit intensity and map completeness. The high entity MFD is usually caused by long edit chains which can result from heavily edited objects. This seems to be a similar picture to factor 1 for the whole dataset. The negative loading for agent power law exponent is not present in the original factor 1. In Chapter 5, we saw that low values for this variable can be caused by dominant users mapping in an area in detail focusing on a single feature type with other users filling in additional features. In an area with more building footprints, it is not surprising that agent power law exponent has a strong loading here.

Factor 2. This factor is very different from the factor 2 for the whole dataset, including the more rural areas. It still has a temporal dimension with loadings for average days since last edit, and new edit count, but these are reversed from the original, as is the loading for maintenance ratio. In contrast

to the original factor 2, high values of this factor are more likely to occur where data has not been edited for some time and is not at its first version. Entity clustering coefficient which has a strong positive loading, can often be high where data has been edited by more than one user. The high clustering coefficient values associated with this factor can be a result of a high ratio of osm:Ways to osm:nodes and the presence of several versions of osm:Ways. The average entity degree also has a positive loading and is also associated with a large number of osm:Ways. The combination of these two loadings suggests that these osm:Ways are mainly building footprints, which usually have fewer osm:nodes than other features. We have also observed high entity degrees associated with major roads and other linear structures which are heavily edited. High values of this factor suggest central urban areas with high building footprint counts and major roads, which have been edited to completion some time ago.

Factor 3. This factor differs from the original factor 3 in that it has no explicitly temporal component. Negative loadings for average activity and agent clustering coefficients indicate that higher values of this factor are likely to occur in areas with editing over time by multiple contributors on the same features in multiple changesets. The positive loading of average agent degree combined with activity power law exponent is consistent with a small number of contributors focusing dominant feature footprints with other scarcer features added by other contributors. The high agent degree would indicate the dominant feature is numerous and likely to be building footprints, and that the bulk of the editing is done by a small number of contributors.

Factor 4. Factor 4 is virtually identical to the original factor.

7.6.1 Summary

Measurement of provenance data using the metrics we have described clearly has the potential to render it amenable to forms of analysis which can uncover insights into the nature of VGI data creation, many of which are not otherwise directly or easily observable. Because of the existence of multicollinearity and varying degrees of communality among our variables we have had to remove a lot of potentially meaningful information. The variables removed were not all strongly correlated but their linear combinations were confounding the analysis. It is likely this is caused by error variance related to external non-provenance related factors, most probably linked to the type of environment and features being mapped. This has meant that using all the variables we have available has not been possible, and it is likely that other latent factors remain undiscovered in the OpenStreetMap provenance data.

However, it has been possible to carry out a robust exploratory factor analysis which has uncovered 4 latent variables. These factors seem to vary with edit intensity over time, levels of collaboration and interaction between contributors, whether editing is recent or ongoing, length of time since mapping began, and rates of change reflecting real world volatility. Performing the exploratory factor analysis on an area with a higher proportion of building footprints yielded different results (see Table 18). It is likely that this would also be the case for road and street networks.

The factor analysis for the output areas with a higher proportion of building footprints (supergroups 2,3,4,5,7 and 8) has some similarities with the original factor analysis, but the variable loadings seem to emphasise aspects of contributing behaviour which are associated with mapping building footprints. For example, in Table 18, the original factor 1 can be interpreted as a measurement of sustained collaborative editing and high completeness. With more building footprints, the second factor 1 also seems to measure sustained editing and edit intensity, but by the dominant user focusing on a single feature type. This is a pattern we have observed in users mapping built-up areas. Factor 2 is completely changed for the area with increased building footprints. Rather than being a simple measure of recent or ongoing intensive editing, it now seems to be measuring building footprint counts and data which is edited to completeness some time ago. The second version of factor 3 remains a measure of collaborative editing but is now measuring the extent of a pattern in which contributors focus on the dominant feature type with other features added by many contributors, another pattern associated with the mapping of built-up areas and residential neighbourhoods.

This confirms previous findings that physical environment characteristics, such as the occurrence of a dominant feature type, leave signals in the structure of a provenance graph. These are a product of capture policies which reflect the geometric structure of the primitives which represent those features. They also reflect the way in which contributors interact with the features they map, often focusing on a single feature type.

Table 18: Factor Characteristics

	Whole Dataset	Urban Areas (Supergroups 2,3,4,5,7, and 8)
Factor 1	Sustained collaborative editing, high completeness, many contributors	Heavily edited objects, map completeness, length of editing history, dominant user focusing on single feature type
Factor 2	Recent, or ongoing intensive editing	Central urban areas, high building footprint counts and major roads which have been edited to completion some time ago
Factor 3	Data first edited a long time ago, gradual editing over a long period with multiple contributors	Collaborative editing over time, most editing done by small number of contributors focusing on dominant feature type with other features added by many contributors
Factor 4	Volatile features	Volatile features

These results also begin to address research question 2. Understanding how the characteristics of features and OSM primitives leave signatures in provenance graphs can inform capture policies which could be tailored to derive specific insights. This might be achieved either by focusing on specific features, or by altering the way in which provenance is recorded for certain primitives. For instance, one might omit to record or aggregate the provenance of the member osm:nodes of osm:Ways.

7.7 Analysing and Comparing Variance

In previous chapters we have identified some of the insights that can be gained from the study of provenance graphs (research question two). We have identified some of the sources of variation in provenance graph measurements. The effects of individual characteristics of OpenStreetMap contributors are apparent, as are the physical properties of the environment and the geometry of its representation in OpenStreetMap. To investigate what effect the demographic characteristics of the coverage area might have on the provenance graphs we carry out our MANOVA procedure as detailed in Chapter 3, Section 3.5.4.

7.7.1 The MANOVA Procedure Results

There are a number of assumptions MANOVA makes about the data which are described and addressed in Chapter 3, Section 3.5.4. The following sections deal with the testing of these assumptions.

Group Sample size. In our study area, the sample size of each OAC supergroup differs. In MANOVA, the sample size should exceed the number of variables in the study [249], and should ideally be more than 30. We therefore exclude group 3 from this investigation. This group also has several outliers which means several of the cases would need to be altered.

Table 19: Group Sample Sizes

OAC	Sample size	%
1: Rural residents	31	2.63
2: cosmopolitans	141	11.97
3: ethnicity central	23	1.95
4: multicultural metropolitans	125	10.61
5: urbanites	356	30.22
6: suburbanites	212	18.00
7: constrained city dwellers	138	11.71
8: hard pressed living	152	12.90

Univariate outliers. The variables in the analysis were tested using box plot in for inspection to identify outliers further than three times the interquartile range from the mean. Most of the variables had an unacceptable number of outliers and so the variables were transformed to reduce their impact. Square root, log 10 or reflection techniques were applied to the variables. Visual inspection of histograms was used to select transformations that produce the most normal-looking distributions. The transformed variables were then reassessed using box plot inspection as described

above. Transformed variables with more than 1% of their values as outliers were eliminated. Consequently, we removed entities, quattrone maturity, transient edit ratio, agent power law exponent, life-cycle edits, assortativity, Data age, average days since last edit, revert count, transient edits, power law exponent, and density.

12 output areas still had extreme values for some of the variables and these were manually examined to investigate why these values were so extreme. Most seem to be caused by peculiarities of the output area zoning algorithm and the way we use its geometry for provenance capture. Many resulted from output area boundaries clipping features such as main roads and railways which tend to be very heavily edited. We also found a case where the output area enclosed a partial footprint of a large residential complex, resulting in the capture of an incomplete feature, which severely skewed the degree distributions.

The demographic properties of the area are still likely to have a role. One might argue that proximity to railways, major roads, and large residential buildings likely to be intersected by output area geometry are demographically related factors. However, these circumstances affect many output areas not flagged as outliers. It is therefore reasonable to suggest that these extreme values are not representative of the data. Their presence, however, is likely to have a disproportionate impact on any modelling, so we feel justified in removing them. Many output areas have extreme values for several variables and so are also likely to be multivariate outliers.

Normality. After outlier removal, the remaining variables were assessed for normality by visual inspection of histograms and QQ plots. The variables were found to be approximately normal in the QQ plots although there were some minor deviations from normality apparent in the histograms. As MANOVA procedures have some degree of robustness to minor normality variations [249], [264]–[266], they were considered acceptable.

Homogeneity of Variance/Covariance. Homogeneity of variance-covariance matrices was assessed using Box's test of equality of covariance matrices. In both cases the significance level (P-value) was less than .001, so we accept the null hypothesis and have violated the homogeneity of covariance assumption. This adds a caveat to the results that the MANOVA tests are less powerful than they might otherwise be.

Linearity. Linear relationships between dependent variables within groups were assessed using Spearman's correlation coefficients. Although all the variables have a linear relationship with at least one other variable, about half had correlation coefficients below 0.2, violating this assumption. This adds the caveat that the power of these MANOVA analysis may be reduced when using some test

statistics. We therefore rely on Pillai's Trace which is generally regarded as robust to this assumption violation [249] (see Chapter 3, Section 3.5.4).

Homogeneity of Variances. We assessed homogeneity of variances using Levene's test and found that for most variables this assumption was violated. This issue does not preclude the use of MANOVA but requires a stricter criterion for statistical significance [296], and so we require a significance level of 0.01 rather than 0.05 to reject the null hypothesis.

Multicollinearity. Multicollinearity was heuristically assessed using Pearson's correlation coefficients. All variables had moderate correlations with several other variables. Average clustering coefficient for entities was removed because of a strong correlation with overall average clustering coefficient. Average editors per cell was removed because of a strong correlation with average agent degree. Interactivity was removed because of a strong correlation with maintenance ratio.

7.7.2 Results

MANOVA Test Statistics. SPSS provides results for Wilk's Lambda, Pillai's trace, Hotelling's trace and Roy's largest root tests, which are shown in Table 20 . Wilks' Lambda is the most popular test statistic, but Pillai's Trace is more robust to assumption violations [265], [267], [268] and is our chosen statistic. All have *P* values < .001, Removal of multivariate outliers had little effect on these results.

Table 20: MANOVA Test Results

		Value	F	Hypothesis df	Error df	Sig.	partial η^2
MV outliers	Pillai's T	0.712	7.84	108	6288	p < 0.0001	0.119
	Wilks' λ	0.455	8.161	108	5984.674	p < 0.0001	0.123
	Hotelling's T	0.877	8.452	108	6248	p < 0.0001	0.127
	Roy's LR	0.363	21.131	18	1048	p < 0.0001	0.266
no MV outliers	Pillai's T	0.685	8.047	108	6744	p < 0.001	0.114
	Wilks' λ	0.47	8.366	108	6420.234	p < 0.001	0.118
	Hotelling's T	0.837	8.654	108	6704	p < 0.001	0.122
	Roy's LR	0.344	21.461	18	1124	p < 0.001	0.256

The results of this MANOVA show that there was a statistically significant difference between output area supergroups based on the combined graph metrics considered in this study: Pillai's T = 0.712, F (108, 6288) = 7.84, P < .001, partial η^2 = 0.119. These results provide sufficient evidence to reject the null hypothesis and conclude that provenance graphs taken from OpenStreetMap provenance graphs differ based on a selection of concrete and abstract provenance graph metrics. The effect size was medium as assessed against Cohen's criteria [269].

7.7.3 Discriminant Function Analysis

To provide further insight into the nature of these results we performed a discriminant function analysis to assess the extent to which the linear combinations from the MANOVA can derive functions which predict group membership. The procedure is essentially a reversal of MANOVA. In MANOVA we discover whether a classification is associated with significant differences in linearly combined dependent variables.

Table 21: Discriminant Function Tests

test function	Wilks' λ	Chi^2	df	Sig.
1 through 6	0.455	829.529	108	< 0.001
2 through 6	0.620	503.321	85	< 0.001
3 through 6	0.768	278.109	64	< 0.001
4 through 6	0.856	163.527	45	< 0.001
5 through 6	0.921	86.678	28	< 0.001
6	0.986	14.996	13	0.308

If the MANOVA procedure finds significant differences these should be able to predict group membership [249]. In discriminant function analysis, a set of functions are derived which use coefficients from the within group covariance matrix to produce a score which is used to classify a data point. The purpose of this analysis is not to achieve efficient classification. Instead we can assess the functions in much the same way a factor analysis can be used to understand underlying structures by looking at the contributions of individual variables. This post hoc use of discriminant function analysis serves as a method of interpreting MANOVA results to gain insights into the factors which drive group differences [248], [249].

The number of functions derived is either the number of groups or the degrees of freedom provided by the grouping variable whichever is least [249]. In our analysis, the procedure has provided 6 discriminant functions. Five of these were found to be statistically significant (Table 21).and most of the variance was explained by the first three. The first explained 41.4% of the variance, canonical $R^2 = 0.516$. The second explained 27.2% of the variance, canonical $R^2 = 0.439$, and the third 13.1% of the variance, canonical $R^2 = 0.321$.

Table 22: Canonical Correlations

Function	Eigenvalue	% Variance	Cumulative %	Canonical R^2
1	.363	41.4	41.4	0.516
2	.238	27.2	68.6	0.439
3	.115	13.1	81.7	0.321
4	.076	8.6	90.3	0.265
5	.070	8.0	98.4	0.256
6	.014	1.6	100.0	0.119

Classification Results. To assess the validity of the functions used for interpretation of the MANOVA results, function 1 is used to predict supergroup membership for each output area provenance graph. These predictions are displayed in Table 23. We validate this model by assessing the prediction accuracy. For any insights to be useful, this should be greater than the probability that a randomly chosen data point is a member of a given group. Because the group sizes are uneven, SPSS calculates the prior probability of group membership, which is shown in Table 24. For example, in Table 23, we see that 60% of the data points identified as belonging to supergroup 1 were correctly identified as such. In Table 24, we can see that the prior probability of a data point belonging to this group is 2.81%. This means that if we used a function that assigned all data points to supergroup 1, it would be correct 2.81% of the time. The function is actually correct 60% of the time, which is a substantial improvement.

Table 23: Classification Results

		Predicted Group Membership							Total
supergroup		**1**	**2**	**4**	**5**	**6**	**7**	**8**	
Count	**1**	**18**	0	0	7	4	0	1	30
	2	6	**71**	9	32	11	2	2	133
	4	2	18	**38**	33	13	7	6	117
	5	8	18	22	**197**	39	23	18	325
	6	13	14	10	74	**66**	3	15	195
	7	0	6	2	50	12	**49**	9	128
	8	2	6	4	60	16	17	**34**	139
%	**1**	**60.0**	0.0	0.0	23.3	13.3	0.0	3.3	100.0
	2	4.5	**53.4**	6.8	24.1	8.3	1.5	1.5	100.0
	4	1.7	15.4	**32.5**	28.2	11.1	6.0	5.1	100.0
	5	2.5	5.5	6.8	**60.6**	12.0	7.1	5.5	100.0
	6	6.7	7.2	5.1	37.9	**33.8**	1.5	7.7	100.0
	7	0.0	4.7	1.6	39.1	9.4	**38.3**	7.0	100.0
	8	1.4	4.3	2.9	43.2	11.5	12.2	**24.5**	100.0

There is also an alternative reading of this matrix. As well as considering accuracy as above, by reading the rows of the matrix, we can also consider the columns to identify errors of commission. For example, 49 graphs were classified as being in supergroup 1, and 18 of these classifications were correct, a rate of only 36%. For a classification task this would be unacceptable. However, this function is still performing better than a classifier that randomly assigned groups, which we would expect to be correct 12.5% of the time. Accuracy and errors of omission rate both show an improvement over random assignment for all of the OAC supergroups, which indicates that the function structure contains information about which variables are playing a role in distinguishing graphs from different output area supergroups.

Table 24: Prior Probabilities for Groups

supergroup	prior (%)	cases
1	2.81	30
2	12.46	133
4	10.97	117
5	30.46	325
6	18.28	195
7	12.00	128
8	13.03	139
Total	100.000	1067

Discriminant Function Structure Analysis. The structure matrix (Figure 50) provides an indication of the contribution each variable makes to each of the discriminant functions. Interpretation is less straightforward than exploratory factor analysis, where we ascribe high or low contributing variable values to a factor to understand it. Here, absolute loading values identify differences between groups. Signed loading values also show us relationships between variables which also distinguish the groups. For example, the greatest distinction we see is between the rural and urbanite supergroups. The main drivers are agent power law exponent, which loads negatively, and average days since last edit, which loads positively.

This represents a distinction between recently edited data which has high agent power law exponents, and less recently edited data with low agent power law exponents. It suggests two contrasting patterns: one, where much of the of the editing has been done by a single user and editing is ongoing, and the other with less recently edited data which was gradually built up by several contributors. This is also supported by the loadings for the agents and activities variables. Both patterns have been seen in Chapter 5 and Chapter 6, where we found they were associated with building footprints and the way contributors interact with them.

Table 25: Structure Matrix

	Function					
	1	2	3	4	5	6
Avg. Days Since Last Edit	**.500***	-0.006	-0.166	-0.259	0.124	0.334
Agent Power Law Exponent	**-.364***	0.094	-0.033	-0.033	0.008	-0.051
Transitivity	-0.184	**.617***	0.170	-0.199	0.265	0.041
Avg. Clustering Entities	0.212	0.044	**-.375***	-0.188	0.164	-0.205
Avg. Creators/Feature	**0.355**	0.010	**-.373***	-0.201	0.029	0.118
Maintenance Ratio	-0.012	-0.022	**.356***	0.172	0.059	-0.020
Avg. Cluster Coefficient	0.192	0.016	**-.318***	-0.225	0.246	-0.193
New Edits	-0.063	0.012	-0.062	**.756***	0.181	-0.244
Activities	0.244	-0.163	-0.056	**.507***	-0.493	0.023
Agents	**0.365**	**-0.285**	-0.060	**.445***	-0.367	0.016
Activity Power Law Exponent	-0.215	0.004	-0.006	-.290*	0.074	-0.147
Entity-Entity MFD	0.061	-0.055	0.030	.275*	-0.146	0.123
Avg. Entity Degree	0.085	-0.040	0.026	.248*	-0.147	0.159
Avg. Clustering Agents	-0.052	-0.104	**0.369**	-0.116	.444*	-0.262
Avg. Agent Degree	-0.050	-0.138	**0.321**	0.068	-.323*	-0.200
Avg. Clustering Activities	0.061	-0.047	0.056	-0.070	.309*	0.098
Avg. Edits/Feature	-0.146	**-0.305**	-0.191	-0.089	0.157	.486*
Avg. Rich Club Coefficient	0.054	0.094	0.056	-0.219	0.222	-.375*

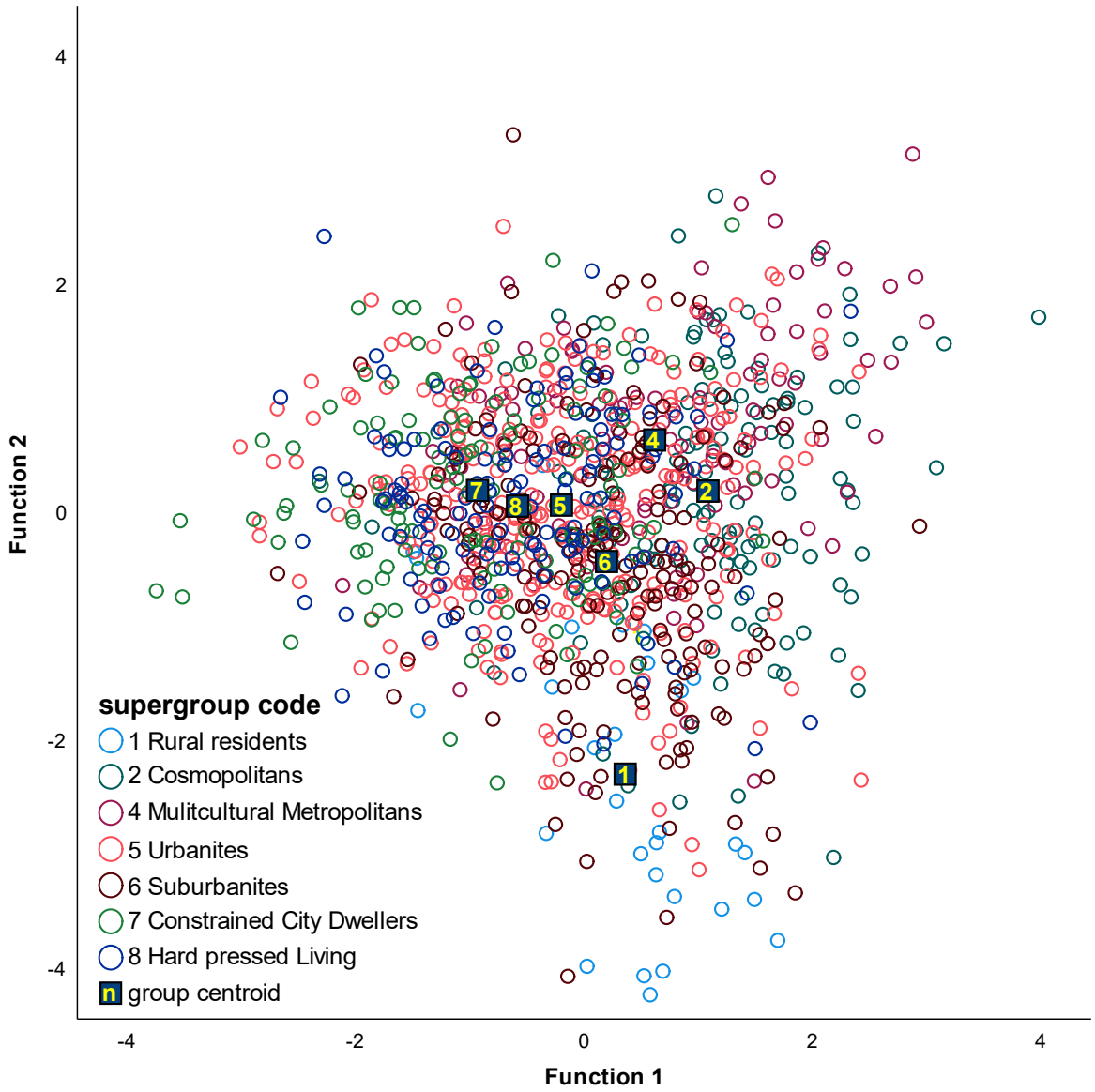
Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.
* Largest absolute correlation between each variable and any discriminant function

Examining function 2, we can see that transitivity is also a distinguishing factor with a high positive loading. Transitivity is a relationship between the number of actual triangles within a graph and triples (potential triangles). It is also related to the ratio of osm:nodes to osm:Ways and has a positive correlation with building count which is stronger in the Rural and Suburbanites supergroups. Transitivity values are higher in the Rural Residents supergroup than in the Urbanites supergroup as confirmed by a t-test: $t(36.3)= 9.312418$, $P <.001$. This variable seems to be strongly influenced by the number and type of features present, and particularly the number of building footprints. Osm:Ways and their versions which share nodes will increase the number of triangles in the graph and transitivity is positively correlated with interactivity, the average feature version number.

These loadings indicate that aspects of the physical and built environment and type of features present in the OSM coverage are substantially responsible for variation in the structure of provenance graphs between census output areas. Building footprints appear to be a major driver for this which is unsurprising given that they are OpenStreetMap's most numerous feature type, accounting for approximately 59% of all osm:Ways at the time of writing. These effects are partly due to the geometric structure of the features and the primitives used to represent them. However this does not entirely explain some of the variation we see. OpenStreetMap contributors mapping practices also vary depending on the type of features they are mapping. For example, in Chapter 5 and Chapter 6 we observed that building footprints are often mapped intensively and to completion by a single user in a very focused way. We also saw how this can leave a distinctive signature in the provenance graph. Variation in built environment features between different OAC supergroups are documented in the ONS pen portraits document [237] and summarised in Chapter 6. These variations are reflected in the discriminant functions identified in this analysis and seem to be driven by both the type of feature being mapped, and its effect on the behaviour of the OpenStreetMap contributor.

Function 1 and 2 Scatterplot. Figure 50 shows a scatterplot of the discriminant function 1 and 2 values . It clearly shows the separation of supergroup one (rural residents), and to a lesser extent supergroup 6. These are both groups with fewer building footprints, containing more land cover characteristic of rural areas. Supergroups 2 and 4, Cosmopolitans and Multicultural Metropolitans respectively tend to be located in more central areas and closer to Southampton's universities, and group 7, 8 and 5, Constrained City Dwellers, Hard-Pressed Living and Urbanites respectively, tend to be in more peripheral urban areas with higher deprivation indices.

Figure 50: Canonical Discriminant Functions



7.8 Conclusion

In this section we have performed a MANOVA which shows that the characteristics of provenance graphs differs between the output area classification supergroups of the output area they are recorded in. We have also used discriminant function analysis to show that these differences can be used to predict the output area supergroup more effectively than random chance. The greatest distinction is between the urban and rural supergroups, and this is likely to be due to the physical environment which affects the structure of provenance graphs due to the geometric structure of the features and how they are reflected by the provenance capture strategy. The type of features being mapped also affect provenance graphs because OSM contributors seem to have different behaviours depending on the features they map. Building footprints are OSM's

most numerous feature and have a profound effect on provenance graph measurements. This is partly due to their geometric characteristics but also the specific editing behaviour associated with the way contributors interact with them. Devising provenance capture policies which suppress the geometric aspects will reveal more about specific contributor behaviour.

Chapter 8 Conclusions

This study of OpenStreetMap was motivated by an interest in provenance network analytics [54], [56]. We have seen a great deal of research which shows that the scientific community are aware of the value of a principled, automated analysis of OpenStreetMap contribution practices. Several studies have also attempted to use provenance type information for OpenStreetMap analysis [58], [58], [61]. All this research has used information extracted from provenance data to make predictions about the *nature of the map*. This usually involves producing some estimate of, or proxy for data quality, which are undoubtedly useful insights. However, they do not provide the deep understanding of the *nature of OpenStreetMap contribution patterns* and what drives their heterogeneity. This was called for by Goodchild and others [33], [94], [110], [258] as an integral component of any quality assurance framework for OpenStreetMap and other VGI.

Whilst there has been some valuable qualitative research aimed at understanding OpenStreetMap contribution practices, these do not operate at the scale of analytics. Provenance data encodes many aspects of the OpenStreetMap creation process and contains a record which reflects much of the variability of its contribution patterns. The current research into OpenStreetMap provenance data analysis only provides a framework for limited predictions about the state of the map. This thesis explores the potential of provenances data to provide a deeper understanding of OpenStreetMap contribution practices and patterns. Using a detailed descriptive analysis [177] of provenance data, we provide a theoretical framework encompassing provenance capture, metrics, and measurement strategies, and what they can reveal about OpenStreetMap contribution practices.

8.1 Research Questions

8.1.1 Research Question One: How Can Different Approaches to the Measurement of a Provenance Graph Produce Useful Insights Into the Nature of VGI/UGC/OpenStreetMap?

We have identified three approaches to the measurement of provenance graphs, described in Chapter 3, Section 3.3 and identified the practical implications of these measurement types, by implementing metrics using all three approaches.

Concrete Metrics. Concrete graph metrics require domain knowledge, and an understanding of some tangible parameter of the phenomenon the graph data is describing. Maturity is our implementation of a concrete metric. It is based on concepts derived from research into open-source content creation, primarily Wikipedia and OpenStreetMap. These were used as a basis to

develop a notion of maturity as a life-cycle stage for VGI map data. It reflects the extent to which data has achieved a complete, stable state, and editing is only required to reflect real world change.

We found that many of the characteristics of maturity in OpenStreetMap are common to other user generated content. Maturity measurements do correlate with summary measures of quality derived from automated error detection and comparison with satellite imagery. However, although correlations exist, assumptions should not be made about maturity metrics based on insights from other domains. For example, Linus's law is an assertion that there is a positive relationship between data quality and the number of humans who have interacted with the data [18]. It has been identified in open source software [18], OpenStreetMap [44] and Wikipedia [297]. In our experiment we found that it does not apply as one might expect. Instead, we found that high-quality data is often produced by a single user editing to a high standard and level of completeness, such that no further editing is required. The lesson here is that although Linus's law correctly identified a relationship, domain specific investigation using inspection of map data and visualisations of RDF network graphs were necessary to understand its nature and direction. The implications for provenance metric design are that although investigations of related domain knowledge and research are useful tools, descriptive data analysis is also required.

Abstract and Semi-Abstract Metrics. Abstract metrics have the advantage of requiring no domain knowledge. An abstract graph metric can just as easily be used to quantify electrical brain activity as it can provenance data. Yet this versatility comes at a price; a lack of accompanying domain knowledge makes these measurements more difficult to interpret. Using graphs encoded with the RDF framework helps to overcome this. Each edge and vertex in an RDF graph is represented by a URI which can be resolved to an attached data point such as an OpenStreetMap feature or a PROV-DM relationship. The graphs can be visualised using the Cytoscape software [232], which allows detailed visual inspection of the network structures and provides access to these URIs. This allows vertices in a provenance graph to be inspected via the OpenStreetMap API and provides access to information about PROV-DM types.

Many of our investigations have shown that signals from variations in the physical environment are a source of variance in provenance metric data. This is discussed in further detail below, but one aspect of this is the role of the different geometric attributes of feature representations in OpenStreetMap. Part of this structure is transferred as a signal into the provenance graph because provenance reflects the actions required to create the OSM data primitives necessary to provide a geometric representation of that feature on the map. This effect is more pronounced in abstract measurements.

General Insights. Provenance capture strategies that suppress the capture of geometric aspects of OSM data primitives could produce different results and insights. This is not to say that these characteristics are always a negative effect. Just that capture policies can be tailored to investigate different aspects of OSM contribution and data. The choice of measurement approach has similar implications, and both can be seen as useful parameters for tailored investigation. Abstract and concrete provenance measurement have different inherent sensitivities and can work well in tandem. The specific details of their implementation are important parameters which can be used to target specific analysis goals. The use of abstract metrics requires careful design of a capture policy; for concrete metrics on the other hand, the emphasis is on metric design.

8.1.2 Research Question Two: What Insights Can Be Demonstrated About User Editing Behaviour and the Mapped Environment Using Provenance From VGI/UGC/OpenStreetMap?

The investigations in this thesis have uncovered number of themes for our understanding of OpenStreetMap contribution practices and their interaction with the mapped environment. The dataset produced for this thesis is a complex one. The provenance capture was a “scattergun” approach designed to produce a graph encoding as much provenance information as possible. Although osm:Relation primitives were omitted, all other features were captured without discrimination. We know from research that OSM contributors interact with features differently depending on the type, having individual preferences for specific features. The presence of those features also varies demographically. To further complicate matters OpenStreetMap contributors also react differently to areas with different demographic characteristics, preferentially editing areas whose population profile is similar to their own demographic. Other altruistically motivated contributors preferentially edit areas they see as being more deprived. These factors exemplify the sources of variance in OpenStreetMap editing practices as a set of complex, interacting phenomena.

Despite this, we have been able to identify several themes related to OpenStreetMap contribution. Some insights can be gained from individual variables. However, many of them cannot be directly observed and exist as latent variables that can be inferred from combined provenance metrics.

OpenStreetMap Contribution: Provenance Network Analytics Themes. The study of both individual and latent variables has resulted in our understanding of the following themes.

Contribution Profiles. These investigations have uncovered several patterns of contribution which are sources of variance in provenance metrics. The variance occurs in individual metrics, but also in

combinations of variables. It is these combinations which help to distinguish the effects as latent variables. Each pattern is composed of :

- variations in collaboration
    - the extent to which multiple users edit the same data
    - the extent to which data is created by a single dominant user
- temporal variations
    - whether most contributions occur steadily over time
    - whether most contributions occur during a short time period
- edit intensity
    - high rate of editing, high-volume over a short period.
- feature focus
    - whether contributors focus on a single feature type

Steady Maintenance Editing. This is a pattern of continuous contribution throughout the lifetime of the data. Several of our investigations have identified this theme. We found maintenance editing had a role in distinguishing urban and rural areas, and to a lesser extent central and peripheral urban areas in the discriminant function analysis in Chapter 7, Section 7.7.3. It also emerged as a latent variable in exploratory factor analysis. Prov:Agent counts, maintenance edit ratio, activity clustering coefficients, rich club coefficients, entity clustering coefficients and activity power law exponents are all affected by this pattern.

Local Editing. Local editing can take the form of substantial maintenance editing by a single user over a long period. Where little work is done by other users, we surmise that this is an experienced contributor editing their local area. This is borne out from inspection of network graphs and osm:Changeset boundaries. As well as other features of maintenance editing, this pattern is characterised by high clustering coefficient. low entity power law exponents result from multiple changesets because the work is carried out over longer time period.

Expert Editing. A pattern which has emerged in several investigations is that of a single contributor who comprehensively maps an area to a high standard over a short period before moving on to another area. The work is carried out to a high standard, such that there is little or no scope for further editing. This intense burst of contribution with little or no activity either side of it leaves a distinctive signature in provenance graphs which can be detected by our metrics. We have seen it in the exploratory factor analysis and the discriminant function analysis. Evidence of it has

also emerged from inspection of network graphs and thematic maps. The metrics involved are clustering coefficients and degree distributions.

Feature Preferences. Individual preferences among contributors for the editing of specific features has been established in other research. In our network graph inspections in Chapter 5 we saw contributors almost exclusively edit building footprints and the effect this had on entity degree distribution and rich club coefficients. This theme also emerged in factor analysis and in the MANOVA post-hoc discriminant function analysis. As building footprints are a dominant feature in OpenStreetMap, other features are not sufficiently numerous for us to be able to detect other effects in our study area.

Collaborative Editing. Levels of collaboration are an important driver of variance in individual provenance metrics, latent variables and in the discriminant, functions identified in Chapter 7, Section 7.7.3. Our primary discriminant function and the first factor in the exploratory factor analysis both relates to levels of prolonged collaborative editing. Metrics affected include clustering coefficients and agent degree distributions

Recent Editing. Factor analysis revealed a latent variable which related to recent or ongoing editing, and we found this to be associated with low average entity degree in our inspection of network graphs in Chapter 5.

Other User Behaviours. Examining provenance metrics on thematic maps shows some specifically spatial behaviours exhibited by contributors. We noted an apparent tendency for contributors to use features on the map to delineate the region they edited in during their session. This is visible both in the graphical representation of aggregated metric data and from inspecting the maps. A good example of this behaviour is the tendency to use a major road, railway, or other linear feature as a boundary. The contributor then maps within the boundary. This is why come of the spatial distribution patterns of provenance metrics seen in thematic maps have descriptive shapes.

Feature Dynamics. Building footprints are OpenStreetMap's most numerous feature. Provenance data for urban and suburban residential areas is completely dominated by these in areas with high completeness. There also preferentially edited by many OSM users and have some distinctive editing dynamics. Factor analysis can yield different results with even a modest increase in the proportion of building footprints. Many of the associated factor loadings show an increased emphasis on patterns associated with building footprint editing.

Other features also leave distinctive features. Street/road networks can have very high node counts and involve a lot of node reuse which results in high average entity degree degrees. Large residential buildings such as tower blocks can result in very small output areas because of the high

population concentration which has a distinctive effect on metrics such as activity power law coefficients.

Spatial Effects. The MAUP (Modifiable Aerial Unit Problem, see Chapter 3, Section 3.4.2) appears to have some effect. Both the measurement of the provenance metrics, and the capture of the provenance itself are forms of spatial aggregation. The use of output area boundaries as aerial units for data aggregation adds an interesting dimension to this. We are aggregating data which varies due to geodemographic factors. To do this we use polygons whose size and shape is a function of population density and housing type, which are also geodemographic factors.

In some cases, our metrics are affected by the geometry used to capture the provenance graphs. The boundary can clip neighbouring features such as major roads and motorways and railways which can skew the data. For example, a tower block will be in a small output area owing to the large population of the building. If the boundary of this output area clips a motorway this will have a disproportionate effect on the data. The effect might appear to be demographic, and in some respects it is. Proximity to such features is likely to bear some relationship to the demography of the area.

Wide area editing is another spatial effect that was identified and can be noticeable in areas with low completeness. This can be the result of data imports, or users who preferentially edit rare features over wide areas. It results in apparently very low degree changesets and prov:Agent counts and is a peculiarity of the way we capture provenance. Because this is done by output area, changesets which in reality have high degrees may only have low degrees when measured locally. This suggests the need a change to measurement algorithms so that degrees can be captured on a local and global basis. Both measurements are likely to be useful ways to study provenance. For example, the local degrees are useful for detecting low building completeness because in this situation the wide area edits affect degree distributions.

Environmental and Demographic Effects. We have investigated what role environmental and demographic factors have had in the variance of our provenance metrics. These factors are quite difficult to separate because of their interdependency. Physical factors such as variations in the built environment are products of and are driven by demographic variations. Both dictate the frequency occurrence of certain OpenStreetMap features and we have seen how the geometry of their representation in OpenStreetMap data can affect provenance graphs. This is evident in the correlations we find between provenance metrics and metrics of the physical environment, and to some extent in the differences we have identified between output area classification supergroups.

The proportion of building footprints in a provenance graph have a profound effect on the variance of provenance metrics and can change the results of factor analysis procedures. It also affects contribution practices. Moreover, the type of building also produces specific patterns. Tower blocks and residential complexes tends to be edited by numerous editors over longer time periods in stark contrast to houses in large residential neighbourhoods. Aggregation using output areas complicates this because the presence of tower blocks affects population density and can lead to output areas with only one building. Business premises and districts also seem to have a distinctive signature. They have a higher rate of change, particularly to tags, than residential premises. This pattern of steady change over long periods is noticeable in provenance metrics. It is a major contributor to one of the discriminant functions we identified in Chapter 7, Section 7.7.3 . Some environmental effects can be controlled by changing provenance capture strategies, e.g. by targeting specific feature types. Alternatively, these effects may be useful targets for analysis raising the possibility of provenance as a remote sensor. For example metric strategies targeting maintenance editing involving metrics such as average rich club coefficient might differentiate commercial from residential buildings.

Data Completeness. Provenance metrics seemed to be quite good at detecting data completeness issues in built-up areas. Built-up areas inevitably contain street networks, and these are often quite intricate and usually mapped before building footprints. In areas with lower building completeness they dominate the provenance graph and can cause extreme values particularly of degree distributions. This occurs because nodes are often shared between streets at intersections which results in some high degree nodes.

8.2 Reflections

The investigations in this thesis have highlighted several practical issues for the exploratory analysis of OpenStreetMap provenance graphs. One of the biggest difficulties was the lack of targeted provenance capture strategy. The initial database captured using the XLT process described in Chapter 4, Section 4.2, captured a complete provenance dataset from the edit history. The area-based graph extraction techniques used to generate the output area provenance graphs then captured as much provenance as possible into a local graph. As we have seen, the drivers for provenance graph variance are highly complex, often interacting seems which need a much more targeted strategy to provide the best insights. Fortunately, the provenance extraction pipeline allows for re-specification of the provenance graph model for further research.

The area-based extraction of provenance generated some large graphs and calculation of metrics did not scale well. One metric, graph diameter, was used in the original provenance network

analytics research [54], [56]. In our study this had to be abandoned because the computation did not scale well with larger provenance graphs. This would be less of an issue with graphs captured on a per feature basis. Feature based graph extraction has its own set of issues discussed in Chapter 3, Section 3.4.1. However, for provenance targeting single features, this may be less problematic. For the study of road networks, area extraction still seems to be a better option. Isolating a road feature in a local study of street networks is problematic because of their linear nature. The provenance for a major road can potentially capture data 50 miles away from the area of interest. Our approach captures a road feature within the area of interest but ignores the portion of the feature outside that area, by only capturing the osm:nodes inside it.

Another issue with area-based provenance capture was caused by aggregation geometry and by the eager provenance capture techniques, which resulted in some extreme measurements and skewed data. Some of the extreme measurements were characteristic of specific features. For example, administrative boundaries were particularly problematic because of the extreme length of their version histories and often complex tagging structures. To ensure the robustness of the MANOVA and factor analysis procedures it was necessary to drop several variables because of these extreme values. More targeted provenance capture will produce cleaner data, providing further and more detailed insights by allowing more variables to be considered.

Both area-based and feature-based provenance have issues with the calculation of changeset and agent degrees. We calculate degrees from the local graph, such that the changeset degree will only reflect edits made within that graph rather than over OpenStreetMap as a whole. There are certain advantages to this approach in that local degrees can reflect local behaviour. However, this fails to identify changesets edits or contributor activity outside of the local provenance graph. Future provenance capture strategies might make use of the OpenStreetMap API to provide these global degree measurements so that changeset and agent degrees can have a local and global dimension.

Although predictive analysis was and still is a motivation for this thesis, we do not offer any of these results or methods as a basis for documenting any aspect of OpenStreetMap contribution. They are a methodological framework for descriptive analysis from which predictive methodologies can be derived. We have identified and overcome numerous practical issues, but others remain. However the work we have presented in this thesis and the insights provided by its results provide the tools to overcome them.

8.3 Contributions

We have implemented a provenance capture pipeline which reconstructs provenance from XML-based edit histories and produces an RDF triple store from which individual provenance graph data can be captured using SPARQL based capture policies. Whilst there have been other attempts at provenance capture from OpenStreetMap edit history, to the best of our knowledge, ours is the only implementation that does so using interoperable standards. This triple store allows the extraction of W3C PROV-DM RDF provenance graphs for a wide range of scientific use cases.

We have provided a methodological framework for the descriptive analysis of OpenStreetMap provenance graph data by identifying and evaluating approaches to the measurement of network graphs. Using these approaches, we have designed metrics which enable a descriptive analysis of OpenStreetMap provenance graph data. The results of that analysis provide insight into the nature of OpenStreetMap editing and the factors which drive its variation, advancing our understanding the insights which can be gained using provenance network analytics. This lays the groundwork for an automated, principled analysis of large volumes of provenance data to provide much-needed functions such as quality/credibility/trust documentation and for providing provenance-based intelligence for humanitarian mapping efforts.

8.4 Future Work

The insights we have gained during the production of this thesis provide opportunities for further studies and for refinement of the existing methods. Addressing scale issues for provenance graphs is a priority and future work should focus on the capture of smaller feature-based provenance graphs. Simplified datasets targeting either single features or a more limited range of feature types will also produce data which is more amenable to analysis.

Further investigations will investigate changes to capture policies to improve the quality of metric data and enhance the insights which can be gained. Provenance capture policies can be designed to capture feature-based provenance and to focus on specific feature types. Capturing provenance exclusively for building footprints is one promising avenue for study. More targeted building features over wide areas can be captured by filtering for specific tags which would enable the study of commercial buildings or purely residential studies.

Further study of demographic characteristics using output areas and output area classifications could still be carried out using feature-based provenance. With many features this could be achieved by calculating a centroid point for each building and identifying its output area. Discovering relationships between provenance metrics and demographic data is a valuable research

avenue. Carrying out these investigations on data from other parts of the world will hopefully reveal themes which generalise to other regions and may allow insights to be gained in into areas which lack geodemographic data coverage. This raises the prospect of VGI provenance as a form of remote sensor from which we can learn more about the world which is being mapped by OpenStreetMap contributors.

References

[1] B. Plewe, 'www94 -- Awards', *First International Conference on the World-Wide Web*, May 1994. http://www94.web.cern.ch/WWW94/Awards0529.html (accessed Sep. 29, 2021).

[2] S. Putz, 'Interactive information services using World-Wide Web hypertext', *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 273–280, Nov. 1994, doi: 10.1016/0169-7552(94)90141-4.

[3] United States. Central Intelligence Agency, 'World Data Bank II: North America, South America, Europe, Africa, Asia'. Inter-university Consortium for Political and Social Research [distributor], 2006. doi: 10.3886/ICPSR08376.v1.

[4] U. S. Geological Survey, '1:2,000,000-scale digital line graph (DLG) data', Earth Science Information Office, 4, 1992. doi: 10.3133/ds4.

[5] M. P. Peterson, 'MapQuest and the beginnings of web cartography', *International Journal of Cartography*, vol. 7, no. 2, pp. 275–281, May 2021, doi: 10.1080/23729333.2021.1925831.

[6] D. DiNucci, 'Fragmented Future', *Print Magazine*, pp. 32, 221, 222, Apr. 1999.

[7] 'Web 2.0 Conference', *Web 2.0 Conference*, 2004. https://web.archive.org/web/20050312204307/http://www.web2con.com/web2con/ (accessed Oct. 14, 2021).

[8] J. J. Garrett, 'Ajax: A New Approach to Web Applications'. Adaptive Path, 2005. Accessed: Aug. 06, 2021. [Online]. Available: https://courses.cs.washington.edu/courses/cse490h/07sp/readings/ajax_adaptive_path.pdf

[9] 'Wiki's Wild World', *Nature*, vol. 438, no. 7070, pp. 890–890, Dec. 2005, doi: 10.1038/438890a.

[10] 'How Many Blogs Are There? (And 141 Other Blogging Stats)', *GrowthBadger*, Jan. 23, 2021. https://growthbadger.com/blog-stats/ (accessed Nov. 10, 2021).

[11] 'Blogging Statistics - Worldometer'. https://www.worldometers.info/blogs/ (accessed Nov. 10, 2021).

[12] 'Media Nations: UK 2019', OFCOM, 2019. Accessed: Nov. 01, 2021. [Online]. Available: https://www.ofcom.org.uk/__data/assets/pdf_file/0019/160714/media-nations-2019-uk-report.pdf

[13] B. Strasser, J. Baudry, D. Mahr, G. Sanchez, and É. Tancoigne, Eds., '"Citizen Science"? Rethinking Science and Public Participation', *Science & Technology Studies*, 2019, doi: 10.23987/sts.60425.

[14] A. Bruns, 'Towards produsage: Futures for user-led content production. Proceedings: Cultural Attitudes towards Communication and Technology', in *In Proceedings: Cultural Attitudes towards Communication and Technology*, 2006. [Online]. Available: http://eprints.qut.edu.au/4863/1/4863_1.pdf

[15] A. Bruns, 'Produsage: Towards a Broader Framework for User-Led Content Creation.', in *Proceedings of 6th ACM SIGCHI Conference on Creativity and Cognition 2007*, New York, NY, USA, 2007, pp. 99–106. doi: 10.1145/1254960.1254975.

[16] OECD, 'Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking', OECD, Sep. 2007. doi: 10.1787/9789264037472-en.

[17] T. K. Naab and A. Sehl, 'Studies of user-generated content: A systematic review', *Journalism*, vol. 18, no. 10, pp. 1256–1273, Nov. 2017, doi: 10.1177/1464884916673557.

[18] E. S. Raymond, *The Cathedral and the Bazaar : Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol, UNITED STATES: O'Reilly Media, Incorporated, 2001. [Online]. Available: http://ebookcentral.proquest.com/lib/soton-ebooks/detail.action?docID=443450

[19] Andrew Turner, *Introduction to Neogeography*. O'Reilly Media, Inc., 2006.

[20] M. Neal, 'The Revolution Will Be Live-Mapped: A Brief History of Protest Maptivism', Sep. 06, 2013. https://www.vice.com/en/article/aeebzp/the-revolution-will-be-live-mapped-a-brief-history-of-protest-maptivism (accessed Nov. 17, 2021).

[21] D. Butler, 'The web-wide world', *Nature*, vol. 439, no. 7078, pp. 776–778, Feb. 2006, doi: 10.1038/439776a.

[22] 'O'Reilly Where 2.0 Conference: The Future of Mapping and Local Search'. https://www.oreilly.com/pub/pr/1532 (accessed Nov. 11, 2021).

[23] M. (Muki) Haklay, 'Neogeography and the Delusion of Democratisation', *Environ Plan A*, vol. 45, no. 1, pp. 55–69, Jan. 2013, doi: 10.1068/a45184.

[24] J. B. Harley, 'Deconstructing the map', *Cartographica: The international journal for geographic information and geovisualization*, vol. 26, no. 2, p. 20, 1989.

[25] ‘What is Free Software? - GNU Project - Free Software Foundation’. https://www.gnu.org/philosophy/free-sw.html (accessed Nov. 25, 2021).

[26] P. Neis and A. Zipf, ‘Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap’, *ISPRS International Journal of Geo-Information*, vol. 1, no. 3, pp. 146–165, Jul. 2012, doi: 10.3390/ijgi1020146.

[27] Z. Gardner, P. Mooney, S. De Sabbata, and L. Dowthwaite, ‘Quantifying gendered participation in OpenStreetMap: responding to theories of female (under) representation in crowdsourced mapping’, *GeoJournal*, vol. 85, no. 6, pp. 1603–1620, Dec. 2020, doi: 10.1007/s10708-019-10035-z.

[28] M. Stephens, ‘Gender and the GeoWeb: divisions in the production of user-generated cartographic information’, *GeoJournal*, vol. 78, no. 6, pp. 981–996, Dec. 2013, doi: 10.1007/s10708-013-9492-z.

[29] J. J. Arsanjani and M. Bakillah, ‘Understanding the potential relationship between the socio-economic variables and contributions to OpenStreetMap’, *International Journal of Digital Earth*, vol. 8, no. 11, pp. 861–876, Nov. 2015, doi: 10.1080/17538947.2014.951081.

[30] N. R. Budhathoki, ‘Participants’ motivations to contribute geographic information in an online community’, University of Illinois at Urbana-Champaign, 2010. Accessed: Apr. 04, 2019. [Online]. Available: https://www.ideals.illinois.edu/handle/2142/16956

[31] D. Bégin, R. Devillers, and S. Roche, ‘Contributors’ enrollment in collaborative online communities: the case of OpenStreetMap’, *Geo-spatial Information Science*, vol. 20, no. 3, pp. 282–295, Jul. 2017, doi: 10.1080/10095020.2017.1370177.

[32] J. Bright, S. De Sabbata, and S. Lee, ‘Geodemographic biases in crowdsourced knowledge websites: Do neighbours fill in the blanks?’, *GeoJournal*, vol. 83, no. 3, pp. 427–440, Jun. 2018, doi: 10.1007/s10708-017-9778-7.

[33] M. E. Haklay, ‘Why is participation inequality important?’, in *In: Capineri, C and Haklay, M and Huang, H and Antoniou, V and Kettunen, J and Ostermann, F and Purves, R, (eds.) European Handbook of Crowdsourced Geographic Information. (pp. 35-44). Ubiquity Press: London, United Kindom. (2016)*, no. 3, C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, Eds. London, United Kindom: Ubiquity Press, 2016, pp. 35–44. Accessed: Jul. 05, 2022. [Online]. Available: http://www.ubiquitypress.com/site/books/detail/28/european-handbook-of-crowdsourced-geographic-information/

[34] C. Bittner, 'OpenStreetMap in Israel and Palestine – "Game changer" or reproducer of contested cartographies?', *Political Geography*, vol. 57, pp. 34–48, Mar. 2017, doi: 10.1016/j.polgeo.2016.11.010.

[35] A. Keen, *The cult of the amateur: how blogs, Myspace, YouTube and the rest of today's user-generated media are destroying our economy, our culture, and our values*, Revised paperback ed., Repr. London: Nicholas Brealey Publ, 2011.

[36] J. Giles, 'Internet encyclopaedias go head to head', *Nature*, Dec. 14, 2005. doi: 10.1038/438900a.

[37] C. D. A. Johnston, 'Instructions to Field Examiners'. Ordnance Survey, 1905. Accessed: Mar. 25, 2018. [Online]. Available: https://www.ordnancesurvey.co.uk/docs/ebooks/historical-instructions-to-field-examiners.pdf

[38] R. Devillers and R. Jeansoulin, Eds., *Fundamentals of spatial data quality*. London ; Newport Beach, CA: ISTE, 2006.

[39] 'ISO 19157:2013(en), Geographic information — Data quality'. https://www.iso.org/obp/ui/#iso:std:iso:19157:ed-1:v1:en (accessed Aug. 03, 2017).

[40] H. Veregin, 'Data Quality Parameters', in *Geographical Information Systems*, 2nd ed., vol. 1, P. Longley, M. Goodchild, D. Maguire, and D. Rhind, Eds. New York: Wiley, 1999.

[41] B. Cipeluch, R. Jacob, P. Mooney, and A. C. Winstanley, 'Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps', in *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, 2010, p. 337. [Online]. Available: http://eprints.maynoothuniversity.ie/2476

[42] J.-F. Girres and G. Touya, 'Quality Assessment of the French OpenStreetMap Dataset', *Transactions in GIS*, vol. 14, no. 4, pp. 435–459, Aug. 2010, doi: 10.1111/j.1467-9671.2010.01203.x.

[43] M. Haklay, 'How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets', *Environment and Planning B: Planning and Design*, vol. 37, no. 4, pp. 682–703, Aug. 2010, doi: 10.1068/b35097.

[44] M. (Muki) Haklay, S. Basiouka, V. Antoniou, and A. Ather, 'How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information', *The Cartographic Journal*, vol. 47, no. 4, pp. 315–322, Nov. 2010, doi: 10.1179/000870410X12911304958827.

[45] P. Mooney, P. Corcoran, and A. C. Winstanley, 'Towards quality metrics for OpenStreetMap', 2010, p. 514. doi: 10.1145/1869790.1869875.

[46] P. Neis, D. Zielstra, and A. Zipf, 'The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011', *Future Internet*, vol. 4, no. 4, pp. 1–21, Dec. 2011, doi: 10.3390/fi4010001.

[47] D. Zielstra and A. Zipf, 'Quantitative studies on the data quality of OpenStreetMap in Germany', in *Proceedings of the Sixth International Conference on Geographic Information Science, GIScience, Zurich, Switzerland*, 2010, pp. 20–26. Accessed: Aug. 09, 2017. [Online]. Available: https://www.researchgate.net/profile/Alexander_Zipf/publication/267989860_Quantitative_Studies_on_the_Data_Quality_of_OpenStreetMap_in_Germany/links/54d99a590cf25013d0426ba0/Quantitative-Studies-on-the-Data-Quality-of-OpenStreetMap-in-Germany.pdf

[48] D. Mcknight and N. Chervany, 'Trust and Distrust Definitions: One Bite at a Time', in *Trust in cyber-societies*, vol. 2246, 2001, pp. 27–54. doi: 10.1007/3-540-45547-7_3.

[49] A. Wierzbicki, *Web Content Credibility*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-77794-8.

[50] A. J. Flanagin and M. J. Metzger, 'The credibility of volunteered geographic information', *GeoJournal*, vol. 72, no. 3–4, pp. 137–148, Aug. 2008, doi: 10.1007/s10708-008-9188-y.

[51] C. I. Hovland, I. L. Janis, and H. H. Kelley, *Communication and persuasion; psychological studies of opinion change.* New Haven, CT, US: Yale University Press, 1953.

[52] M. J. Metzger, 'Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research', *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 13, pp. 2078–2091, Nov. 2007, doi: 10.1002/asi.v58:13.

[53] A. Wierzbicki, 'Credibility of Social Media', in *Web Content Credibility*, A. Wierzbicki, Ed. Cham: Springer International Publishing, 2018, pp. 131–153. doi: 10.1007/978-3-319-77794-8_4.

[54] T. D. Huynh, M. Ebden, J. Fischer, S. Roberts, and L. Moreau, 'Provenance Network Analytics: An approach to data analytics using data provenance', *Data Mining and Knowledge Discovery*, Feb. 2018, doi: 10.1007/s10618-017-0549-3.

[55] T. D. Huynh, M. Ebden, S. Ramchurn, S. Roberts, and L. Moreau, 'Data Quality Assessment From Provenance Graphs', presented at the Provenance Analytics 2014, Jun. 2014, p. 4. [Online]. Available: https://eprints.soton.ac.uk/365510/

[56] M. Ebden, T. Huynh, L. Moreau, S. Ramchurn, and S. Roberts, 'Network analysis on provenance graphs from a crowdsourcing application', in *Provenance and Annotation of Data and Processes*, 2012, pp. 168–182. [Online]. Available: http://link.springer.com/content/pdf/10.1007/978-3-642-34222-6.pdf#page=179

[57] 'Testbed 10 Provenence Engineering Report', OGC, OGC Engineering Report OGC 14-001, Jul. 2014. Accessed: Dec. 08, 2017. [Online]. Available: http://www.opengis.net/doc/ER/testbed10/provenance

[58] F. D'Antonio, P. Fogliaroni, and T. Kauppinen, 'VGI Edit History Reveals Data Trustworthiness and User Reputation', p. 5, 2014.

[59] P. Fogliaroni, F. D'Antonio, and E. Clementini, 'Data trustworthiness and user reputation as indicators of VGI quality', *Geo-spatial Information Science*, vol. 21, no. 3, pp. 213–233, Jul. 2018, doi: 10.1080/10095020.2018.1496556.

[60] C. Keßler and R. T. A. de Groot, 'Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap', in *Geographic Information Science at the Heart of Europe*, D. Vandenbroucke, B. Bucher, and J. Crompvoets, Eds. Cham: Springer International Publishing, 2013, pp. 21–37. doi: 10.1007/978-3-319-00615-4_2.

[61] C. Keßler, J. Trame, and T. Kauppinen, 'Provenance and Trust in Volunteered Geographic Information: The Case of OpenStreetMap', in *Proceedings of Workshop on Identifying Objects, Processes and Events in SpatioTemporally Distributed Data (IOPE 2011)*, 2011, p. 3.

[62] Q. T. Truong, C. de Runz, and G. Touya, 'Analysis of collaboration networks in OpenStreetMap through weighted social multigraph mining', *International Journal of Geographical Information Science*, vol. 33, no. 8, pp. 1651–1682, Aug. 2019, doi: 10.1080/13658816.2018.1556395.

[63] Jacob Nielsen, 'Participation Inequality: The 90-9-1 Rule for Social Features', *Nielsen Norman Group*. https://www.nngroup.com/articles/participation-inequality/ (accessed Jul. 05, 2022).

[64] D. Begin, R. Devillers, and S. Roche, 'The Life Cycle of Volunteered Geographic Information (VGI) Contributors: the OpenStreetMap Example', *International Conference on GIScience Short Paper Proceedings*, vol. 1, no. 1, 2016, doi: 10.21433/B31146p8p31g.

[65] A. Yang, H. Fan, and N. Jing, 'Amateur or Professional: Assessing the Expertise of Major Contributors in OpenStreetMap Based on Contributing Behaviors', *ISPRS International Journal of Geo-Information*, vol. 5, no. 2, Art. no. 2, Feb. 2016, doi: 10.3390/ijgi5020021.

[66] Z. Gardner, P. Mooney, L. Dowthwaite, and G. Foody, 'Gender differences in OpenStreetMap contributor activity, editing and tagging behaviour', in *Schedule - GISRUK 2018*, 2018, p. 6. [Online]. Available: http://gisruk.org/ProceedingsGISRUK2018/GISRUK2018_Contribution_045.pdf

[67] M. Schmidt and S. Klettner, 'Gender and Experience-Related Motivators for Contributing to OpenStreetMap', in *Online proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information*, Leuven, Belgium, 2013, p. 4. [Online]. Available: https://flrec.ifas.ufl.edu/geomatics/agile2013/presentations/ACTIVITY_WS_AGILE_2013_SESSION_1_Schmidt.pdf

[68] G. Quattrone, A. Mashhadi, and L. Capra, 'Mind the map: the impact of culture and economic affluence on crowd-mapping behaviours', in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, Baltimore Maryland USA, Feb. 2014, pp. 934–944. doi: 10.1145/2531602.2531713.

[69] 'Verifiability - OpenStreetMap Wiki'. https://wiki.openstreetmap.org/wiki/Verifiability (accessed Jul. 21, 2022).

[70] D. Coleman, Y. Georgiadou, and J. Labonte, 'Volunteered Geographic Information: the nature and motivation of produsers', *International Journal of Spatial Data Infrastructures Research*, vol. 4, no. 4, Art. no. 4, Apr. 2009.

[71] N. R. Budhathoki and C. Haythornthwaite, 'Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap', *American Behavioral Scientist*, vol. 57, no. 5, pp. 548–575, May 2013, doi: 10.1177/0002764212469364.

[72] P. Mooney and P. Corcoran, 'How social is OpenStreetMap?', in *Proceedings of the AGILE'2012 International Conference on Geographic Information Science*, Avignon, 2012, p. 6.

[73] R. Steinmann, S. Gröchenig, K. Rehrl, and R. Brunauer, 'Contribution profiles of voluntary mappers in OpenStreetMap', in *Proceedings of Action and Interaction in Volunteered Geographic Information (ACTIVITY) Workshop at AGILE*, Leuven, Belgium, 2013, pp. 1–6.

[74] Y.-W. Lin, 'A qualitative enquiry into OpenStreetMap making', *New Review of Hypermedia and Multimedia*, vol. 17, no. 1, pp. 53–71, Apr. 2011, doi: 10.1080/13614568.2011.552647.

[75] W. Lin, 'Revealing the making of OpenStreetMap: A limited account', *The Canadian Geographer / Le Géographe canadien*, vol. 59, no. 1, pp. 69–81, 2015, doi: 10.1111/cag.12137.

[76] J. Anderson, D. Sarkar, and L. Palen, 'Corporate Editors in the Evolving Landscape of OpenStreetMap', *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, Art. no. 5, May 2019, doi: 10.3390/ijgi8050232.

[77] 'Organised Editing Guidelines'. OpenStreetMap Foundation, 2018. Accessed: Jul. 17, 2022. [Online]. Available: https://wiki.osmfoundation.org/wiki/Organised_Editing_Guidelines

[78] T. W. Bank, 'Ramani Huria : the atlas of flood resilience in Dar es Salaam', The World Bank Group, Washington D.C, Working Paper 125586, Jan. 2018. Accessed: Sep. 17, 2019. [Online]. Available: http://documents.worldbank.org/curated/en/200421524092301920/Ramani-Huria-the-atlas-of-flood-resilience-in-Dar-es-Salaam

[79] E. Hagen, 'Open mapping from the ground up: learning from Map Kibera', Institute of Development Studies, Brighton, Making All Voices Count Research Report, 2017. [Online]. Available: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/123456789/13244/RReport_MapKibera_Online.pdf

[80] M. Kogan, J. Anderson, L. Palen, K. M. Anderson, and R. Soden, 'Finding the Way to OSM Mapping Practices: Bounding Large Crisis Datasets for Qualitative Investigation', in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2016, pp. 2783–2795. doi: 10.1145/2858036.2858371.

[81] M. (Muki) Haklay *et al.*, 'Identifying Success Factors in Crowdsourced Geographic Information Use in Government', Global Facility for Disaster Reduction & Recovery, World Bank, Working Paper 139461, Dec. 2018.

[82] D. Zielstra, H. H. Hochmair, P. Neis, and F. Tonini, 'Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap', *ISPRS International Journal of Geo-Information*, vol. 3, no. 4, Art. no. 4, Dec. 2014, doi: 10.3390/ijgi3041211.

[83] M. Napolitano and P. Mooney, 'MVP OSM: a tool to identify areas of high quality contributor activity in OpenStreetMap', *The Bulletin of the Society of Cartographers*, vol. 45, no. 1, Art. no. 1, 2012.

[84] E. Strano, V. Nicosia, V. Latora, S. Porta, and M. Barthélemy, 'Elementary processes governing the evolution of road networks', *Sci Rep*, vol. 2, no. 1, p. 296, Mar. 2012, doi: 10.1038/srep00296.

[85] P. Corcoran, P. Mooney, and M. Bertolotto, 'Analysing the growth of OpenStreetMap networks', *Spatial Statistics*, vol. 3, pp. 21–32, Feb. 2013, doi: 10.1016/j.spasta.2013.01.002.

[86] J. J. Arsanjani, M. Helbich, M. Bakillah, and L. Loos, 'The emergence and evolution of OpenStreetMap: a cellular automata approach', *International Journal of Digital Earth*, vol. 8, no. 1, pp. 76–90, Jan. 2015, doi: 10.1080/17538947.2013.847125.

[87] J. J. Arsanjani, P. Mooney, M. Helbich, and A. Zipf, 'An Exploration of Future Patterns of the Contributions to OpenStreetMap and Development of a Contribution Index', *Transactions in GIS*, vol. 19, no. 6, pp. 896–914, Dec. 2015, doi: 10.1111/tgis.12139.

[88] V. Antoniou and C. Schlieder, 'Addressing Uneven Participation Patterns in VGI Through Gamification Mechanisms', in *Geogames and Geoplay: Game-based Approaches to the Analysis of Geo-Information*, O. Ahlqvist and C. Schlieder, Eds. Cham: Springer International Publishing, 2018, pp. 91–110. doi: 10.1007/978-3-319-22774-0_5.

[89] G. Quattrone, A. Mashhadi, D. Quercia, C. Smith-Clarke, and L. Capra, 'Modelling Growth of Urban Crowd-sourced Information', in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2014, pp. 563–572. doi: 10.1145/2556195.2556244.

[90] D. Bégin, R. Devillers, and S. Roche, 'Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours', in *Proceedings of the 8th international symposium on spatial data quality ISSDQ*, 2013, pp. 149–154. [Online]. Available: http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-2-W1/149/2013/isprsarchives-XL-2-W1-149-2013.pdf

[91] A. Comber, P. Mooney, R. S. Purves, D. Rocchini, and A. Walz, 'Crowdsourcing: It Matters Who the Crowd Are. The Impacts of between Group Variations in Recording Land Cover', *PLOS ONE*, vol. 11, no. 7, p. e0158329, Jul. 2016, doi: 10.1371/journal.pone.0158329.

[92] D. M. Wilkinson and B. A. Huberman, 'Cooperation and Quality in Wikipedia', in *Proceedings of the 2007 International Symposium on Wikis*, New York, NY, USA, 2007, pp. 157–164. doi: 10.1145/1296951.1296968.

[93] M. Goodchild, 'NeoGeography and the nature of geographic expertise', *Journal of Location Based Services*, vol. 3, no. 2, pp. 82–96, Jun. 2009, doi: 10.1080/17489720902950374.

[94] P. Maué and S. Schade, 'Quality Of Geographic Information Patchworks', in *Agile - Proceedings 2008*, Girona, Spain, 2008, p. 8.

[95] C. Barrington-Leigh and A. Millard-Ball, 'The world's user-generated road map is more than 80% complete', *PLOS ONE*, vol. 12, no. 8, p. e0180698, Aug. 2017, doi: 10.1371/journal.pone.0180698.

[96] H. Fan, A. Zipf, Q. Fu, and P. Neis, 'Quality assessment for building footprints data on OpenStreetMap', *International Journal of Geographical Information Science*, vol. 28, pp. 700–719, Apr. 2014, doi: 10.1080/13658816.2013.867495.

[97] R. Hecht, C. Kunze, and S. Hahmann, 'Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time', *ISPRS International Journal of Geo-Information*, vol. 2, no. 4, Art. no. 4, Dec. 2013, doi: 10.3390/ijgi2041066.

[98] H. Dorn, T. Törnros, and A. Zipf, 'Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany', *ISPRS International Journal of Geo-Information*, vol. 4, no. 3, Art. no. 3, Sep. 2015, doi: 10.3390/ijgi4031657.

[99] M. Haklay, 'How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets', *Environment and Planning B: Planning and Design*, vol. 37, no. 4, pp. 682–703, Aug. 2010, doi: 10.1068/b35097.

[100] P. Mooney and P. Corcoran, 'The Annotation Process in OpenStreetMap', *Transactions in GIS*, vol. 16, no. 4, pp. 561–579, Aug. 2012, doi: 10.1111/j.1467-9671.2012.01306.x.

[101] J. Bright, S. De Sabbata, S. Lee, B. Ganesh, and D. K. Humphreys, 'OpenStreetMap data for alcohol research: Reliability assessment and quality indicators', *Health & Place*, vol. 50, pp. 130–136, Mar. 2018, doi: 10.1016/j.healthplace.2018.01.009.

[102] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad, 'Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?', Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2595096, 2011. doi: 10.2139/ssrn.2595096.

[103] M. Jilani, P. Corcoran, and M. Bertolotto, 'Multi-granular Street Network Representation towards Quality Assessment of OpenStreetMap Data', in *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science*, New York, NY, USA, Nov. 2013, pp. 19–24. doi: 10.1145/2533828.2533833.

[104] Christopher Braune and Jens Klump, *Exploring the Quality and Usability of OpenStreetMap Data*. 2014.

[105] J. Harding, 'Usability of geographic information – Factors identified from qualitative analysis of task-focused user interviews', *Applied Ergonomics*, vol. 44, no. 6, pp. 940–947, Nov. 2013, doi: 10.1016/j.apergo.2012.11.013.

[106] M. Brown *et al.*, 'Usability of Geographic Information: Current challenges and future directions', *Applied Ergonomics*, vol. 44, no. 6, pp. 855–865, Nov. 2013, doi: 10.1016/j.apergo.2012.10.013.

[107] C. Barron, P. Neis, and A. Zipf, 'A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis: A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis', *Transactions in GIS*, vol. 18, no. 6, pp. 877–895, Dec. 2014, doi: 10.1111/tgis.12073.

[108] C. Keßler, J. Trame, and T. Kauppinen, 'Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap', presented at the Workshop at COSIT 2011: 10th International Conference on Spatial Information Theory, 2011, p. 8. [Online]. Available: https://carsten.io/wp-content/uploads/papers/iope2011.pd

[109] A. Mobasheri, Y. Sun, L. Loos, and A. Ali, 'Are Crowdsourced Datasets Suitable for Specialized Routing Services? Case Study of OpenStreetMap for Routing of People with Limited Mobility', *Sustainability*, vol. 9, no. 7, p. 997, Jun. 2017, doi: 10.3390/su9060997.

[110] M. F. Goodchild, 'Citizens as sensors: the world of volunteered geography', *GeoJournal*, vol. 69, no. 4, pp. 211–221, Nov. 2007, doi: 10.1007/s10708-007-9111-y.

[111] P. Buneman, S. Khanna, and W.-C. Tan, 'Data Provenance: Some Basic Issues', in *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, vol. 1974, S. Kapoor and S. Prasad, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 87–93. doi: 10.1007/3-540-44450-5_6.

[112] Luc Moreai and Paolo Missier, Eds., 'PROV-DM: The PROV Data Model'. W3C, 2013. Accessed: Jan. 19, 2018. [Online]. Available: https://www.w3.org/TR/2013/REC-prov-dm-20130430/

[113] Y. Cui and J. Widom, 'Practical lineage tracing in data warehouses', 2000, pp. 367–378. doi: 10.1109/ICDE.2000.839437.

[114] Y. Cui, J. Widom, and J. L. Wiener, 'Tracing the lineage of view data in a warehousing environment', *ACM Transactions on Database Systems*, vol. 25, no. 2, pp. 179–227, Jun. 2000, doi: 10.1145/357775.357777.

[115] A. Woodruff and M. Stonebraker, 'Supporting Fine-Grained Data Lineage in a Database Visualization Environment', in *Proceedings 13th International Conference on Data Engineering*, Apr. 1997, pp. 91–102. doi: 10.1109/ICDE.1997.581742.

[116] P. Buneman, S. Khanna, and T. Wang-Chiew, 'Why and Where: A Characterization of Data Provenance', in *Database Theory — ICDT 2001*, Berlin, Heidelberg, 2001, pp. 316–330. doi: 10.1007/3-540-44503-X_20.

[117] J. Cheney, L. Chiticariu, and W.-C. Tan, 'Provenance in Databases: Why, How, and Where', *Foundations and Trends in Databases*, vol. 1, no. 4, pp. 379–474, 2007, doi: 10.1561/1900000006.

[118] A. Chapman and H. V. Jagadish, 'Why Not?', in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2009, pp. 523–534. doi: 10.1145/1559845.1559901.

[119] Y. Gil *et al.*, 'Examining the Challenges of Scientific Workflows', *Computer*, vol. 40, no. 12, pp. 24–32, Dec. 2007, doi: 10.1109/MC.2007.421.

[120] T. Fahringer *et al.*, 'ASKALON: a Grid application development and computing environment', 2005, p. 10 pp. doi: 10.1109/GRID.2005.1542733.

[121] C. Goble, C. Wroe, and R. Stevens, *The myGrid Project: Services, Architecture and Demonstrator*. 2003.

[122] Oinn Tom *et al.*, 'Taverna: lessons in creating a workflow environment for the life sciences', *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1067–1100, Dec. 2005, doi: 10.1002/cpe.993.

[123] B. Ludäscher *et al.*, 'Scientific workflow management and the Kepler system', *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1039–1065, Aug. 2006, doi: 10.1002/cpe.994.

[124] I. Foster, J. Vockler, M. Wilde, and Y. Zhao, 'Chimera: a virtual data system for representing, querying, and automating data derivation', in *Proceedings 14th International Conference on Scientific and Statistical Database Management*, 2002, pp. 37–46. doi: 10.1109/SSDM.2002.1029704.

[125] E. Deelman *et al.*, 'Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems', *Scientific Programming*, vol. 13, no. 3, pp. 219–237, 2005, doi: 10.1155/2005/128026.

[126] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, 'VisTrails: visualization meets data management', 2006, p. 745. doi: 10.1145/1142473.1142574.

[127] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo, 'Managing Rapidly-Evolving Scientific Workflows', in *Provenance and Annotation of Data*, Berlin, Heidelberg, 2006, vol. 4145, pp. 10–18. doi: 10.1007/11890850_2.

[128] J. Zhao, C. Wroe, C. Goble, R. Stevens, D. Quan, and M. Greenwood, 'Using Semantic Web Technologies for Representing E-science Provenance', in *The Semantic Web – ISWC 2004*, 2004, pp. 92–106.

[129] S. Bowers, T. McPhillips, and B. Ludäscher, 'Validation and Inference of Schema-Level Workflow Data-Dependency Annotations', in *Provenance and Annotation of Data and Processes*, 2018, pp. 128–141.

[130] P. Alper, K. Belhajjame, C. A. Goble, and P. Karagoz, 'Enhancing and abstracting scientific workflow provenance for data publishing', 2013, p. 313. doi: 10.1145/2457317.2457370.

[131] S. Bharathi, A. Chervenak, E. Deelman, G. Mehta, M. Su, and K. Vahi, 'Characterization of scientific workflows', in *2008 Third Workshop on Workflows in Support of Large-Scale Science*, Nov. 2008, pp. 1–10. doi: 10.1109/WORKS.2008.4723958.

[132] L. Ramakrishnan and B. Plale, 'A Multi-dimensional Classification Model for Scientific Workflow Characteristics', in *Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science*, New York, NY, USA, 2010, p. 4:1-4:12. doi: 10.1145/1833398.1833402.

[133] S. Bowers and B. Ludäscher, 'Actor-Oriented Design of Scientific Workflows', in *Conceptual Modeling – ER 2005*, 2005, pp. 369–384.

[134] J. Ferrante, K. J. Ottenstein, and J. D. Warren, 'The Program Dependence Graph and Its Use in Optimization', *ACM Trans. Program. Lang. Syst.*, vol. 9, no. 3, pp. 319–349, Jul. 1987, doi: 10.1145/24039.24041.

[135] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer, 'Provenance-Aware Storage Systems', in *Proceedings of the Annual Conference on USENIX '06 Annual Technical Conference*, Berkeley, CA, USA, 2006, p. 4. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267359.1267363

[136] N. Balakrishnan, T. Bytheway, R. Sohan, and A. Hopper, 'OPUS: A Lightweight System for Observational Provenance in User Space', in *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance*, Berkeley, CA, USA, 2013, p. 4. [Online]. Available: https://dl.acm.org/doi/10.5555/2482613.2482621

[137] A. Gehani and D. Tariq, 'SPADE: Support for Provenance Auditing in Distributed Environments', in *Middleware 2012*, Berlin, Heidelberg, 2012, pp. 101–120.

[138] P. Groth, S. Miles, W. Fang, S. C. Wong, K. Zauner, and L. Moreau, 'Recording and using provenance in a protein compressibility experiment', 2005, pp. 201–208. doi: 10.1109/HPDC.2005.1520960.

[139] S. C. Chan, A. Gehani, J. Cheney, R. Sohan, and H. Irshad, 'Expressiveness Benchmarking for System-Level Provenance', in *9th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2017)*, Seattle, WA, 2017. [Online]. Available: https://www.usenix.org/conference/tapp17/workshop-program/presentation/chan

[140] M. D. Allen, L. Seligman, B. Blaustein, and A. Chapman, 'Provenance Capture and Use: A Practical Guide'. MITRE, 2010. [Online]. Available: https://www.mitre.org/publications/technical-papers/provenance-capture-and-use-a-practical-guide

[141] L. Moreau *et al.*, 'The Open Provenance Model core specification (v1.1)', *Future Generation Computer Systems*, vol. 27, no. 6, pp. 743–756, Jun. 2011, doi: 10.1016/j.future.2010.07.005.

[142] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, 'The Open Provenance Model: An Overview', in *Provenance and Annotation of Data and Processes*, Berlin, Heidelberg, 2008, pp. 323–326. doi: 10.1007/978-3-540-89965-5_31.

[143] OpenStreetMap Wiki Contributors, 'Any tags you like - OpenStreetMap Wiki', *OpenStreetMap Wiki*. http://wiki.openstreetmap.org/wiki/Any_tags_you_like (accessed Aug. 30, 2017).

[144] S. Miles, P. Groth, S. Munroe, and L. Moreau, 'PrIMe: A methodology for developing provenance-aware applications', *ACM Transactions on Software Engineering and Methodology*, vol. 20, no. 3, pp. 1–42, Aug. 2011, doi: 10.1145/2000791.2000792.

[145] L. Murta, V. Braganholo, F. Chirigati, D. Koop, and J. Freire, 'noWorkflow: Capturing and Analyzing Provenance of Scripts', in *Provenance and Annotation of Data and Processes*, Cologne, 2014, pp. 71–83. doi: 10.1007/978-3-319-16462-5_6.

[146] T. McPhillips *et al.*, 'YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts', *IJDC*, vol. 10, no. 1, pp. 298–313, Feb. 2015, doi: 10.2218/ijdc.v10i1.370.

[147] T. McPhillips, S. Bowers, K. Belhajjame, and B. Ludäscher, 'Retrospective Provenance Without a Runtime Provenance Recorder', in *7th USENIX Workshop on the Theory and Practice of Provenance (TaPP 15)*, Edinburgh, Scotland, 2015. [Online]. Available: https://www.usenix.org/conference/tapp15/workshop-program/presentation/mcphillips

[148] M. D. Allen, A. Chapman, B. Blaustein, and L. Seligman, 'Capturing provenance in the wild', in *International Provenance and Annotation Workshop*, 2010, pp. 98–101.

[149] A. Chapman, B. T. Blaustein, L. Seligman, and M. D. Allen, 'PLUS: A provenance manager for integrated information', in *2011 IEEE International Conference on Information Reuse Integration*, Las Vegas, Aug. 2011, pp. 269–275. doi: 10.1109/IRI.2011.6009558.

[150] M. Whittaker, C. Teodoropol, P. Alvaro, and J. M. Hellerstein, 'Debugging Distributed Systems with Why-Across-Time Provenance', in *Proceedings of the ACM Symposium on Cloud Computing - SoCC '18*, Carlsbad, CA, USA, 2018, pp. 333–346. doi: 10.1145/3267809.3267839.

[151] D. Ghoshal and B. Plale, 'Provenance from Log Files: a BigData Problem', in *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, New York, NY, USA, Mar. 2013, pp. 290–297. doi: 10.1145/2457317.2457366.

[152] S. Magliacane and P. Groth, 'Towards Reconstructing the Provenance of Clinical Guidelines', in *Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences*, Amsterdam, 2016, p. 4. [Online]. Available: http://ceur-ws.org/Vol-952/paper_36.pdf

[153] A. Aierken, D. B. Davis, Q. Zhang, K. Gupta, A. Wong, and H. U. Asuncion, 'A Multi-level Funneling Approach to Data Provenance Reconstruction', in *2014 IEEE 10th International Conference on e-Science*, Oct. 2014, vol. 2, pp. 71–74. doi: 10.1109/eScience.2014.54.

[154] S. Vasudevan, W. Pfeffer, D. Davis, and H. Asuncion, 'Improving data provenance reconstruction via a multi-level funneling approach', in *2016 IEEE 12th International Conference on e-Science (e-Science)*, Oct. 2016, pp. 175–184. doi: 10.1109/eScience.2016.7870898.

[155] T. D. Nies, S. Magliacane, R. Verborgh, S. Coppens, E. Mannens, and R. V. de Walle, 'Git2PROV: Exposing Version Control System Content as W3C PROV', in *Proceedings of the ISWC 2013*, Sydney, Australia, 2013, vol. 1035, p. 4. [Online]. Available: http://ceur-ws.org/Vol-1035/iswc2013_demo_32.pdf

[156] I. Taxidou, T. De Nies, R. Verborgh, P. M. Fischer, E. Mannens, and R. Van de Walle, 'Modeling Information Diffusion in Social Media As Provenance with W3C PROV', in *Proceedings of*

the 24th International Conference on World Wide Web, New York, NY, USA, 2015, pp. 819–824. doi: 10.1145/2740908.2742475.

[157] Y. Yan and T. McLane, 'Metadata Management and Revision History Tracking for Spatial Data and GIS Map Figures', in *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*, New York, NY, USA, 2012, p. 35:1-35:2. doi: 10.1145/2345316.2345357.

[158] 'About tracking an editor's changes to data—Help | ArcGIS for Desktop', *ArcGIS for Desktop*, 2016. http://desktop.arcgis.com/en/arcmap/10.3/manage-data/editing-fundamentals/about-tracking-an-editor-s-changes-to-data.htm (accessed May 02, 2019).

[159] 'GRASS GIS Manual: Vector data processing in GRASS GIS', *Grass GIS Manual*, 2018. https://grass.osgeo.org/grass70/manuals/vectorintro.html (accessed May 02, 2019).

[160] P. Groth and L. Moreau, 'PROV-Overview', 2013. https://www.w3.org/TR/prov-overview/

[161] 'ISO 19115-1:2014 - Geographic information -- Metadata -- Part 1: Fundamentals'. Accessed: Aug. 11, 2017. [Online]. Available: https://www.iso.org/standard/53798.html

[162] 'OWS-9 Cross Community Interoperability (CCI) Conflation with Provenance Engineering Report', OGC, OGC Engineering Report OGC 12-159, Feb. 2013. Accessed: Dec. 08, 2017. [Online]. Available: https://portal.opengeospatial.org/files/?artifact_id=51818

[163] L. Moreau and P. Groth, *Provenance: an introduction to PROV*. San Rafael, Calif.: Morgan & Claypool, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6812734?arnumber=6812734

[164] L. Jiang, P. Yue, W. Kuhn, C. Zhang, C. Yu, and X. Guo, 'Advancing interoperability of geospatial data provenance on the web: Gap analysis and strategies', *Computers & Geosciences*, vol. 117, pp. 21–31, Aug. 2018, doi: 10.1016/j.cageo.2018.05.001.

[165] A. J. Simmons, S. Barnett, S. Vajda, and R. Vasa, 'Data Provenance for Sport', *arXiv:1812.05804 [cs]*, Dec. 2018, Accessed: Mar. 22, 2019. [Online]. Available: http://arxiv.org/abs/1812.05804

[166] P. Macko and M. Seltzer, 'Provenance Map Orbiter: Interactive Exploration of Large Provenance Graphs', in *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance*, 2011, p. 6.

[167] R. Hoekstra and P. Groth, 'PROV-O-Viz - Understanding the Role of Activities in Provenance', in *Provenance and Annotation of Data and Processes*, vol. 8628, B. Ludäscher and B. Plale, Eds. Cham: Springer International Publishing, 2015, pp. 215–220. doi: 10.1007/978-3-319-16462-5_18.

[168] T. Kohwalter, T. Oliveira, J. Freire, E. Clua, and L. Murta, 'Prov Viewer: A Graph-Based Visualization Tool for Interactive Exploration of Provenance Data', in *Provenance and Annotation of Data and Processes*, Cham, 2016, vol. 9672, pp. 71–82. doi: 10.1007/978-3-319-40593-3_6.

[169] O. Hartig and J. Zhao, 'Using web data provenance for quality assessment', in *In: Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*, 2009.

[170] J. E. G. Malaverri, C. B. Medeiros, and R. C. Lamparelli, 'A provenance approach to assess the quality of geospatial data', 2012, p. 2043. doi: 10.1145/2245276.2232116.

[171] R. Ikeda and J. Widom, 'Data lineage: A survey', 2009.

[172] T. Pasquier *et al.*, 'Practical Whole-System Provenance Capture', *arXiv:1711.05296 [cs]*, pp. 405–418, 2017, doi: 10.1145/3127479.3129249.

[173] G. B. Coe, R. C. Doty, M. D. Allen, and A. P. Chapman, 'Provenance Capture Disparities Highlighted through Datasets', p. 4, 2014.

[174] P. Yue, M. Zhang, X. Guo, and Z. Tan, 'Granularity of geospatial data provenance', in *2014 IEEE Geoscience and Remote Sensing Symposium*, Jul. 2014, pp. 4492–4495. doi: 10.1109/IGARSS.2014.6947490.

[175] P. Missier, J. Bryans, C. Gamble, V. Curcin, and R. Danger, 'ProvAbs: Model, Policy, and Tooling for Abstracting PROV Graphs', in *Provenance and Annotation of Data and Processes*, Jun. 2014, pp. 3–15. doi: 10.1007/978-3-319-16462-5_1.

[176] T. D. Huynh, M. Ebden, M. Venanzi, S. D. Ramchurn, S. Roberts, and L. Moreau, 'Interpretation of crowdsourced activities using provenance network analysis', in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7388

[177] R. J. Webber, 'Making the Most of the Census for Strategic Analysis', *The Town Planning Review*, vol. 49, no. 3, pp. 274–284, 1978.

[178] J. L. Falguera, C. Martínez-Vidal, and G. Rosen, 'Abstract Objects', in *The Stanford Encyclopedia of Philosophy*, Winter 2021., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford

University, 2021. Accessed: Feb. 15, 2022. [Online]. Available: https://plato.stanford.edu/archives/win2021/entries/abstract-objects/

[179] D. K. Lewis, *On the plurality of worlds*. Malden, Mass: Blackwell Publishers, 2001.

[180] P. Jamieson, 'More graph comparison techniques on mind maps to provide students with feedback', in *2013 IEEE Frontiers in Education Conference (FIE)*, Oct. 2013, pp. 992–998. doi: 10.1109/FIE.2013.6684976.

[181] L. H. Gomes, R. B. Almeida, L. M. A. Bettencourt, V. A. F. Almeida, and J. M. Almeida, 'Comparative Graph Theoretical Characterization of Networks of Spam', in *CEAS*, 2005.

[182] M. Jahnke, C. Thul, and P. Martini, 'Graph based Metrics for Intrusion Response Measures in Computer Networks', in *32nd IEEE Conference on Local Computer Networks (LCN 2007)*, Oct. 2007, pp. 1035–1042. doi: 10.1109/LCN.2007.45.

[183] L. Deuker *et al.*, 'Reproducibility of graph metrics of human brain functional networks', *NeuroImage*, vol. 47, no. 4, pp. 1460–1468, Oct. 2009, doi: 10.1016/j.neuroimage.2009.05.035.

[184] M. E. J. Newman, *Networks: an introduction*. Oxford ; New York: Oxford University Press, 2010.

[185] M. E. J. Newman, 'The Structure and Function of Complex Networks', *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, Jan. 2003, doi: 10.1137/S003614450342480.

[186] A. Hagberg, P. Swart, and D. S Chult, 'Exploring network structure, dynamics, and function using NetworkX', Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[187] M. E. J. Newman, 'Mixing patterns in networks', *Phys. Rev. E*, vol. 67, no. 2, p. 026126, Feb. 2003, doi: 10.1103/PhysRevE.67.026126.

[188] 'Britannica attacks', *Nature*, vol. 440, no. 7084, pp. 582–582, Mar. 2006, doi: 10.1038/440582b.

[189] T. Chesney, 'An empirical examination of Wikipedia's credibility', *First Monday*, Nov. 2006, doi: 10.5210/fm.v11i11.1413.

[190] M. Messner and M. W. DiStaso, 'Wikipedia versus Encyclopedia Britannica: A Longitudinal Analysis to Identify the Impact of Social Media on the Standards of Knowledge', *Mass Communication and Society*, vol. 16, no. 4, pp. 465–486, Jul. 2013, doi: 10.1080/15205436.2012.732649.

[191] ‘mature, adj. and n.’, *OED Online*. Oxford University Press. Accessed: Mar. 01, 2022. [Online]. Available: https://www.oed.com/view/Entry/115114

[192] ‘What Is Mature Economy?’, *Investopedia*. https://www.investopedia.com/terms/m/mature-economy.asp (accessed Mar. 01, 2022).

[193] ‘Wikipedia:Featured articles’, *Wikipedia*. Mar. 01, 2022. Accessed: Mar. 01, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia:Featured_articles&oldid=1074581139

[194] Stephen Maguire and M. Tomko, ‘Ripe for the picking? Dataset maturity assessment based on temporal dynamics of feature definitions’, *International Journal of Geographical Information Science*, vol. 31, no. 7, pp. 1334–1358, Jul. 2017, doi: 10.1080/13658816.2017.1287370.

[195] ‘Wikipedia:Content assessment’, *Wikipedia*. Feb. 18, 2022. Accessed: Mar. 01, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=1072540861

[196] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, ‘Information quality work organization in wikipedia’, *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 983–1001, 2008, doi: 10.1002/asi.20813.

[197] ‘Wikipedia:Statistics’, *Wikipedia*. Feb. 02, 2020. Accessed: Feb. 06, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia:Statistics&oldid=938748498

[198] M. Anderka, B. Stein, and N. Lipka, ‘Predicting Quality Flaws in User-generated Content: The Case of Wikipedia’, in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2012, pp. 981–990. doi: 10.1145/2348283.2348413.

[199] D. Hasan Dalip, M. André Gonçalves, M. Cristo, and P. Calado, ‘Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of Wikipedia’, in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2009, pp. 295–304. doi: 10.1145/1555400.1555449.

[200] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, ‘Assessing information quality of a community-based encyclopedia’, in *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005*, 2005. [Online]. Available: http://www.scopus.com/inward/record.url?scp=84871554587&partnerID=8YFLogxK

[201] A. Lih, 'Wikipedia as Participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource', in *In Proceedings of the 5th International Symposium on Online Journalism*, 2004, pp. 16–17.

[202] T. Wöhner and R. Peters, 'Assessing the Quality of Wikipedia Articles with Lifecycle Based Metrics', in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, New York, NY, USA, 2009, p. 16:1-16:10. doi: 10.1145/1641309.1641333.

[203] S. Gröchenig, R. Brunauer, and K. Rehrl, 'Estimating Completeness of VGI Datasets by Analyzing Community Activity Over Time Periods', in *Connecting a Digital Europe Through Location and Place*, J. Huerta, S. Schade, and C. Granell, Eds. Cham: Springer International Publishing, 2014, pp. 3–18. doi: 10.1007/978-3-319-03611-3_1.

[204] K. Rehrl, S. Gröechenig, H. Hochmair, S. Leitinger, R. Steinmann, and A. Wagner, 'A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI', in *Progress in Location-Based Services*, J. M. Krisp, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 373–388. doi: 10.1007/978-3-642-34203-5_21.

[205] G. Quattrone, M. Dittus, and L. Capra, 'Work Always in Progress: Analysing Maintenance Practices in Spatial Crowd-sourced Datasets', in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, New York, NY, USA, 2017, pp. 1876–1889. doi: 10.1145/2998181.2998267.

[206] P. Mooney and P. Corcoran, 'Characteristics of Heavily Edited Objects in OpenStreetMap', *Future Internet*, vol. 4, no. 1, pp. 285–305, Mar. 2012, doi: 10.3390/fi4010285.

[207] X. Li, Z. Luo, K. Pang, and T. Wang, 'A Lifecycle Analysis of the Revision Behavior of Featured Articles on Wikipedia', in *2013 International Conference on Information Science and Cloud Computing Companion*, Dec. 2013, pp. 846–851. doi: 10.1109/ISCC-C.2013.16.

[208] A. Gehani, H. Kazmi, and H. Irshad, 'Scaling SPADE to "Big Provenance"', p. 8.

[209] L. Moreau, 'Aggregation by Provenance Types: A Technique for Summarising Provenance Graphs', *Electron. Proc. Theor. Comput. Sci.*, vol. 181, pp. 129–144, Apr. 2015, doi: 10.4204/EPTCS.181.9.

[210] B. Glavic, 'Big Data Provenance: Challenges and Implications for Benchmarking', in *Specifying Big Data Benchmarks*, Berlin, Heidelberg, 2014, pp. 72–80. doi: 10.1007/978-3-642-53974-9_7.

[211] G. Closa, J. Masó, B. Proß, and X. Pons, 'W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment', *Computers, Environment and Urban Systems*, vol. 64, pp. 103–117, Jul. 2017, doi: 10.1016/j.compenvurbsys.2017.01.008.

[212] J. Zhao, C. Goble, and R. Stevens, 'An Identity Crisis in the Life Sciences', in *Provenance and Annotation of Data*, Berlin, Heidelberg, 2006, pp. 254–269. doi: 10.1007/11890850_26.

[213] W. S. Robinson, 'Ecological Correlations and the Behavior of Individuals', *American Sociological Review*, vol. 15, no. 3, pp. 351–357, 1950, doi: 10.2307/2087176.

[214] C. D. Lloyd, *Exploring spatial scale in geography*. Chichester, West Sussex, ; Hoboken, NJ: Wiley Blackwell, 2014.

[215] C. E. Gehlke and K. Biehl, 'Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material', *Journal of the American Statistical Association*, vol. 29, no. 185, pp. 169–170, 1934, doi: 10.2307/2277827.

[216] G. U. Yule and M. G. Kendall, 'Correlation and Regression: Some Practical Problems', in *An Introduction to the Theory of Statistics*, 14th ed., Dehli: Universal Book Stall, 1951.

[217] G. U. Yule and M. G. Kendall, 'An Introduction to the Theory of Statistics', *Journal of Symbolic Logic*, vol. 16, no. 1, pp. 51–51, 1951.

[218] H. M. Blalock, *Causal inferences in nonexperimental research*. Chapel Hill: Univ. of North Carolina Press, 1964.

[219] W. a. V. Clark and K. L. Avery, 'The Effects of Data Aggregation in Statistical Analysis', *Geographical Analysis*, vol. 8, no. 4, pp. 428–438, 1976, doi: 10.1111/j.1538-4632.1976.tb00549.x.

[220] H. C. Selvin, 'Durkheim's Suicide and Problems of Empirical Research', *American Journal of Sociology*, vol. 63, no. 6, pp. 607–619, 1958.

[221] S. Openshaw, *The Modifiable Areal Unit Problem*. Norwich [Norfolk]: Geo Books, 1984.

[222] S. Openshaw, 'An Empirical Study of Some Zone-Design Criteria', *Environ Plan A*, vol. 10, no. 7, pp. 781–794, Jul. 1978, doi: 10.1068/a100781.

[223] S. Openshaw and R. S. Baxter, 'Algorithm 3: A Procedure to Generate Pseudo-Random Aggregations of N Zones into M Zones, Where M is Less Than N', *Environ Plan A*, vol. 9, no. 12, pp. 1423–1428, Dec. 1977, doi: 10.1068/a091423.

[224] D. Martin, 'Optimizing census geography: the separation of collection and output geographies', *International Journal of Geographical Information Science*, vol. 12, no. 7, pp. 673–685, Nov. 1998, doi: 10.1080/136588198241590.

[225] D. Martin, A. Nolan, and M. Tranmer, 'The Application of Zone-Design Methodology in the 2001 UK Census', *Environ Plan A*, vol. 33, no. 11, pp. 1949–1962, Nov. 2001, doi: 10.1068/a3497.

[226] R. Flowerdew, 'How serious is the Modifiable Areal Unit Problem for analysis of English census data?', *Popul Trends*, vol. 145, no. 1, pp. 106–118, Sep. 2011, doi: 10.1057/pt.2011.20.

[227] D. W. MacDonald, 'Beyond the Group: The Implications of Roderick D. McKenzie's Human Ecology for Reconceptualizing Society and the Social', *Nature and Culture*, vol. 6, no. 3, pp. 263–284, Dec. 2011, doi: 10.3167/nc.2011.060304.

[228] A. D. Singleton and S. E. Spielman, 'The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom', *The Professional Geographer*, vol. 66, no. 4, pp. 558–567, Oct. 2014, doi: 10.1080/00330124.2013.848764.

[229] D. Vickers and P. Rees, 'Creating the UK National Statistics 2001 output area classification', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 2, pp. 379–403, 2007, doi: 10.1111/j.1467-985X.2007.00466.x.

[230] G. W. Milligan and M. C. Cooper, 'Methodology Review: Clustering Methods', *Applied Psychological Measurement*, vol. 11, no. 4, pp. 329–354, Dec. 1987, doi: 10.1177/014662168701100401.

[231] C. G. Gale, A. D. Singleton, A. G. Bates, and P. A. Longley, 'Creating the 2011 area classification for output areas (2011 OAC)', *J. Spat. Int. Sci.*, no. 12, pp. 1–27, 2016, doi: 10.5311/JOSIS.2016.12.232.

[232] P. Shannon *et al.*, 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[233] QGIS.org, 'QGIS Geographic Information System'. QGIS Association, 2022. [Online]. Available: http://www.qgis.org

[234] G. Closa, J. Masó, A. Zabala, L. Pesquer, and X. Pons, 'A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation', *Transactions in GIS*, vol. 0, no. 0, doi: 10.1111/tgis.12555.

[235] S. Wang, A. Padmanabhan, J. Myers, W. Tang, and Y. Liu, *Towards provenance-aware geographic information systems*. 2008. doi: 10.1145/1463434.1463515.

[236] L. Anselin, S. J. Rey, and W. Li, 'Metadata and provenance for spatial analysis: the case of spatial weights', *International Journal of Geographical Information Science*, vol. 28, no. 11, pp. 2261–2280, Nov. 2014, doi: 10.1080/13658816.2014.917313.

[237] A. Bates, 'Pen Portraits for the 2011 Area Classification for Output Areas', ONS, 2015.

[238] K. A. Bollen, 'Latent Variables in Psychology and the Social Sciences', *Annual Review of Psychology*, vol. 53, no. 1, pp. 605–634, 2002, doi: 10.1146/annurev.psych.53.100901.135239.

[239] C. Spearman, '"General Intelligence," Objectively Determined and Measured', *The American Journal of Psychology*, vol. 15, no. 2, pp. 201–292, 1904, doi: 10.2307/1412107.

[240] L. L. Thurstone, 'The vectors of mind.', *Psychological Review*, vol. 41, no. 1, pp. 1–32, 1934, doi: 10.1037/h0075959.

[241] D. J. Bartholomew, 'Factor Analysis and Latent Structure: Overview', in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 5249–5254. doi: 10.1016/B0-08-043076-7/00425-3.

[242] E. C. Tupes and R. E. Christal, 'Recurrent Personality Factors Based on Trait Ratings', *Journal of Personality*, vol. 60, no. 2, pp. 225–251, 1961, doi: 10.1111/j.1467-6494.1992.tb00973.x.

[243] R. R. McCrae and O. P. John, 'An introduction to the five-factor model and its applications.', *Journal of Personality*, Jun. 1992, doi: 10.1111/j.1467-6494.1992.tb00970.x.

[244] H. J. Eysenck, *The structure of human personality*, 3rd ed. London: Routledge, 2013.

[245] A. G. C. Wright, 'The Current State and Future of Factor Analysis in Personality Disorder Research', *Personal Disord*, vol. 8, no. 1, pp. 14–25, Jan. 2017, doi: 10.1037/per0000216.

[246] R. Howard, O. Almeida, and R. Levy, 'Phenomenology, demography and diagnosis in late paraphrenia', *Psychological Medicine*, vol. 24, no. 2, pp. 397–410, May 1994, doi: 10.1017/S0033291700027379.

[247] G. Remafedi, M. Resnick, R. Blum, and L. Harris, 'Demography of Sexual Orientation in Adolescents', *Pediatrics*, vol. 89, no. 4, pp. 714–721, Apr. 1992, doi: 10.1542/peds.89.4.714.

[248] A. Field, *Discovering statistics using IBM SPSS statistics*, 5th edition. Thousand Oaks, CA: SAGE Publications, 2017.

[249] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*, 4th ed. Boston, MA: Allyn and Bacon, 2001.

[250] E. Guadagnoli and W. Velicer, 'Relation of Sample Size to the Stability of Component Patterns', *Psychological bulletin*, vol. 103, pp. 265–75, Apr. 1988, doi: 10.1037//0033-2909.103.2.265.

[251] W. Tobler, 'On the First Law of Geography: A Reply', *Annals of the Association of American Geographers*, vol. 94, no. 2, pp. 304–310, Jun. 2004, doi: 10.1111/j.1467-8306.2004.09402009.x.

[252] W. Tobler, 'Linear pycnophylactic reallocation comment on a paper by D. Martin', *International Journal of Geographical Information Science*, vol. 13, no. 1, pp. 85–90, Jan. 1999, doi: 10.1080/136588199241472.

[253] C. G. Thompson, R. S. Kim, A. M. Aloe, and B. J. Becker, 'Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results', *Basic and Applied Social Psychology*, vol. 39, no. 2, pp. 81–90, Mar. 2017, doi: 10.1080/01973533.2016.1277529.

[254] H. F. Kaiser, 'An index of factorial simplicity', *Psychometrika*, vol. 39, no. 1, pp. 31–36, Mar. 1974, doi: 10.1007/BF02291575.

[255] K. J. Preacher and R. C. MacCallum, 'Repairing Tom Swift's Electric Factor Analysis Machine', *Understanding Statistics*, vol. 2, no. 1, pp. 13–43, Feb. 2003, doi: 10.1207/S15328031US0201_02.

[256] L. R. Tucker, 'The Objective Definition of Simple Structure in Linear Factor Analysis*', *ETS Research Bulletin Series*, vol. 1954, no. 2, pp. i–32, 1954, doi: https://doi.org/10.1002/j.2333-8504.1954.tb00486.x.

[257] L. L. Thurstone, *Multiple factor analysis.* University of Chicago Press: Chicago, 1947.

[258] W. F. Mullen, S. P. Jackson, A. Croitoru, A. Crooks, A. Stefanidis, and P. Agouris, 'Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features', *GeoJournal*, vol. 80, no. 4, pp. 587–605, Aug. 2015, doi: 10.1007/s10708-014-9564-8.

[259] J. W. Tukey, *Exploratory data analysis*. 1977.

[260] D. C. Hoaglin and B. Iglewicz, ‘Fine-Tuning Some Resistant Rules for Outlier Labeling’, *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1147–1149, 1987, doi: 10.2307/2289392.

[261] J. Laurikkala, M. Juhola, and E. Kentala, *Informal Identification of Outliers in Medical Data*. 2000.

[262] ‘Laerd Statistics’, *One-way MANOVA using SPSS Statistics. Statistical tutorials and software guides.*, 2015. https://statistics.laerd.com/ (accessed Sep. 05, 2021).

[263] C. J. Huberty and M. D. Petoskey, ‘Multivariate Analysis of Variance and Covariance’, in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, H. E. A. Tinsley and S. D. Brown, Eds. San Diego: Academic Press, 2000, pp. 183–208. doi: 10.1016/B978-012691360-6/50008-2.

[264] T. Kariya, ‘Robustness of Multivariate Tests’, *The Annals of Statistics*, vol. 9, no. 6, pp. 1267–1275, Nov. 1981, doi: 10.1214/aos/1176345643.

[265] H. Finch and B. French, ‘A Monte Carlo Comparison of Robust MANOVA Test Statistics’, *J. Mod. App. Stat. Meth.*, vol. 12, no. 2, pp. 35–81, Nov. 2013, doi: 10.22237/jmasm/1383278580.

[266] A. Khan and G. D. Rayner, ‘Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem’, *Journal Of Applied Mathematics And Decision Sciences*, vol. 7, p. 21.

[267] C. L. Olson, ‘On choosing a test statistic in multivariate analysis of variance.’, *Psychological Bulletin*, vol. 83, no. 4, p. 579, 1976, doi: 10.1037/0033-2909.83.4.579.

[268] C. L. Olson, ‘Comparative Robustness of Six Tests in Multivariate Analysis of Variance’, *Journal of the American Statistical Association*, vol. 69, no. 348, pp. 894–908, 1974, doi: 10.2307/2286159.

[269] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, N.J: L. Erlbaum Associates, 1988.

[270] ‘RDF 1.1 Concepts and Abstract Syntax’, *World Wide Web Consortium (W3C)*, 2014. https://www.w3.org/TR/rdf11-concepts/ (accessed Sep. 21, 2022).

[271] ‘RDF 1.1 XML Syntax’, *World Wide Web Consortium (W3C)*, 2014. https://www.w3.org/TR/rdf-syntax-grammar/ (accessed Sep. 21, 2022).

[272] ‘OWL 2 Web Ontology Language Primer (Second Edition)’, *World Wide Web Consortium (W3C)*, 2012. https://www.w3.org/TR/owl-primer/#What_is_OWL_2.3F (accessed Jan. 13, 2017).

[273] Timothy Lebo, Satya Sahoo, and Deborah McGuinness, Eds., ‘PROV-O: The PROV Ontology’. W3C, 2013. Accessed: Aug. 08, 2017. [Online]. Available: https://www.w3.org/TR/prov-o/

[274] ‘SPARQL 1.1 Overview’, *World Wide Web Consortium (W3C)*. https://www.w3.org/TR/sparql11-overview/ (accessed Sep. 21, 2022).

[275] J. Bennett, *OpenStreetMap: be your own cartographer*. Birmingham, U.K.: Packt Pub., 2010. Accessed: Jun. 27, 2017. [Online]. Available: http://site.ebrary.com/id/10428653

[276] B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov, ‘OWLIM: A Family of Scalable Semantic Repositories’, *Semant. web*, vol. 2, no. 1, pp. 33–42, Jan. 2011.

[277] Ontotext, ‘GraphDB Documentation’, *GraphDB Documentation*, 2017. https://graphdb.ontotext.com/documentation/8.0/free/ (accessed Sep. 20, 2022).

[278] D. Hunter, *Beginning XML*. Indianapolis, Ind.: Wiley, 2007.

[279] D. Livingston, *Essential XML for Web professionals*. Upper Saddle River, NJ: Prentice Hall PTR, 2002.

[280] A. Lohfink and D. McPhee, ‘A Mechanism for the Representation of Versions in Linked Administrative Geographic Data’, in *Proceedings of the GIS Research UK 18th Annual Conference*, 2010, p. 10.

[281] Michael Kay, ‘Saxon-HE’. Saxonica, 2018. [Online]. Available: https://www.saxonica.com/welcome/welcome.xml

[282] T. D. Huynh, ‘prov: A library for W3C Provenance Data Model supporting PROV-JSON, PROV-XML and PROV-O (RDF)’. Accessed: Sep. 21, 2022. [OS Independent]. Available: https://github.com/trungdong/prov

[283] G. Kocevar *et al.*, ‘Graph Theory-Based Brain Connectivity for Automatic Classification of Multiple Sclerosis Clinical Courses’, *Front. Neurosci.*, vol. 10, p. 478, Oct. 2016, doi: 10.3389/fnins.2016.00478.

[284] D. Bégin, R. Devillers, and S. Roche, ‘The life cycle of contributors in collaborative online communities -the case of OpenStreetMap’, *International Journal of Geographical Information Science*, vol. 32, no. 8, pp. 1611–1630, Aug. 2018, doi: 10.1080/13658816.2018.1458312.

[285] P. Mooney and P. Corcoran, 'Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors', *Transactions in GIS*, vol. 18, no. 5, pp. 633–659, 2014, doi: 10.1111/tgis.12051.

[286] A. Venerandi, G. Quattrone, L. Capra, D. Quercia, and D. Saez-Trumper, 'Measuring Urban Deprivation from User Generated Content', in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, Vancouver, BC, Canada, Feb. 2015, pp. 254–264. doi: 10.1145/2675133.2675233.

[287] D. Martin, '2001 Census output areas: from concept to prototype', *Popul Trends*, no. 94, pp. 19–24, 1998.

[288] R. Ewing and R. Cervero, 'Travel and the Built Environment: A Synthesis', *Transportation Research Record*, vol. 1780, no. 1, pp. 87–114, Jan. 2001, doi: 10.3141/1780-10.

[289] C. Sarkar, J. Gallacher, and C. Webster, 'Built environment configuration and change in body mass index: The Caerphilly Prospective Study (CaPS)', *Health & Place*, vol. 19, pp. 33–44, Jan. 2013, doi: 10.1016/j.healthplace.2012.10.001.

[290] D. C. van Wijk, J. O. Groeniger, F. J. van Lenthe, and C. B. M. Kamphuis, 'The role of the built environment in explaining educational inequalities in walking and cycling among adults in the Netherlands', *International Journal of Health Geographics*, vol. 16, no. 1, p. 10, Mar. 2017, doi: 10.1186/s12942-017-0083-y.

[291] Q. Zhou, 'Exploring the relationship between density and completeness of urban building data in OpenStreetMap for quality estimation', *International Journal of Geographical Information Science*, vol. 32, no. 2, pp. 257–281, Feb. 2018, doi: 10.1080/13658816.2017.1395883.

[292] 'OS MasterMap Topography Layer Technical Specification', Ordnance Survey, Southampton, 2017. [Online]. Available: https://www.ordnancesurvey.co.uk/docs/technical-specifications/os-mastermap-topography-layer-technical-specification.pdf

[293] B. Efron, 'Better Bootstrap Confidence Intervals', *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 171–185, 1987, doi: 10.2307/2289144.

[294] F. D. Schönbrodt and M. Perugini, 'At what sample size do correlations stabilize?', *Journal of Research in Personality*, vol. 47, no. 5, pp. 609–612, Oct. 2013, doi: 10.1016/j.jrp.2013.05.009.

[295] J. Cohen, 'A power primer.', *Psychological Bulletin*, vol. 112, no. 1, pp. 155–159, 1992, doi: 10.1037/0033-2909.112.1.155.

[296] P. J. Allen and K. Bennett, *SPSS for the health & behavioural sciences*. Australia: Thomson, 2008.

[297] A. Kittur and R. E. Kraut, 'Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination', in *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, New York, NY, USA, 2008, pp. 37–46. doi: 10.1145/1460563.1460572.