Taylor & Francis
Taylor & Francis Group

Annals of GIS

ISSN: 1947-5683 (Print) 1947-5691 (Online) Journal homepage: www.tandfonline.com/journals/tagi20

‘Research ready’ geographically enabled smart data

Paul A. Longley, James Cheshire & Alex Singleton

To cite this article: Paul A. Longley, James Cheshire & Alex Singleton (2024) ‘Research ready’ geographically enabled smart data, Annals of GIS, 30:3, 267-273, DOI: 10.1080/19475683.2024.2353035

To link to this article: https://doi.org/10.1080/19475683.2024.2353035

Published online: 15 May 2024.

ANNALS OF GIS
2024, VOL. 30, NO. 3, 267–273
https://doi.org/10.1080/19475683.2024.2353035

Taylor & Francis
Taylor & Francis Group

'Research ready' geographically enabled smart data

Paul A. Longley[a,b], James Cheshire[a] and Alex Singleton[c]

[a]Department of Geography, University College London, London, UK; [b]School of Geography, Nanjing Normal University, Nanjing, China; [c]Department of Geography and Planning, University of Liverpool, Liverpool, UK

ABSTRACT
This paper reviews and assesses the prospects for developing geographically enabled research ready data (RRD) with reference to current UK initiatives. Examples of projects for which such data have been provisioned are given.

ARTICLE HISTORY
Received 5 January 2024
Accepted 5 May 2024

KEYWORDS
Smart data; research ready data; social investigation

1. Smart data

This is a paper about using new sources and forms of data to develop socially inclusive representations of humans and their geographic interactions, and then delivering these representations for analysis by a greatly extended mass research culture of users. At its core, any quantitative generalization about the world requires focus upon specific human characteristics and activities at known levels of detail, to enable shared understanding of the way that the world looks and works. Generalized representation is informed by existing knowledge and directed towards its extension, cognizant of complexity of individual human characteristics and agency. In this context, the advent of Big Data has presented a dual-edged sword: on the one hand, richer and more timely recording of human agency and environments has created an ocean of new opportunity for harvesting salient behavioural descriptors and predictors, but on the other, the challenges of data selectivity, preparation for purpose, management and linkage have multiplied exponentially.

Current UK initiatives in what has been termed 'Smart Data' (e.g. Department for Business, Energy and Industrial Strategy 2021) are both a response to the technical challenges and the growing public appetite for better control and governance of data collected about them. In this context, our operational perspective on Smart Data is designed to achieve systematic and scientific consolidation of raw data arising from human interactions with all manner of digital devices. Consolidation may itself entail 'smart' analysis using AI-driven methods, cognizant of societal structures and effective governance of data sourced from multiple third-party providers (TPPs).

Seen from this perspective, the Smart Data research agenda emerging in the UK reflects international developments in data collection, curation and analysis that require focus upon:

(a) uncertainty: developing an understanding of its nature and detail with particular attention to conveying which individuals, characteristics and places are under-, over- or mis-represented. This requires rethinking of the remit of metadata, along with recognition that the nature and impacts of uncertainty differs amongst research applications.
(b) linkage: preventing the ambiguity of aggregation and ecological fallacy by focusing upon unique and identifiable individuals. Only linkage at the level of the individual can rectify what Goodchild (2015, 2022) has described as 'Balkanisation of the quantitative self' in separate smart data holdings, often only in anonymized or aggregated form.
(c) improved access: commercial sensitivities render most Smart Data inaccessible to the academic community, and sector-wide data licencing is required to address this.
(d) infrastructure: presenting inclusive representations of all of society brings issues of effective data governance, sound research ethics and effective disclosure control. It also requires developing appropriate spatial and temporal aggregations for dissemination of desensitized derivative datasets. And finally
(e) preparation for purpose: developing Smart Data pre-processing procedures to avoid duplication of

CONTACT Paul A. Longley p.longley@ucl.ac.uk Department of Geography, University College London, Gower Street, London WC1E 6BT, UK

effort by the wide and growing community of data users. This will also improve the safety, integrity and applicability of smart data to new research applications.

Here we reflect on the evolution of social science research practices and propose new ones for the Smart Data era. We describe how 'research ready data' (RRD) may be developed to support these new practices as Smart Data infrastructure underpinned by innovation-led data services and assess some implications for sustainable, timely and data rich research.

2. Smart data and social investigation

Moser and Kalton's (1971) *Survey Methods in Social Investigation* remains a convincing manifesto for the design and implementation of inclusive social science, founded upon data collected in ways that are robust, transparent and open to scrutiny. Core to this conception of social science, 'investigations' entail primary data collection, not in controlled laboratory conditions but in the observable world. Research design is used to ensure that sampled respondents are representative of the known populations from which they are drawn. Each element of the population has a known and prespecified chance of selection. This enables hypothesis led research, using the apparatus of inferential statistics to model likely characteristics of the vast majority of the unobserved population. The extent and scope of such investigations are ultimately determined by available resources, which in practice may limit the value of research resources to a tiny minority of researchers. For the adequately resourced, linear project design proceeds through successive steps of population definition, drawing and supplementing of samples, checking for patterning in non-response, inferring from samples to populations and report-writing based on the results.

Although such surveys remain important today (albeit plagued by falling response rates), this world fundamentally changed at the end of the last century with the advent of networked computing and data warehouses (Goodchild and Longley 1999). While enabling a mass research culture through re-use of multiple large datasets, this innovation brought considerable diminution of control of research design. Data proliferation has further accelerated these changes in the smart data era, bringing still greater broadening of research opportunity but greater potential risks to inclusivity of social science research.

3. Research ready data

Smart Data can deliver rich and timely representations of what is going on in society, albeit that the provenance of these representations is much less well understood than those developed from survey-led investigation. Data availability has been further decoupled from research design, creating algorithmic bias in data driven research that may severely limit validity, interpretability and applicability. Such outcomes can be avoided only if data content and coverage are understood prior to analysis, requiring pre-processing and preparation for purpose. Pre-processing requirements may be specific to a particular project with unique and distinctive characteristics, or more generic to projects that share concerns with society-wide representation. In the latter case, pre-processing of RRD can realize economies of scale for a broad range of applications. Carefully managed, such RRD can attain known levels of internal consistency, resonant of data collection using the linear project design of survey-led social investigation.

The creation of smart RRD has up-front costs that are quite different from those of conventional linear research design surveys. In the latter, costs are borne by the funder, and justified by the value of the end-uses to which the data are put. In contrast, Smart Data (in common with administrative data) are collected for different and probably unrelated purposes, and can be acquired for research use only pursuant to acquisition from data owners (Lansley and Cheshire 2018). They usually also require data subject legal consents. Creation of RRD from smart data thus begins with negotiation of data access under multi-lateral data licencing agreements (DLAs), with due regard to all anticipated uses.

Perceptions of data protection legislation and penalties for misuse provide speedbumps to data acquisition, since sharing creates risk, and thus data partnerships are likely to be forged only with trusted data partners. Trust is typically built up over time through authentic engagements within the context of projects that address shared concerns and realize mutual benefits. Such interactions both underscore the value of collaborative activities and the outputs that these generate, and cumulative interaction builds the trust required to enable sector-wide agreements and a culture of collaboration. Data partnerships and innovation in smart data can also create virtuous cycles of identifying and then exploiting new potential data sources, alongside assessment of their likely value and usefulness across social science research.

The complexity of negotiating data acquisition may also be eased through provision of secure trusted research environments (TREs) and contract provisions

that potential disclosure risks will be minimized through strong data governance and day-to-day management by accredited trusted researchers. Data services providing access to such facilities typically have up-front costs of platform development, ensuring safe researcher training, implementing strong governance of the application processes, and output checking.

RRD can thus arise from the application of widely accepted procedures to render Smart Data usable for a predefined range of Social Science applications. Many assemblages of Smart Data are best thought of as digital traces of human actions and occurrences, only some of which may be pertinent to important research questions (Cheshire 2020). The creation of smart RRD is the outcome of selectivity and preparation for purpose, including ascertaining the extent of population coverage. This typically requires triangulation of Smart Data with other sources of known representation (e.g. censuses) which, although typically much less frequently collected and lacking rich detail, can nevertheless frame smart data coverage (Gibbs et al. 2023). Quantitative and qualitative documentation of coverage can be used to communicate the usefulness of smart RRD in developing different research applications.

RRD may be developed from a single smart data source or may be derived from concatenation and conflation of multiple sources – including smart, administrative, or conventional statistical components. Linkage should be at the most granular level of detail possible to avoid ecological fallacy or scale and aggregation effects. This argues for retention of Personal Data (as defined by General Data Protection Regulation, GDPR) at the level of the individual or the use of pseudonymisation procedures that do not force reliance upon artificial aggregations – since these may preclude further linkage with other data sources pursuant to further RRD creation.

4. Delivering smart RRD

Innovation in RRD creation is a necessary but not sufficient condition for enabling the widest community of smart data researchers: data services need to become similarly innovative and attuned to current and envisaged user requirements. This is crucial, since the content and coverage of RRD very much shapes the nature of the research hypotheses that can be investigated subsequently.

A successful data service will develop and engage a community of practice around existing and prospective data assets, to the greater good of the research community. Numerically, the greatest benefits will accrue to the many RRD users, but service-led innovation should also avail specialist users of the tacit knowledge necessary to hone raw (or less pre-processed) smart data to more specialized research tasks. This not only enables potentially transformative new research but also traces a path for further service led RRD innovation.

Flexible specialization in RRD creation and maintenance argues for consolidation of data services at specialized centres, with systematic development of data pipelines using common procedures. McGrath-Lone et al (2022) propose a framework for effective provision of RRD derived from administrative sources, which we adapt for FAIR (findable, accessible, interoperable, reusable) Smart Data RRD in Figure 1.

The first and foremost requirement for smart data RRD is for vision and **breadth**. Commonalties of interest that motivate pilot projects or specialist collaborations between smart data owners and academics may be developed through bilateral data sharing agreements for specific geographic localities or customer segments. Such pilots usefully establish proof of concept but value for a broader constituency of research users requires community-wide multilateral licences for national or other significant administrative jurisdictions. A strategic and cost-effective approach to smart RRD creation and maintenance also requires that acquisitions can be linked – through an integrated smart data spine that readily enables cross-classification with new or updated acquisitions. Whilst it can be more challenging to secure multiple data licencing agreements for the full extent of datasets, ambition in acquisition has the potential to drive RRD creation at rates commensurate with the explosion in available smart data. This is most readily achieved if data are neither anonymized nor aggregated. An innovation-led data service should also develop a schedule of data linkage and updates in response to monitoring of patterns of user applications and known requirements that remain unmet.

The second consideration is **curation** of RRD. This entails updating of the data infrastructure underpinning existing RRD, including but not limited to the addition of new data. It includes procedures of internal validation through longitudinal profiling of underpinning infrastructure, and external validation with respect to framework data (such as censuses and property gazetteers) or other ancillary sources. This enables improved harmonization of the multiple smart data sources used to create smart RRD and suggests potential new RRD products that the research community would find useful. It also anticipates the need for effective data governance and disclosure control.

Third, innovation-led data services must render **accessible** the pipeline of smart RRD to a large and diverse research user community. Proliferation of data sources creates a requirement for data catalogues that

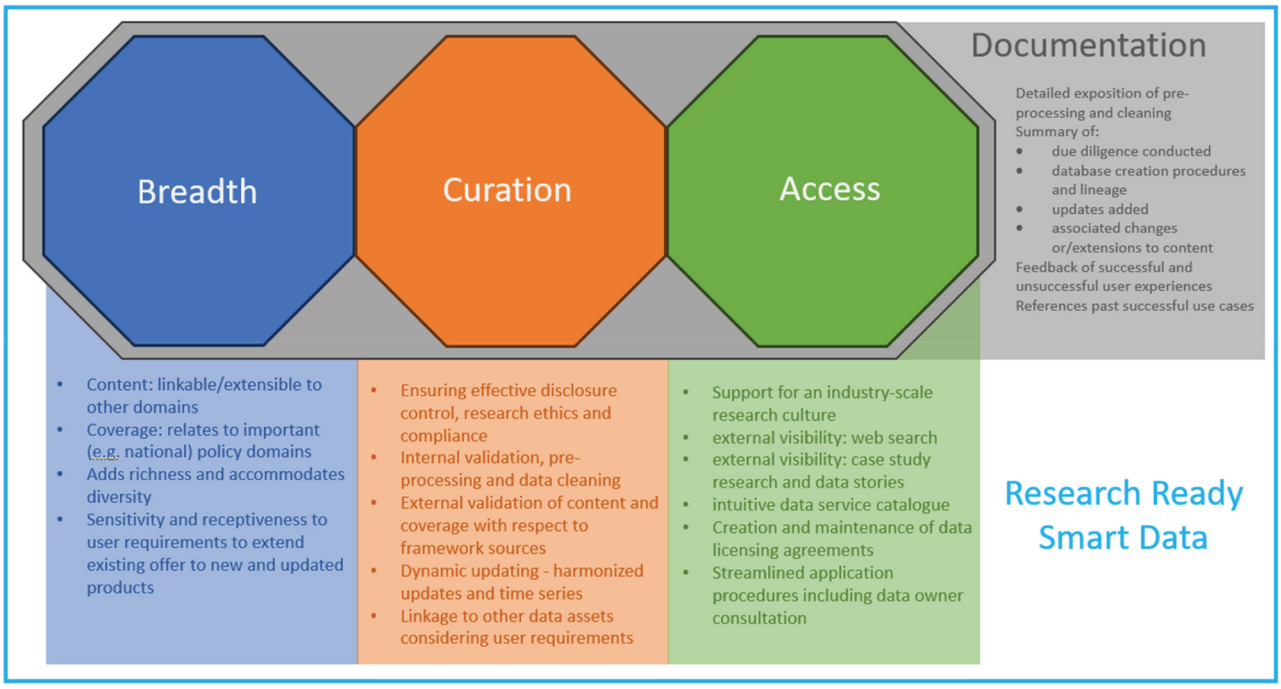

Figure 1. The effective provision of smart data RRD.

are prominent in search engines and accessible to users. Data stories of use cases can provide less experienced prospective users with guides to the relevance of resources to their own research. Application procedures should be streamlined, while background service work provides continuity of data licencing – which is likely to become complex where RRD are derived from multiple sources. Ideally data should be available under open access licences, but issues of ownership, service funding and licencing will typically necessitate additional access protocols – for example providing for data owner clearance in circumstances in which there are potential commercial conflicts of interest.

Data service breadth, curation and accessibility each also require smart RRD **documentation** of work undertaken prior to and following acquisition. This includes summaries of due diligence conducted prior to acquisition, database creation procedures and pipelines used, updates added, and associated changes or extensions to content considering user feedback. The latter should also include references to published work. Crucially, many owners of smart data are private sector operators in markets in which they have no monopoly of provision. As a result, data are likely to be biased in content and may be geographically restricted in coverage. This requires documentation in quantitative and qualitative terms – the former with respect to available representative validation data such as censuses and property gazetteers, and the latter with respect to how and why the data were generated and the mission statements of the data provider. This underpinning work may entail new research on linkage and validation, and such work should ideally be subjected to critical external scrutiny – for example through external peer review of research papers.

These detailed and specific research-led processes are all required for creation, maintenance and delivery of smart RRD. The requirement is to retain the richness of Smart Data while establishing a reliable basis for generalization when data are exploited in new research settings. Innovation-led data services will improve technical aspects of data manageability, without unduly restricting the range of research questions which the raw data might be used to address. Accommodating feedback of the experience and requirements of users is integral to this process. A successful data centre is thus both technocratic and research focused, balancing the opportunity costs and benefits of prioritizing different RRD products.

5. Spatial data infrastructure underpinning smart RRD

Realization of innovation-led smart RRD will typically require effective linkage of smart data sources with each other, as well as with other RRD validation sources. Geography provides an obvious linkage medium, yet geographic detail is almost invariably the first casualty of disclosure control and privacy maintenance. This is despite accumulated evidence that TREs are effective at

Table 1. Illustrative applications built using RRD derived from the LCRs.

- Gentrification, displacement, and the impacts of council estate renewal in 21st century London
- Tower Hamlets (London) joint needs strategic assessments: health services
- Examining the relationships between population churn, socioeconomic status and inpatient admissions in Wales
- The geography of bus crime: a case study of Greater Manchester
- Assessing domestic abuse revictimisation risk using police data and open data sources
- Understanding ethnic variation in use of health services
- Disclosure of spatial peculiarities of Brexit
- Small area estimation for crime surveys
- The spatial and temporal patterns of residential house prices and housing affordability in England
- Small area estimates of psycho-social constructs in Wales
- School quality, school choice and patterns of residential mobility
- Evaluating the reliability of using consumer generated data on residential mobility patterns to identify studentified areas in UK cities
- Stratford town centre masterplan – socio-economic baseline
- Social boundaries in the street
- Exploring changes to the spatial heterogeneity and clustering of ethnic groups using Electoral Roll and consumer register data
- Rapid analysis of ethnic variation in COVID-19 outcomes
- The geography of risk to COVID-19 infections and benchmarking incidence rates against post-2011 estimates of neighbourhood composition
- An econometric analysis of household recycling rates: comparing London to the rest of England.
- Exploring the relationship between ethnic heterogeneity, intergroup relations and stress
- Understanding inequalities in access to public transport
- Pushed to the margins: race, class and gentrification in London
- Understanding the impact of COVID-19 within Solihull's communities
- Hybrid geodemographics and creation of the 2021 Output Area Classification
- Area stability measures over time
- Exploring links between deprivation, ethnicity, travel behaviour and public transport provision
- The structural transformation of the public sphere: high street changes and populist vote
- Supporting the Chagossian community in Wythenshawe
- Are certain patient ethnicities waiting longer for treatment?
- Investigating the spatial and socio-demographic heterogeneity in covid-related changes in mobility behaviour
- Deriving small area estimates of chronic pain and disability
- Eye health inequalities in England
- Analysing the relationship between ethnic minority proportions and tree canopy cover in London at neighbourhood level
- Giving nature a home in Cardiff
- Creating a mapping tool to predict areas of low influenza vaccination take-up to guide targeted vaccination efforts.
- Activity, mobility, demographics and the formation of COVID-19 hotspots
- Urban mobility inequalities in London: extending public transport accessibility
- Investigating socioeconomic and geographical disparities in vascular surgery rates and outcomes in England
- Open space and psychological wellbeing during COVID-19 lockdowns
- Combining spatial and sociodemographic regression techniques to predict dwelling fire risk at neighbourhood level in the UK
- Assessing the factors associated with fertility declines in London, 2010–2019.
- Understanding the changing market structure of the UK alcohol outlet industry: an epidemiological approach
- Estimating the distribution of ethnic groups across new parliamentary constituency boundaries
- Culture and community spaces at risk
- Measuring changes in the urban environment in London using street view imagery
- A spatio-temporal analysis of environmental inequality in Dorset, UK
- Intergenerational inequality, age concentration, and amenities
- The "reach" of London's developing cycle network
- Developing a social vulnerability to air pollution index for greater London
- Examining social disorganization theory in Nottingham in the COVID-19 lockdown periods
- School quality and income segregation as manifest through house prices.
- Well-being and the ethnic composition of British neighbourhoods
- Identifying educationally isolated schools: the creation and use of a composite indicator of educational isolation
- An analysis of fire risk in the ethnic communities in Humberside
- Transit-induced-gentrification? Evidence from the greater London area
- Modelling gentrification in contemporary London: an agent-based approach.
- City-wide re-sorting effects of estate regeneration projects
- The distribution of gentrification in London
- Mapping gentrification in Greater Manchester
- The algorithms behind car insurance premiums
- Environmental quality, residential mobility, and human capital
- The residential displacement effects of accessibility in transport
- The effects of gentrification on voter behaviour in Greater London

managing risks associated with service and research applications. Concurrently, the widespread adoption of advanced high-resolution georeferencing technologies offers a powerful tool for quantifying and accommodating gaps in smart data representations. Such developments enhance our ability to understand the limitations and biases inherent in smart data, while also demonstrating the potential of georeferencing to link historical and present-day data sources (e.g. Lan, van Dijk, and Longley 2021; Longley, van Dijk, and Lan 2021).

The value of using TREs to retain georeferenced individual-level data has been demonstrated by Lansley et al (2019) and van Dijk et al (2021) in the creation of a 'Smart Data spine' of linked consumer registers (LCRs) for the

UK. This research has assembled and modelled population-wide data pertaining to adult individuals and the properties in which they live from smart and administrative sources, including consumer lifestyle surveys, Electoral Registers, domestic energy performance certificates, rental listings and property sales during the last quarter century. Completeness varies between sources for the composite LCRs, but geographic referencing of individual records enables scale-free measures of coverage to be created. Linkage of annualized smart and administrative data transactions enables updates, alongside internal validation and the synthetic estimation of missing data by borrowing strength from the rich longitudinal profile of existing records. Some 'less often heard' individuals are nevertheless missing from the consolidated database, but the aggregate composition of these omissions is ascertained through external validation with infrequently collected framework data sources, such as censuses. In this context, a future challenge for geo-AI will be synthesis of missing individuals and their characteristics, where precise residential locations are known (Singleton, Alexiou, and Savani 2020). Micro geodemographic patterns of social similarities, derived from the geodemographics literature (building upon Wyszomierski 2024, Chapter 6), will be integral to these endeavours.

6. Prospects

Here, we have presented a UK perspective on the direction of travel towards Smart Data infrastructure. We have created a number of derivative datasets from our LCR datasets, which enable updates to conventional statistics (e.g. annual estimates of residential turnover, and small area modelled ethnicity proportions), more granular estimates (e.g. distances of residential moves into, out of and within administrative units), and linked data not otherwise available (e.g. changes in neighbourhood deprivation, or hardship, experienced following residential moves). Our experience is that selective abstractions from the LCRs that we have created are of wide use within the academic community, as illustrated in Table 1.

Future development will require a research agenda that will upscale smart RRD creation, validation and maintenance with respect to clearly defined populations, augmented by AI-driven synthesis of missing elements. Retention of detailed geography will be integral to the design of the underpinning infrastructure but will not create disclosure control issues in derivative RRD. Database design will thus support inclusive social science and diminish risks of algorithmic bias arising through uncritical and data driven application of computational models to data of unknown provenance. As with the linear research design, substitution of missing elements will be cognizant of known or estimated heterogeneity of the underlying population, in scale-free geographic context.

This vision is of spatial data infrastructure that can frame any smart data source to known levels of precision. Linked assemblages of Smart Data may then be sliced and diced in secure research environments into RRD products of known provenance and in accordance with the widest range of research user needs. Data governance and licencing will span multiple constituent datasets and data centres – the current LCRs, for example, are supported by more than a dozen inter-linked licencing agreements and some data are syndicated from other centres. Successful service delivery will require multilateral DLAs that can support highly disaggregate spatial linkage of individual level geographic data. From this perspective, innovations in service delivery for RRD will require extended legal and ethical support alongside improved technical capacity.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge funding for this paper from the Economic and Social Research Council via the Consumer Data Research Centre grant, reference [ES/L011840/1] (Longley, Cheshire and Singleton) and the Chinese 111 Program (Longley).

Data availability statement

The data will be made available upon request by contacting the corresponding author.

References

Cheshire, J. 2020. "Code/Space." In *International Encyclopaedia of Human Geography*, edited by A. Kobayashi, 303–308. 2nd ed. Vol. 2. Elsevier. https://doi.org/10.1016/B978-0-08-102295-5.10524-4.

Department for Business, Energy and Industrial Strategy. 2021. *Smart Data Working Group Spring 2021 Report*. assets.publishing.service.gov.uk/media/60c72e058fa8f57ce3773c2d/smart-data-working-group-report-2021.pdf.

Gibbs, H., P. Ballatyne, J. Cheshire, A. Singleton, and M. Green. 2023. "Harnessing Mobility Data to Capture Changing Work from Home Behaviours Between Censuses." *The Geographical Journal* 190 (2). https://doi.org/10.1111/geoj.12555.

Goodchild, M. F. 2015. "Four Thoughts on the Future of GIS." ArcWatch. Accessed December 22, 2023. https://www.esri.com/about/newsroom/arcwatch/four-thoughts-on-the-future-of-gis/.

Goodchild, M. F. 2022. "Elements of an Infrastructure for Big Urban Data." *Urban Informatics* 1 (1): 3. https://doi.org/10.1007/s44212-022-00001-5.

Goodchild, M. F., and P. A. Longley. 1999. "The Future of GIS and Spatial Analysis." In *Geographical Information Systems: Principles, Techniques, Management and Applications New York*, edited by P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, 567–580. Hoboken, NJ: Wiley.

Lansley, G., and J. Cheshire. 2018. "Challenges to Representing the Population from New Forms of Consumer Data." *Geography Compass* 12 (7). https://doi.org/10.1111/gec3.12374.

Lansley, G., P. A. Longley, and W. Li. 2019. "Creating a Linked Consumer Register for Granular Demographic Analysis." *Journal of the Royal Statistical Society Series A: Statistics in Society* 182 (4): 1587–1605. https://doi.org/10.1111/rssa.12476.

Lan, T., J. van Dijk, and P. A. Longley. 2021. "Family Names, City Size Distributions and Residential Differentiation in Great Britain, 1881–1901." *Urban Studies* 59 (10): 2110–2128. https://doi.org/10.1177/00420980211025721.

Longley, P. A., J. van Dijk, and T. Lan. 2021. "The Geography of Inter-Generational Social Mobility in Britain." *Nature Communications* 12 (1): 6050. https://doi.org/10.1038/s41467-021-26185-z.

McGrath-Lone, L., M. A. Jay, R. Blackburn, E. Gordon, A. Zylbersztejn, L. Wijlaars, and R. Gilbert. 2022. "What Makes Administrative Data Research-Ready? A Systematic Review and Thematic Analysis of Published Literature." *International Journal of Population Data Science* 7 (1). https://doi.org/10.23889/ijpds.v7i1.1718.

Moser, C. A., and G. Kalton. 1971. *Survey Methods in Social Investigation*. London: Routledge. https://doi.org/10.4324/9781315241999.

Singleton, A., A. Alexiou, and R. Savani. 2020. "Mapping the Geodemographics of Digital Inequality in Great Britain: An Integration of Machine Learning into Small Area Estimation." *Computers, Environment and Urban Systems* 82:101486. https://doi.org/10.1016/j.compenvurbsys.2020.101486.

van Dijk, J., G. Lansley, and P. A. Longley. 2021. "Using Linked Consumer Registers to Estimate Residential Moves in the United Kingdom." *Journal of the Royal Statistical Society Series A: Statistics in Society* 184 (4): 1452–1474. https://doi.org/10.1111/rssa.12713.

Wyszomierski, J. 2024. "Micro-Geodemographics and Deharmonisation of the 2021/2 Output Area Classification." Ph.D. thesis, London: UCL.