Institution: University of Liverpool

Unit of Assessment: 11 (Computer Science and Informatics)

Title of case study: Geodemographic classifications boost public service provision, encourage sustainable travel, promote the arts, and aid advocacy efforts

Period when the underpinning research was undertaken: 2011 – present

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Prof Alex Singleton	Lecturer - Professor	2010 - present
Dr Dani Arribas-Bel	Lecturer - Senior Lecturer	2015 - present
Prof Rahul Savani	Lecturer - Professor	2009 - present

Period when the claimed impact occurred: August 2013 – December 2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact

The University of Liverpool's Geographic Data Science Lab (GDSL) applied machine learning to create geodemographic classifications from high-dimensional spatial data. These classifications are open, reproducible, and more accurate than their predecessors. They have generated extensive impact in the public, private, and third sectors, including on public health, sustainable travel, and promotion of the arts, for example:

- Hull City Council built a classification using GDSL's methodology to boost services and promote the arts for ~260K residents.
- Transport for London collaborated with GDSL to create a classification that transformed travel policy development and implementation, benefitting those making ~29M daily trips.
- The Office for National Statistics enhanced six impactful surveys using a GDSL classification.
- A West End theatre production used a GDSL classification for the regional rollout of shows.
- US-based company CARTO used GDSL's classifications in their platform (300K+ users).

2. Underpinning research

The University of Liverpool's Geographic Data Science Lab (GDSL), which integrates Computer Science, Data Science, and Geography, developed new techniques to create geodemographic classifications using machine learning **[3.1]**. These classifications provide summary indicators of the demographic and built characteristics of small areas. For illustration, the *Transport Classification of Londoners* partitions London into small areas and categorises each area as primarily comprising one of nine population categories (for example, "Students/Graduates," "Settled Suburbia," "Detached Retirement.")

GDSL extended existing clustering-based approaches for building geodemographic classifications to make their classifications more accurate, open, and reproducible. The main innovations relate to the pre-processing of input data **[3.2, 3.3]**, the stochastic nature of certain clustering algorithms **[3.2, 3.4]**, and a framework for open geographic data science **[3.5]**. GDSL's methodology was developed through the creation of five classifications, described next.

The *2011 Output Area Classification* **[3.2]**, co-produced with UCL and the Office for National Statistics between 2010-16, used 2011 Census data to create a national classification. It resolved major issues with its predecessor, the *2001 Output Area Classification*, by being more accurate and fully reproducible. A greater number of transformation and rate calculation techniques were evaluated (log10; Box-Cox; inverse hyperbolic sine). 10,000 initial partitions with random seeds were run and applied to a model selection technique, providing an optimum partitioning of the output areas in the 2011 Census. Other key classifications include:

- *London Output Area Classification* (2014) **[3.6]**: in a collaboration with UCL and Greater London Authority; the first published instance of a regionally tailored classification.
- *US Geodemographic Classification* (2015) **[3.4]**: created with the University of Colorado in 2015. It is the first open classification of the US Census.
- *Internet User Classification* (2014-15) **[3.3]**: the first classification to aggregate population profiles of Internet engagement.

- *Transport Classification of Londoners* (2016-17): developed with Transport for London.

Two clustering algorithms were used to create Liverpool's classifications: *k*-means (used in all) and Ward's algorithm (*US Classification*). Two specific examples of computing challenges on data pre-processing and stochastic clustering that GDSL addressed were:

1. Small-area estimation (estimation of statistics for small sub-populations): when creating classifications, survey data only detail the characteristics of a small but representative subset of people and areas. But measures for a full extent (like the whole UK) are needed for input into the clustering algorithms. GDSL pioneered the first use of supervised machine learning for small area estimation, where a model is trained to predict measures for all local areas. This technique was applied in the creation of the *Internet User Classification* **[3.3]** to create full coverage of online engagement data and consumption measures.
2. High variability in the quality of output from stochastic clustering methods (such as *k*-means) over different runs on the same data: 10,000 repetitions with different random seeds were performed and goodness of fit was applied to the original data for model selection. However, high dimensionality of input measures (extreme for the *US Classification*) meant prohibitively large running times for existing code bases. Several innovations addressed this challenge, including the use of the parallel multi-core platform H2O **[3.4]**.

3. References to the research

[3.1] – **A. Singleton; D. Arribas-Bel** (2019), 'Geographic Data Science,' *Geographical Analysis*, 1-15. https://doi.org/10.1111/gean.12194

[3.2] – C. Gale; **A. Singleton**; P. Longley (2016), 'Creating the 2011 Area Classification for Output Areas (2011 OAC),' *Journal of Spatial Information Science* 12, 1-27. https://doi.org/10.5311/JOSIS.2016.12.232

[3.3] – **A. Singleton**; A. Alexiou; **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, 1-20. https://doi.org/10.1016/j.compenvurbsys.2020.101486

[3.4] – S. Spielman; **A. Singleton** (2015), 'Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach,' *Annals of the Association of American Geographers* 105:5, 1003-1025. https://doi.org/10.1080/00045608.2015.1052335

[3.5] – **A. Singleton**; S. Spielman, C. Brunsdon (2016), 'Establishing a Framework for Open Geographic Information Science,' *International Journal of Geographical Information Science* 30, 1507-1521. https://doi.org/10.1080/13658816.2015.1137579

[3.6] – **A. Singleton**; P. Longley (2015), 'The Internal Structure of Greater London: A Comparison of National and Regional Geodemographic Models,' *Geo: Geography and Environment* 21, 69-87. https://doi.org/10.1002/geo2.7

4. Details of the impact

The University of Liverpool's Geographic Data Science Lab (GDSL) drove a step change in the use of geodemographic classifications by government and have improved the efficacy of the targeting of resources in the public, private, and third sectors. Effective policymaking and commercial strategy require a full understanding of small areas and their populations, which only high-quality classifications provide. Suboptimal classifications lead to poor decision making, which wastes budgets, limits profits, and has negative implications for individuals' life chances.

GDSL's outputs are the only open and reproducible residential classifications in the UK, in a landscape of expensive commercial alternatives. Moreover, commercial classifications typically provide limited methodological detail beyond their surface-level representations. By contrast, the classifications developed by GDSL are based on a transparent and reusable methodology. They have created significant impact in the public, private, and third sectors both directly and as a reproducible framework for creating bespoke classifications. GDSL's impact will be illustrated through the following beneficiaries: local and regional authorities (Hull City Council and Transport for London); the Office for National Statistics and users; a West End theatre production; and a leading US-based location intelligence platform and its users.

4.1 Improving Policy Implementation and Engagement with the Arts in Hull. Local authorities used GDSL's methodology to create derivative models. To illustrate, in 2013, Hull City Council, which serves 260K residents, "*ruled out commercial classifications … due both to their high licensing and maintenance costs… and the loss of accuracy / reliability when mapping national samples to local communities*." For Hull's City Council's Customer Insight Analyst, GDSL's "*free-to-use, open and reproducible methodology of the 2011 Output Area Classification*" was "*absolutely crucial*" to create Hull City Council's own model and "*ensured that* [it] *was robust and based on a sound and tested process*" **[5.1]**. Hull City Council avoided commercial alternatives with costs ranging from £10-100K **[5.2a]**.

<u>Improving Service Delivery:</u> As one example, in 2015, Hull City Council used their classification to improve service delivery in its North Carr Ward. The classification showed that residents preferred co-located, in-person services. Hull City Council acted on this to better meet residents' needs cost effectively. For example, as residents often used North Carr's new health centre, council staff were relocated there. Co-location provided residents with a holistic service and ended a 15-year lease at approximately £70K per year. In total, Hull City Council have saved approximately £164K per year and residents of North Carr gained improved service provision **[5.1; 5.2b]**. The classification's success has "*embedded bespoke customer insight into* [Hull City Council's] *decision making as a default approach*" **[5.1]**.

<u>Promoting the Arts:</u> After a successful bid for UK City of Culture (2017), Hull City Council created an arts-focused classification that identified where underrepresented audiences lived. Consequently, three specific wards were targeted to boost participation. Arts Council-funded groups like *Back to Ours* established Community Hubs and "*shifted emphasis away from traditional centres (museums, galleries)*" by bringing events to these low engagement wards **[5.1]**. The success of this approach is evidenced by a 2018 study: the percentage of Hull residents who visited a cultural venue increased from 36% in 2013 to 95% in 2017; residents' confidence in attending events increased from 43% in 2015 to 52% in 2018; and 70% of residents felt that the City of Culture positively impacted them **[5.2c]**.

Hull City Council have since advised other local authorities (like Leicester City Council) on the creation of their own models **[5.2a]**.

4.2 Creating A New Tool to Increase Sustainable Travel in London. In 2015-16, Transport for London (TfL) collaborated with GDSL's Singleton to create the *Transport Classification of Londoners* (TCoL), which was a "*key improvement on TfL's historic dependence on non-London specific classifications*." TfL's Director of City Planning confirmed: *"We approached Singleton for his expert guidance in creating TCoL. TCoL would not be possible without LOAC* [the London Output Area Classification] *… Singleton's LOAC research and data science guidance in facilitating the development of TCoL has proved invaluable for TfL*" **[5.3]**. TCoL has impacted 1) evidence-based policy development and 2) policy implementation.

<u>Policy Development:</u> TCoL was made "*out of a pressing need to understand transport-based behaviours and motivations to help* [TfL] *plan more effectively for London."* TCoL was *"crucial in mapping small-area level differences in Londoners' propensity to adopt active travel"* which identified key targets for policy intervention. In 2016-17, TCoL underpinned five reports that examined walking and cycling behaviour and strategies for increasing modal shift, along with a new cycling model. The new model accurately reflects actual route choice decisions made by cyclists "*improving our ability to model cycle demand and to present the case for cycling investment*" compared to previous models, which "*rarely capture cycling trends accurately*" **[5.3]**.

These resources were a significant component of the evidence base of the *Mayor's Transport Strategy* (2017-2041). The strategy's central aim is for 80% of all trips to be on foot, by cycle, or using public transport by 2041: *"TCoL has been key in highlighting the policy mix vital to achieve the local contributions towards active travel necessary to enable the London-wide aim"* **[5.3]**.

<u>Policy Implementation:</u> TCoL is used to implement policy in two ways. First, since 2017, investments costing £200K+ on TFL's road network were required to undergo a "Healthy Streets

Check" to ensure they met TfL's targets for improving sustainable travel: *"TCoL is an essential component of this* [check]*."* Since 2019, the check has been applied to all investments, irrespective of size. The check has been successful: in 2019-20, TfL reported that checked investments resulted in an average increase of 15% in TfL's health metric **[5.3]**.

Second, TCoL underpins the implementation of the "Liveable Neighbourhoods Scheme" (since 2017). This scheme offers grants of up to £10M for borough-level projects. One use of TCoL is to help assess proposed projects according to the expected improvement delivered. The most recent round of this scheme corresponded to £53M of TfL funding. An example of a funded project is the development in Lambeth of new pedestrian crossings, improved cycle routes and footways, low traffic neighbourhoods, and a new cycle-only route (£9.5M from TfL) **[5.3]**.

TfL's Director of City Planning summarised: TCoL *"continues to be a valuable framework in the formulation of research & policy; improving integrated land use and transport planning; tailoring and targeting schemes and investment; and underpinning the promotion of active travel"* **[5.3]**.

4.3 A More Accurate Classification for the Office of National Statistics (ONS). The 2011 Output Area Classification **[3.2]** was adopted by the ONS as its official classification **[5.4a]**. The improved accuracy of the 2011 Output Area Classification had wide-ranging impacts via at least six ONS national surveys that used the classification in their non-response models **[5.4b; 5.4c]**. Non-response of selected participants is a critical issue as it can severely bias survey estimates; non-response models are used to mitigate this.

For illustration, the "Dental Health Survey of Children and Young People" (2013) was carried out by the ONS on behalf of NHS Digital. The survey helps a wide range of users, including policymakers and the NHS, to understand how the dental health of each generation of children across the UK is changing. It serves as a crucial component of the evidence base that is used for policymaking over the ten-year period 2013-23. In the context of tooth decay, the survey underpinned guidance to local authorities from Public Health England (PHE) for setting up toothbrushing schemes **[5.4d]**. A June 2018 PHE report noted that 51% of English local authorities had adopted such schemes **[5.4e]**.

4.4 Transforming Marketing Strategies for Award-Winning Theatre in England. GDSL's *Internet User Classification* (IUC) transformed the marketing strategy of the West End theatre production The Book of Mormon. In 2017, the marketing team was challenged to form a digital advertising strategy for the regional rollout of the shows across England (for example, Manchester, Newcastle, and Sunderland). The resulting strategy was based on the Internet User Classification **[3.3]**: aggregate population profiles of Internet engagement were used to target specific customer segments. The Marketing Director affirmed that this constituted the *"first instance of this approach being used in live entertainment marketing."* The IUC was commended for being "*more detailed than any other tool in the arts marketer's toolbox over the last fifteen years of our company's operation*." **[5.5]**

<u>Cost Savings and Revenue Growth:</u> The regional rollout *"prospered as a result… breaking box office records, playing to 96% capacity and grossing close to £20M from 2019 to lockdown in 2020."* The IUC delivered significant financial benefit, increased the show's reach, and helped to engage new audiences: *"51% of customers to The Book of Mormon have been first time visitors to the venues and 57% under the age of 35"* **[5.5]**.

The Marketing Director summarised: *"The application of research by the Geographic Data Science Lab has offered us an unparalleled understanding of potential audiences for our regional expansion and has had a direct impact on our ability to market these shows effectively. Our use of the Internet User Classification improved both cost-effectiveness (directly saving the production £60,000 relative to a commercial offering) and was worth multiple hundreds of thousands of pounds in terms of increased revenue… The success we experienced in marketing The Book of Mormon meant that the IUC was applied to four other high-profile projects; and is now integrated into the online booking system and intelligence platform of one of our key business partners, TickX"* **[5.5]**.

4.5 Powering GIS Software and Improving User Confidence in the Private Sector. CARTO is a US-based company in spatial analysis and Geographic Information Systems (GIS) development. Their GIS platform has more than 300,000 users. In 2016, GDSL collaborated with CARTO to integrate its *US Geodemographic Classification* into their GIS platform. The Head of Data Science found Singleton's article **[3.5]** on the classification "*invaluable in understanding the development of geographic segmentation*". The classification was chosen instead of commercial alternatives, due to "*both the transparency of the methods and the open source nature of the code and the data"* **[5.6]**.

Within CARTO's platform, the US Geodemographic Classification is widely used by the public, private, and third sectors. Here, we illustrate a third sector example: The US-based Anti-Eviction Mapping Project (AEMP) uses the classification to understand the effects of gentrification and support those affected by eviction. The AEMP primarily operate in San Francisco, Los Angeles, and New York City. It guides advocacy by mapping patterns of housing discrimination for use by its partners, such as the San Francisco-based "Tenants Together". To this end, the AEMP enriches eviction data with extensive demographic and census information from the US Geodemographic Classification. CARTO confirmed: "*It is a great dataset for enrichment because it distils much of the complexity of the demographic variables down into human-readable descriptions of the underlying demographic information at different locations"* **[5.7].** The AEMP has been promoted by major NGOs and media platforms, including the World Economic Forum, the Guardian, and the Los Angeles Times.

4.6 Further Reach. From October 2015 to December 2020, the 2011 Output Area Classification, London Output Area Classification, and the Internet User Classification were downloaded 23,220 times. From May 2017, 507 of these were by local authorities **[5.8].**

5. Sources to corroborate the impact

[5.1] Testimonial from Hull City Council Customer Insight Analyst (2020), supporting impact on service provision and promotion of the arts, and Hull residents.
[5.2] Evidence of Hull City Council impact:
a) extract: 'Developing a Customer Classification Tool' (2013), supporting cost savings and Leicester City Council's development of a bespoke classification.
b) extract: 'Customer Segmentation Toolkit' (2016), illustrating North Carr case study.
c) extract: 'Cultural Transformations: The Impact of UK City of Culture 2017' (2018), support uplift in Hull residents' engagement with the arts.
[5.3] Testimonial from Transport for London Director of City Planning (2020), supporting impact on TfL's sustainable travel policies and London residents.
[5.4] Evidence of impact on Office for National Statistics' surveying, associated policymakers/service providers, and subsequent beneficiaries.
a) '2011 Residential-Based Area Classifications', showing ONS adoption of 2011 OAC.
b) ONS surveys using 2011 OAC.
c) extract: 'Children's Dental Health Survey Technical Report' (2013), showing 2011 OAC use.
d) extract: Local Authorities Improving Oral Health (2014), supporting use of 2011 OAC-underpinned Dental Health Survey as evidence for policymaking/service provision.
e) extract: 'Local Authority Variation in the Oral Health of Five-Year-Olds' (2018), evidencing uptake of schemes underpinned by Dental Health Survey.
[5.5] Testimonial from Mormon Trail Ltd. Marketing Director (2020), supporting impact on The Book of Mormon's entrepreneurial activity, new audiences, and profits.
[5.6] Testimonial from CARTO Head of Data Science (2018), supporting impact on CARTO's GIS platform and clients.
[5.7] extract: 'Anti-Eviction Mapping Project and CARTO Expose Realities of Gentrification and Evictions in San Francisco' (2016), supporting impact on Anti-Eviction Mapping Project.
[5.8] – Web Usage Statistics Spreadsheet, indicating further reach of GDSL research.