Update on methodology for the Digital domain of the Hard to Count index for the 2021 Census

Ercilia Dini, MDR/Methodology, Office for National Statistics
May 2019

Summary

The 2021 Census, unlike previous censuses, will undertake the collection using an online questionnaire as the primary response mode. This change in the basic collection mode means that the non-response patterns observed in certain population groups in the 2001 and 2011 censuses may be different in the 2021 Census.

In addition to 'willingness' – as an innate propensity to respond to the census there will be an additional component which is the ability to respond, driven by access and use of digital technology. People who are 'digitally excluded' will require digital assistance or a way to respond to the census that may not be via the primary mode.

The ONS is conducting research to develop a Hard-to-Count (HtC) index to:

a) Identify geographical Lower Super Output Areas (LSOAs) in England and Wales at risk of census non-response
b) Identify LSOAs in England and Wales according to their 'digital ability'

These will be used as a tool in the 2021 Census to support pre-planning of field follow-up, areas where to send paper questionnaires as first approach to households and where assisted digital centres should be situated. It will also be used as a stratification variable in the sample design of the 2021 Census Coverage Survey and as a covariate in the census estimation and adjustment methods. The HtC index will be key to ONS achieving high quality census estimates.

The HtC index is composed of two domains: the *willingness to self-respond* domain and the *digital* domain. Each of these domains may be used as an index per se.
The willingness domain is constructed using an area level (Lower Super Output) model that predicts non-response by day 10 after census day. The covariates used to build the model parameters are from previous census and updateable administrative data sources.

The initial version of the digital domain was built as an indicator (at LSOA level of geography) to measure infra-structures available which enables households to respond to an online census. New administrative data from the Driver and Vehicle Licensing Agency (DVLA) became available to ONS in November 2018. The data give information on people's use of internet to conduct online transactions to send personal information to a government website. It can potentially be used as a proxy to the likelihood of LSOAs according to their 'digital ability to respond to an online census.

This paper presents the methodology used to develop the digital domain of the HtC index using this new available administrative data, the results obtained and recommendations for its use.

CRAG members are invited to comment on the methodology and recommendations given on this paper.

Table of contents **Page**

1. Introduction

The full report with the methodology developed to build the two domains of the HtC index for the 2021 Census was presented to the External Assurance Panel (EAP) in October 2018 (Dini, 2018).

At the time, the digital domain of the HtC index measured one of the infra-structures available which enables households to respond to an online census. It was measured (at LSOA level) by the percentage of households with access to the internet, estimated using the number of fixed line broadband connections to premises (households and small businesses) (Ofcom, 2016) and number of occupied households (Census 2011). The indicator was the best proxy to support the census planning for areas where digital assistance may be required and where census forms are to be sent as first option. More detail on how the indicator was built is available in Annex A of the full report (Dini, 2018).

Responding to an online census will be a specific process where people will need to use the internet and more specifically, make use of the internet to interact with a government website to submit their personal information. In the report presented to the EAP we suggested future use of fit for purpose data that would give information on peoples' use of internet to interact with government websites. Fit for purpose data from the Driver and Vehicle Licensing Agency (DVLA) on number of online transactions to apply for or renew a driving licence became available to ONS in November 2018.

This paper presents the methodology used to develop the digital domain of the HtC index using the new available administrative data, the results obtained and recommendations for its use.

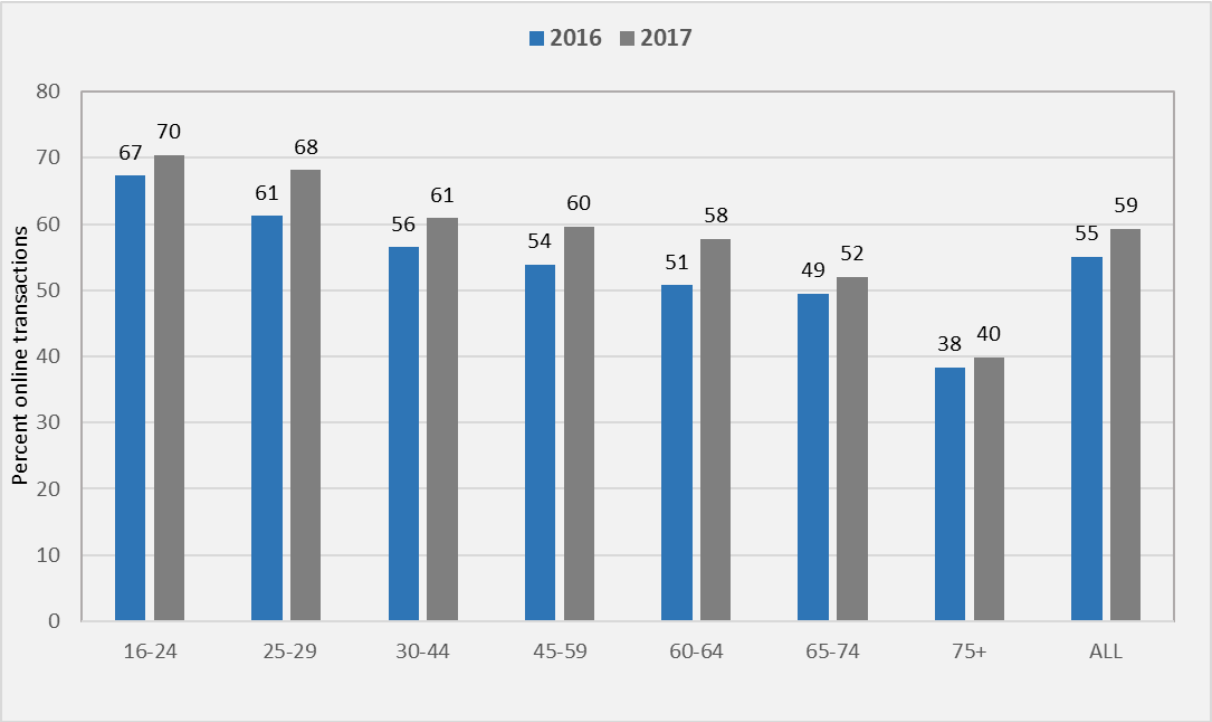
Section 2 of this paper presents the background of DVLA data provided to ONS and an analysis of the data for years 2016 and 2017. Section 3 presents the methodology used to build the 'new' digital domain of the HtC index and the method used to assign HtC categories into the 'new' digital domain.
Section 4 presents the results obtained for the digital domain of the HtC index and compares them with the previous digital domain.
Section 5 presents results of online returns in the 2017 Census test as a proportion of all returns (online and paper) from addresses that were sent a Unique Access Code (UAC) as the first approach. The results are post stratified by the 'new' digital domain of the HtC index.

2. Background on DVLA data

The DVLA data provided to ONS contains England and Wales (E&W) aggregated (LSOA level) number of transactions to apply for or to renew a driving licence. The transactions are for people aged 16 and over by single year of age (SYOA) and sex and split by mode of transaction (paper or online). DVLA data is provided to ONS yearly and the first year of data received was back to 2016. The analysis presented in this section refers to transactions conducted in 2016 and 2017.

In 2016 there were 7.6 million transactions (16 per cent of the population in E&W aged 16 and over) and 55 per cent of the transactions were using online mode. In 2017 there were 7.9 million transactions (17 per cent of the population in E&W aged 16 and over) and 59 per cent of the transactions were via online mode. Approximately one per cent of the transactions in 2016 were missing age/sex information; this percentage was smaller in 2017, at 0.5 per cent. Figure 1 presents the percentage of online transactions by age group in 2016 and 2017.

Figure 1: Percentage of online transactions by age group in 2016 and 2017. England and Wales.



DVLA data has partial population coverage. It includes only people aged 16 and over who apply/renew a driving licence. It does not include people who never applied/renewed a driving licence. In addition, there are restrictions for online application; you cannot apply online if your name or title has changed and you also need to hold a valid UK passport (Gov.UK, 2019).

These restrictions may cause bias towards for example women who changed their title after marriage and to foreigners who, despite residing in the UK, do not hold a UK passport. Therefore, caution is needed when using the data as a proxy for areas' digital ability to respond to a census.

In both years, 2016 and 2017, the percentage of transactions decreased with age. The age group 75 and over had the lowest percentage of online transactions in both years. This was expected as there is evidence that younger age groups are more likely to conduct online transactions than older age groups. In addition, interacting with a government website and applying online for a driving licence may be seen as bringing an immediate benefit to younger age groups. As for answering to an online census, the benefit may not be seen as immediate and/or not well understood.

3. Method

3.1. Model for the 'new' digital domain of the HtC index

An area level (LSOA) model was used to predict areas according to their digital 'ability'. The dependent variable of the model was the 2017 DVLA percentage of online transactions. All DVLA online transactions within an LSOA were considered equivalent, all experiencing the same conditions (area level covariates) and displaying the same probability *p* of conducting online transaction to apply for or to renew a driving licence.
The covariates in the model were the 2017 percentage of mid-year population estimates by age group at LSOA level (ONS, 2018), broadband access to internet in 2017 at LSOA level (Ofcom, 2017) and English regions and Wales.

The predicted percentage of online transactions at LSOA level given by the model were then used as a proxy predictor for the 2021 Census areas' digital ability.

The DVLA percentage of online transactions had a normal distribution and varied nonlinearly with some of the covariates chosen. A generalised model with a normal distribution and a log function was used

to model the probability *p* of conducting online transaction to apply for or to renew a driving licence. X describes the explanatory or covariates and the β's are the coefficients used to convert the prevalence of the factor into an effect upon the response variable (propensity of areas' digital ability).

The model is given by:

$\log p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

where β_0 is the intercept and $\beta_1, \beta_{2...}\beta_n$ are the coefficients used to convert the prevalence of the covariates into an effect upon the outcome variable.

The predicted 2021 Census areas' digital ability given by the model should not be taken as the actual online responses to the 2021 Census. The predicted values are to be used only as the likelihood of LSOAs (a rank of) areas according to their digital ability to respond to an online census.

Information on the model parameters is given in Annex A.

3.2. Assigning HtC categories to the 'new' digital domain

The predicted percentage of online transactions at LSOA level given by the model was ranked from the highest to the lowest.

The ranked values were then split into 5 categories, the highest – category 5 - being the least digitally able areas. The percentages of LSOAs used in the split were: 40% in HtC1, 40% in HtC2, 10% in HtC3, 8% in HtC4 and 2% in HtC5.

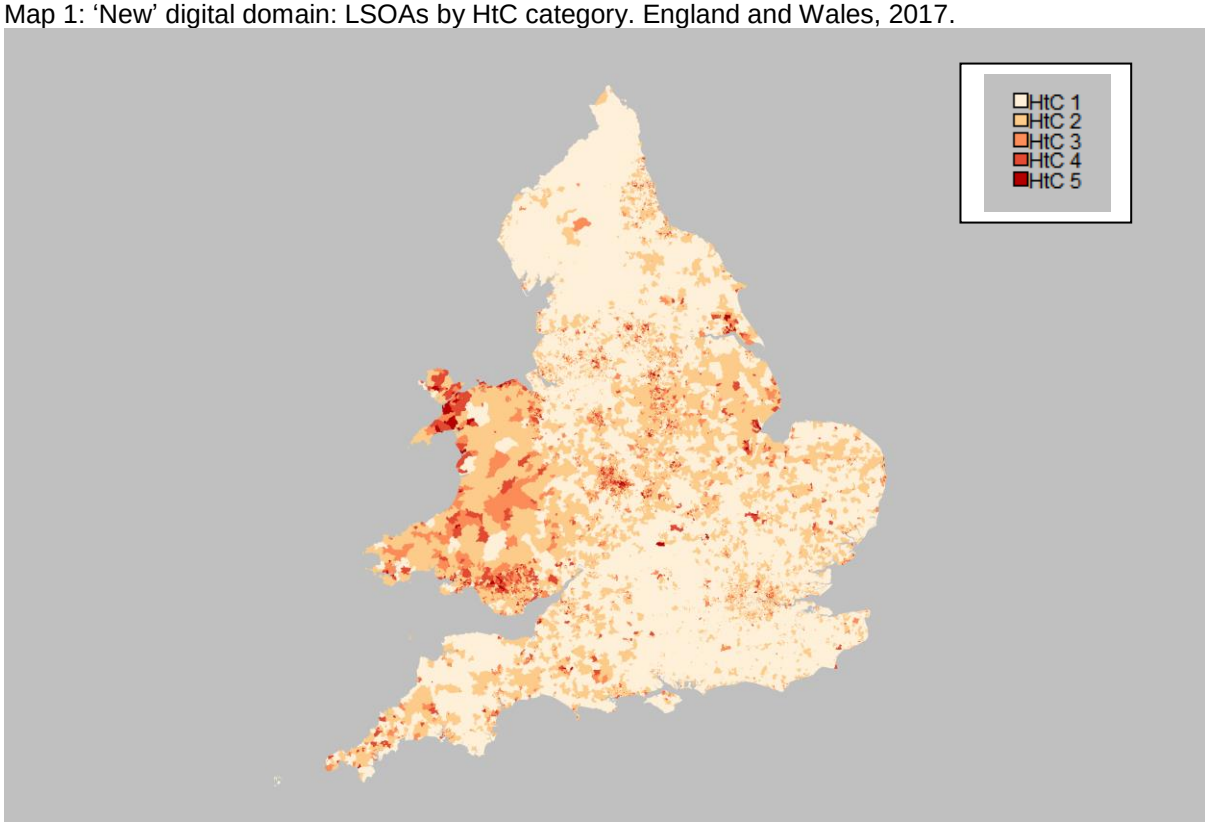
4. Results

4.1. 'New' digital domain of the HtC index

Map 1 shows the distribution of LSOAs in England and Wales according to the 5 categories of the 'new' digital domain of the HtC index.

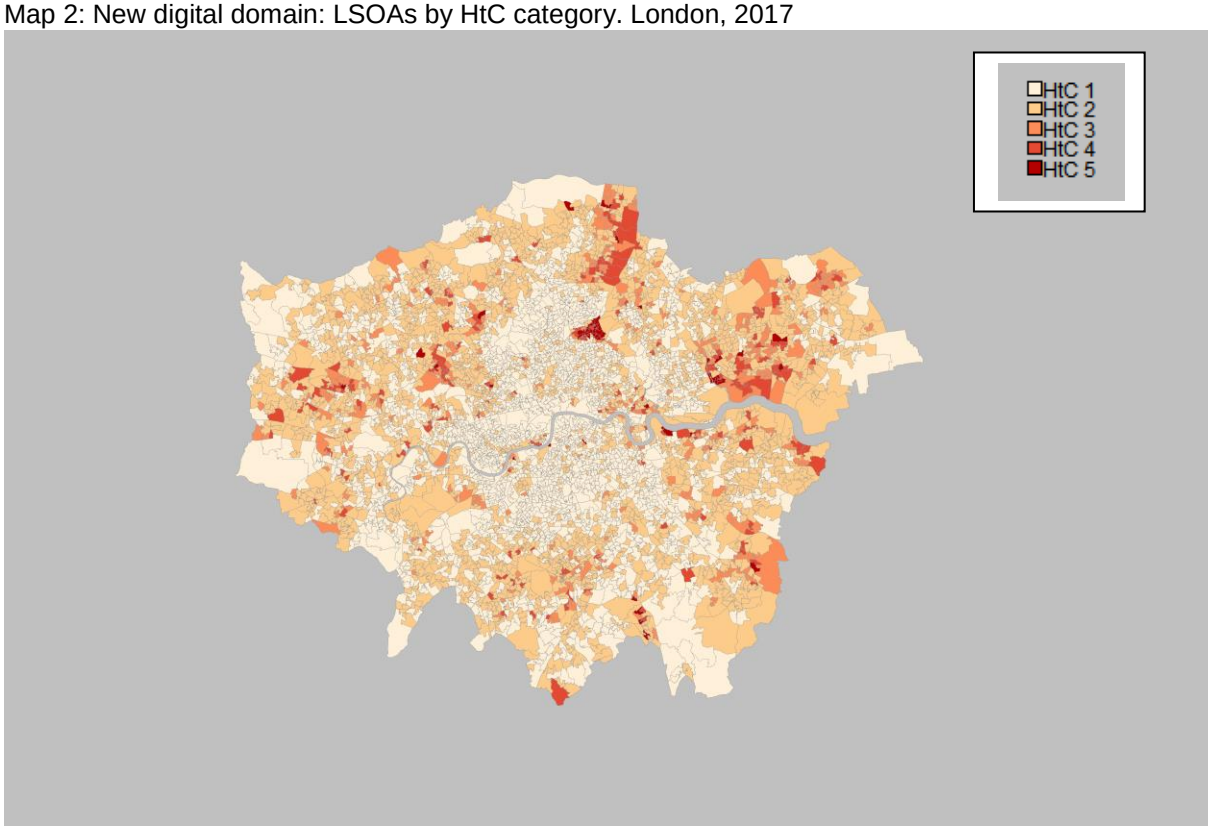
Map 2 shows the distribution of LSOAs in Inner and Outer London according to the 5 categories of the 'new' digital domain of the HtC index.

Map 1: ‘New’ digital domain: LSOAs by HtC category. England and Wales, 2017.



Note: HtC 5 = least digitally able areas
Map: Contains Ordnance Survey data © Crown copyright 2018
Source: Office for National Statistics licensed under the Open Government Licence v.3.0

Map 2: New digital domain: LSOAs by HtC category. London, 2017



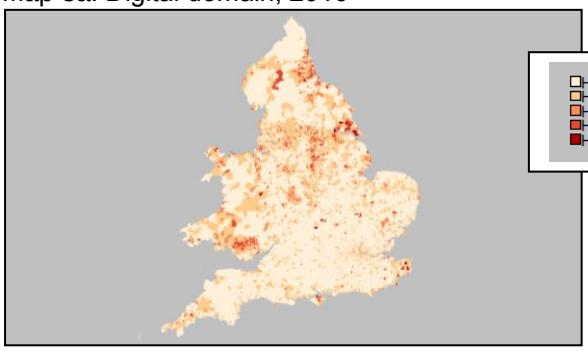
Map: Contains Ordnance Survey data © Crown copyright 2018
Source: Office for National Statistics licensed under the Open Government Licence v.3.0

4.2. Comparison between the previous and the new' digital domain of the HtC index

Maps 3a and 3b show the distribution of LSOAs in England and Wales according to the 5 categories of the digital domain of the HtC index as given by the previous indicator (percentage of households with access to the internet, estimated using the number of fixed line broadband connections to premises (Ofcom, 2016) and the 'new' digital domain of the HtC index given by modelling DVLA data, respectively.

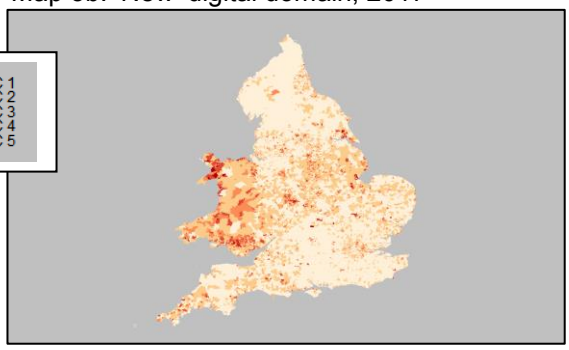
Maps 4a and 4b show the distribution of LSOAs in Inner and Outer London according to the 5 categories of the digital domain of the HtC index as given by the previous indicator and as per modelling DVLA data, respectively.

Map 3a: Digital domain, 2016



Note: HtC 5 = areas with the lowest access to internet

Map 3b: 'New' digital domain, 2017



Note: HtC 5 = least digitally able areas

Map 4a: Digital domain. London, 2016.

Note: HtC 5 = areas with the lowest access to internet

Map 4b: 'New' digital domain, 2017.

Note: HtC 5 = least digitally able areas

5. 2017 Census test online returns and the 'new' digital domain

This section presents analysis using data from the designed experiments within the 2017 Census Test.

The sample in Component 2 Part 1 was designed specifically to test the impact of sending a paper questionnaire versus a letter containing a Unique Access Code (UAC) as the first approach to households. The total sample in Component 2 Part 1 was approximately 60,000 addresses in England and Wales. Half of the sampled addresses were sent a paper questionnaire first and the other half received a UAC. Households in addresses that were sent a UAC first could request a paper questionnaire, but this was intentionally not made obvious. Households in addresses that were initially sent a paper questionnaire were also given a UAC, on the first page of the paper questionnaire, so it was clear that they could respond online if they chose to do so. Three reminders were sent to non-responding addresses in both groups, the first, 3 weeks after the initial questionnaire/UAC letter had been sent. All the reminders in both groups were via a letter containing the UAC only.

The 2017 Census Test sample was stratified using 9 categories constructed at LSOA level by combining 3 categories on likelihood to self-respond (based on return rate by day 10 in 2011 Census) and 3 digital categories using information on broadband internet availability (Ofcom, 2014). The cut points for the 3 categories for each domain were constructed using the Dalenius-Hodges method of stratification (Dalenius and Hodges, 1959) independently on each margin. The categories from each domain were then combined. For more information on the 2017 Census Test design and sample see ONS (2017).
Previous analysis of the test (Corps et al, 2017) showed that households were more likely to respond if they were sent a paper questionnaire first. This pattern of response was seen across the 9 categories of the sample stratification. The analysis also showed that the digital domain of the sample stratification used in the 2017 Census Test was a good predictor of online response and particularly by age group.

Further analysis was conducted to investigate the percentage of online returns as a proportion of all returns (online and paper) from addresses that were sent a UAC as first approach. The returns are post post-stratified by the 'new' digital domain of the HtC index (Table 1).
The rationale for using the post stratification analysis was to check if levels of online returns in the 2017 Census test were consistent with the HtC categories in the 'new' digital domain; that is the higher the HtC category the lower the online return rate.

Table 1: Online and paper returns and percentage of online returns from all returns from addresses that were sent a UAC as first approach, by HtC categories given by the 'new' digital domain.

HtC categories	online returns	paper returns	% online returns
1	2,753	252	91.6
2	2,614	381	87.3
3	730	125	85.4
4	626	104	85.8
5	117	27	81.3

This analysis used post stratification and the sample of the 2017 Census test had not been designed specifically to validate the 5 HtC strata of the 'new' digital domain. Therefore, we need caution in interpreting these results.

In addition, we need to consider that the census test was a voluntary survey; overall online returns were very low and likely to be from the most willing to participate. This consequently leads to response bias. The 2021 Census will be mandatory, there will be media communication to gain people's compliance with the online census. It is expected that household's online responses will be higher.

Despite all these constraints with the use of the 2017 Census test data and while the differences observed by some of the HtC categories may be small, the results are reassuring. The results of online returns as a proportion of all returns from addresses that were sent a UAC as first approach suggest that the 'new' digital domain is a reasonable predictor of areas' digital ability. With the exception of categories 3 and 4 of the HtC that had similar level of returns, the other HtC categories showed the expected monotonically decreasing pattern in online returns.

6. Recommendation

Based on the results obtained by the model and the post stratification analysis conducted using the 2017 Census test online responses by the HtC categories we make the following recommendations:

- We recommend the use of the predicted percentages of DVLA online transactions given by the model to attribute a 'digital ability' score (HtC categories) to LSOAs in England and Wales.
- The predicted areas' digital ability given by the model should not be taken as the actual online responses to the 2021 Census. The predicted values are to be used only as the likelihood of LSOAs (a rank of) areas according to their 'digital ability' to respond to an online census.
- The predicted values should be used for planning areas where digital assistance is likely to be needed.

- The predicted values should be used in conjunction with the 'Willingness' domain of the HtC index for planning areas where paper questionnaires are to be sent as first approach to households.
- If more administrative data (e.g. electoral register data at LSOA level) becomes available, they could be tested and included if proved to improve the fit of the model.

7. References

Corps D, Fraser O and Meirinhos V (ONS, 2017). Test Analysis – Analysis of online and paper respondents. ONS Internal report. Available at: https://share.sp.ons.statistics.gov.uk/sites/cen/csod/CSOD_Stats_Design/Statistical_Design/2017%20Test/Reports/CRAG_paper_Nov17.docx

Dalenius T and Hodges J L Jr (1959). Minimum variance stratification. Journal of the American Statistical Association, 54, 88–101.

Dini E (2018). Hard to Count index for the 2021 Census. Available at: https://share.sp.ons.statistics.gov.uk/sites/MTH/Cen/2021/Sample_Design_and_Estimation/Hard_to_Count/HtC_index_report_v5_for_External_Assurance_Panel.docx

GOV.UK (2019). Renew your driving licence. Available at: https://www.gov.uk/renew-driving-licence

Office for National Statistics (2017). ONS Census Transformation Programme. 2017 Census Test. Available at: https://www.ons.gov.uk/census/censustransformationprogramme

Office for National Statistics (2018). Annual Small Area Population Estimates: mid 2017. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualsmallareapopulationestimates/mid2017

Office of Communications (Ofcom) (2014). Infrastructure Report 2014. Available at: https://www.ofcom.org.uk/research-and-data/multi-sector-research/infrastructure-research/infrastructure-2014

Office of Communications (Ofcom, 2016). Infrastructure report 2016: downloads. Available at: https://www.ofcom.org.uk/research-and-data/multi-sector-research/infrastructure-research/connected-nations-2016/downloads

Office of Communications (Ofcom, 2017). Connected nations 2017: Main report. Available at: https://www.ofcom.org.uk/research-and-data/multi-sector-research/infrastructure-research/connected-nations-2017/concise-summary

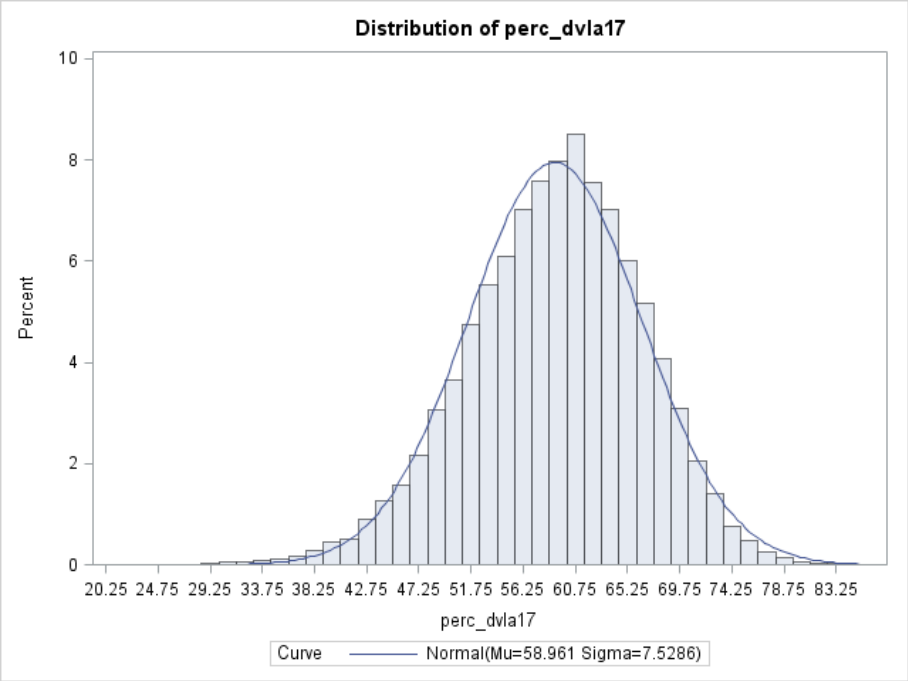
8. Annex A

Annex A provides information on outcome variable and the parameters of the model.

Outcome variable

The dependent variable of the model was the 2017 DVLA percentage of online transactions. Figure 1 shows the distribution of LSOAs by the dependent variable. The distribution is normal with a mean of 59 per cent and a standard deviation of 7.5 per cent.

Figure A1: Distribution of LSOAs by DVLA online transactions. England and Wales, 2017.



Model covariates

The covariates (explanatory variables) and data source used in the model were:

- Percentage of people aged 16-29, 30-44, 45-64 and 65 and over (ONS Mid-year population estimates at LSOA level of geography)
- Percentage of households with access to internet (Ofcom 2017 and 2011 Census number of occupied households). The method used to estimate this percentage is available in Annex A of the full report presented to the EAP in October 2018. (https://share.sp.ons.statistics.gov.uk/sites/MTH/Cen/2021/Sample_Design_and_Estimation/Hard_to_Count/HtC_index_report_v5_for_External_Assurance_Panel.docx)

- English regions and Wales

The covariates were chosen to answer how much area level age groups, access to broadband internet and region/country affect DVLA percentage of online transactions.

Model output

Model parameters and fit

Table A1 shows the parameter estimates obtained by the model and Figure A2 shows the correlation between observed and predicted values.
The model explained 33 per cent of the variance (deviance). In other words, the model 'explained' 33 per cent of the variation in the total LSOAs' DVLA online transactions observed in 2017. It suggests a moderate fit to predict areas digital ability.

Table A1: Model parameter estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.559	0.009	-1.576	-1.542	32520	<.0001
perc_16_29		1	0.006	0.000	0.006	0.006	4487	<.0001
perc_30_44		1	0.008	0.000	0.008	0.008	3217	<.0001
perc_45_64		1	0.012	0.000	0.011	0.012	7714	<.0001
perc_65p		1	0.004	0.000	0.004	0.004	1944	<.0001
perc_ofcom17		1	0.005	0.000	0.005	0.005	11480	<.0001
region_n	EE	1	-0.042	0.002	-0.045	-0.038	437	<.0001
region_n	EM	1	-0.064	0.002	-0.068	-0.059	906	<.0001
region_n	NE	1	-0.010	0.002	-0.015	-0.006	19	<.0001
region_n	NW	1	-0.009	0.002	-0.013	-0.005	21	<.0001
region_n	OL	1	-0.058	0.002	-0.062	-0.054	858	<.0001
region_n	SE	1	0.001	0.002	-0.003	0.004	0	0.8
region_n	SW	1	-0.052	0.002	-0.056	-0.048	634	<.0001
region_n	W	1	-0.082	0.002	-0.087	-0.078	1234	<.0001
region_n	WM	1	-0.055	0.002	-0.059	-0.051	739	<.0001
region_n	YH	1	-0.015	0.002	-0.019	-0.011	54	<.0001
region_n	IL	0	0.000	0.000	0.000	0.000	.	.
Scale		1	0.374	0.001	0.373	0.376		

Figure A2: Correlation between observed and predicted values.

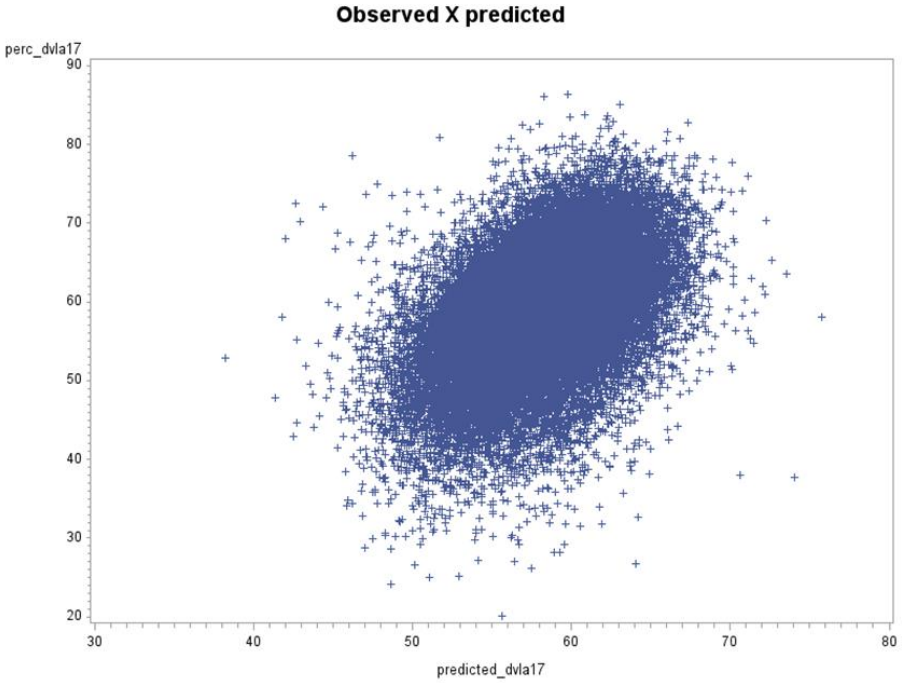


Figure A3 shows the distribution of residuals or unexplained part of the variance of the model at the LSOA level. The residual is given by the observed minus predicted values. The histogram (Figure A3) shows the distribution of the residuals with a normal distribution overlaid. The mean distance of observed to predicted value for an LSOA is +0.7 and a mean standard deviation of 6.5 per cent. This indicates that the model is slightly pessimistic, that is the observed value was higher than the predicted value. The slightly large mean standard deviation suggests that the model is not a perfect fit for the data. Ideally, we should have more covariates in the model to improve its fit.

Figure A3: Distribution of residuals (observed minus predicted values).

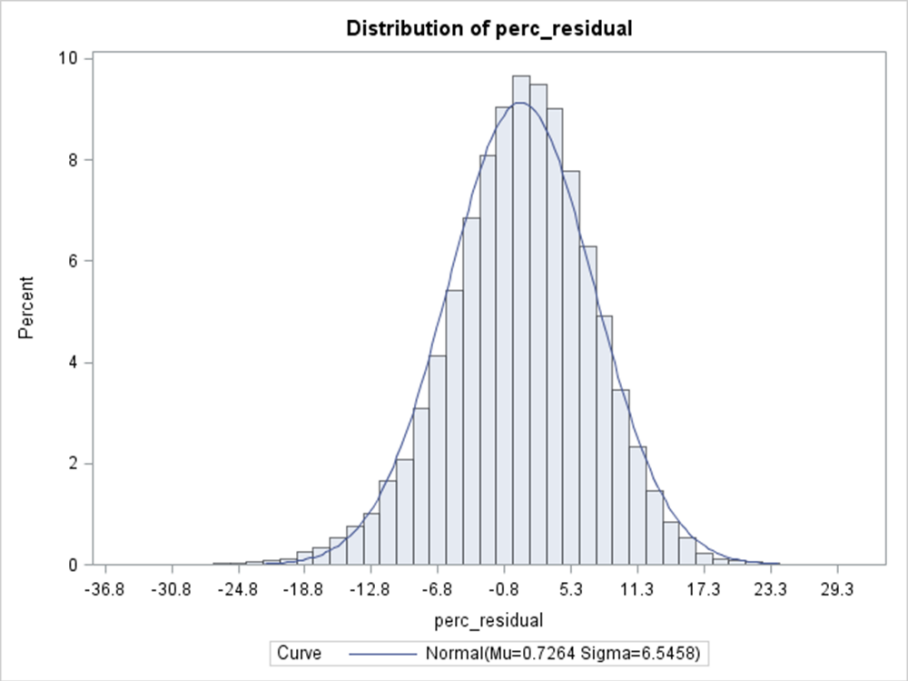


Figure A4: Distribution of LSOAs by predicted values. England and Wales.

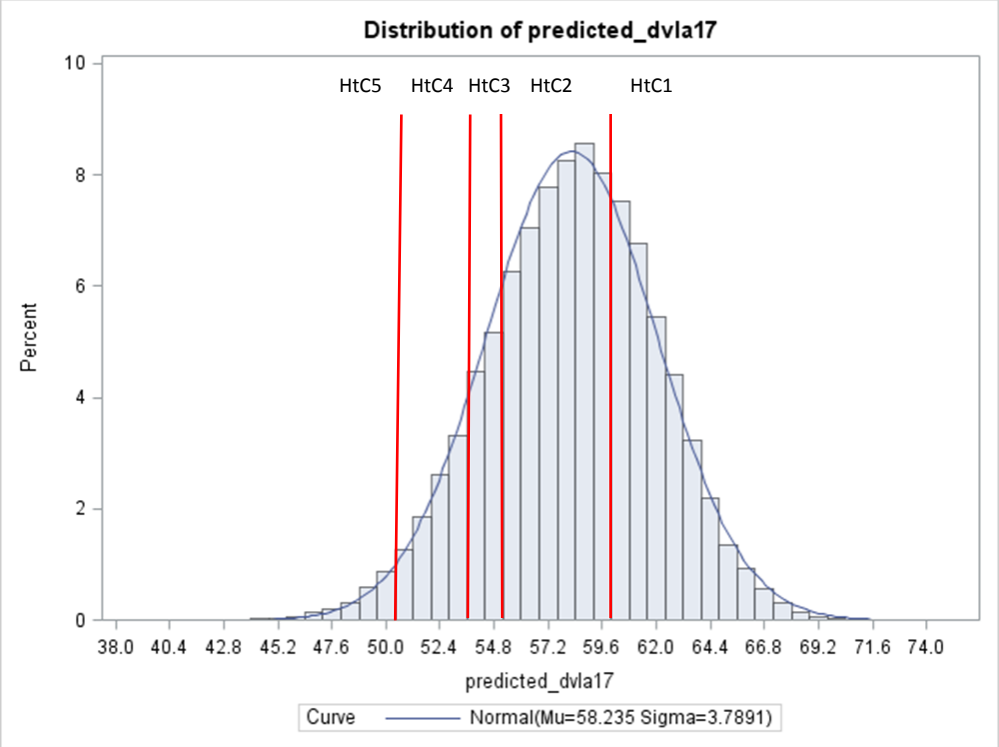


Figure A4 shows the distribution of LSOAs by predicted percentage of DVLA online transactions. The red bars in Figure A4 are a rough illustration where, in the predicted rates, the 40%, 40%, 10%, 8% and 2% cut off points fall in each of the 5 categories of the HtC. There is overlap as for the same predicted percentages falling for example in two contiguous categories of the HtC. Despite this constraint the cut off points show that the split used seems to work well. The analysis using the 2017 Census test showed that the 'new' digital domain performed reasonably well as a discriminator for areas' digital ability.