theark
DATA QUALITY SPECIALISTS

COVID-19 dataset

Powerful COVID-19 data to improve business decisioning

Guide	The Ark COVID-19 Dataset	
Executive Summary	The Ark is making available a COVID-19 dataset that provides estimates of risk factors and COVID-19 infection rates at a range of local geographies (Ward, Parliamentary Constituency and Clinical Commission Group). The plan is to update the data periodically to enable the tracking of infection rates at a local level over time across England and the wider UK. The Ark has used its existing and new datasets to create 14 risk measures that it considers to be relevant to COVID-19. As part of this we have applied our disaggregation method to estimate cumulative infection rates on an “as is” basis in England and on a “timeline adjusted” basis across the whole of the UK. The comparison of risk factors and infection rates at a local level suggests that there are plausible associations between the two. The high infection rates seen to date are dominated by London and are occurring in locations with higher overcrowding risks, and higher overall engagement risk (indicating adherence to the lockdown advice may be less rigorous in these locations). The time adjusted pattern for the UK as a whole, shifts the risk profile to areas with poorer health and lower incomes, but with a hot spot still clearly associated with overcrowding that is consistent with the “as is” profile. A further comparison of modelled infection rates to the Output Area Classifications provided by ONS show that categories in super groups 3: Ethnicity Central, 4: Multicultural metropolitans and 7: Constrained city dwellers, are over indexed either on the “as is” or timeline adjusted models. Sub-groups are generally over-indexed compared to their parent Super Group where they have residents who live in more overcrowded conditions and / or use public transport more and / or have a higher proportion of workers in industries engaging with the general public (e.g. accommodation, food service). To support the wider analytical community investigating COVID-19, we are making our datasets at Ward, Parliamentary Constituency and CCG level freely available under the Creative Commons Attribution-Non Commercial 4.0 International Licence. For this study we have only used aggregated open-source data, which means that there are no GDPR implications clients need to be concerned with when using our COVID-19 datasets.	
Abbreviations	BMJ	British Medical Journal
	CCG	Clinical Commission Group
	OAC	Output Area Classification
	ONS	Office of National Statistics
	UTLA	Upper Tier Local Authority

Introduction

The Ark provides innovative data solutions for a wide range of business and market sectors, from financial services to charities, energy providers and retailers.

In response to the challenges posed by COVID-19, we are offering a selection of our modelled data to anyone who can make good use of it. Our hope is that our data will provide a range or useful measures at a local neighbourhood level that are associated with an increased risk of being infected by COVID-19 or affected by the actions taken to mitigate its spread. The range of measures provided includes age and household risks, health and mortality risks, economic risks and engagement risks.

To enhance our risk data, we have estimated COVID-19 infection rates in the population starting in week 01 at 5th April 2020 and planned to be updated weekly. The source data for this analysis is taken from the Upper Tier Local Authority (UTLA) data series published by Public Health England. In order to estimate infection rates for a range of different geographies other than UTLA, we disaggregate our UTLA infection estimates down to Output Area. We re-aggregate these local estimates by other geographies of interest (e.g. at Ward level) which we publish alongside the associated risk measures outlined above to provide a more complete picture of COVID-19 risks and impacts.

We outline the method and assumptions used to calculate infection rates later, but it is important to note that our infection rate estimates should be treated as indicative only. It is particularly important to note that the source data on COVID-19 cases obtained from PHE may be affected by collection bias. We estimate that only about 1 in 35 of people who are infected is tested currently, and under current NHS protocols the tested group are a subset of the infected population who are presenting with symptoms. A current hot topic is whether people from the BAME community are more prone to become seriously ill with COVID-19, and if this were to be the case, our estimates on infection rates are likely to overestimate the underlying rate of infections in this segment of the UK population.

The possibility of bias in any particular measure has prompted us to take a more holistic view by publishing a broad range of data related to COVID-19. The different measures we are providing are aimed at giving a rounded view of COVID-19 risks that are relevant to a wide range of organisations across all parts of the UK. We are pleased to offer our data to any and all of you who can make good use of it at this challenging time.

Overview of the The Ark COVID-19 Data

The main dataset we are making available is a multi-dimensional dataset that ranks all Wards across the UK. Separate rankings are also being made available tagged by Clinical Commissioning Groups (CCG) and Parliamentary Constituencies and are supplied in a single Excel Workbook.

The data we have used for this exercise is our existing datasets derived exclusively from aggregated Open Source data. There are no GDPR implications that users should concern themselves with as we have not used any personal data or PII data in creating this output.

The data series for risk measures are ranked versions of our modelled data by percentile, with 1 = lowest risk and 100 highest risk for Wards. The rank scale is from 1 to 20 in the case of parliamentary constituencies and CCGs.

All of these datasets include our estimated values for the "as is" infection rate starting at week 01 at 5th April 2020 modelled from PHE data for England. A second "time adjusted" estimate of COVID-19 infection rates is also included. These values are calculated on the assumption that all neighbourhoods in all parts of the UK are infected at exactly the same time point. The resulting modelled output then estimates resulting infection rates that occur within neighbourhoods thereafter. This aims to provide a "level playing field" view of transmission risks across the UK as a whole. These estimates have been based only on the data for England and applied across the UK. As mentioned earlier the values obtained may be subject to bias caused by the current NHS testing protocols focussing on people who present with symptoms.

In total we provide risk rankings by 14 variables which we have grouped into five over-arching dimensions.

Risk ranking: Age and Household	This is aimed at identifying locations with a higher proportion of older people, those living in larger households, and those living in small spaces with a high number of residents per room. The four measures that sit within this dimension are • **All Age Risk** derived from ONS Age data weighted by COVID-19 death rates by age band. Includes communal residents. • **Household Age Risk** is derived from ONS Age data weighted by COVID-19 death rates by age band. Includes household residents only. • **Room Risk** is derived from ONS data for the number of household residents divided by the number of rooms in residential properties. This ratio gives an indication of overcrowding within properties. • **Resident Risk** is derived from ONS data for the number of household residents divided by the number of residential properties. This ratio gives an indication of household size with larger households having greater risk of catching COVID-19 from others in the household.
Risk ranking: Mortality and Co-morbidity	This is aimed at identifying locations with a higher proportion of the population who have high health risk factors. All of the variables used for these rankings are age adjusted. The three measures that sit within this risk dimension are • **Mortality Risk** is derived from a The Ark disaggregation of ONS published population death counts • **Obesity Risk** is derived from a The Ark disaggregation of PHE population overweight proportion • **Smoker Risk** is derived from a The Ark disaggregation of PHE smoker proportion and ONS lifestyle data.
Risk ranking: Economic Resilience	This is aimed at identifying those locations with low wealth and low Incomes before the COVID-19 outbreak who have fewer financial reserves to call on during the lockdown. In addition, we have analysed those neighbourhoods that are most likely to suffer additional hardships because of potential financial hardships caused by the lockdown. This has a differential impact on those working in particular sectors of the economy defined by combinations of Industry Sector Risk and economic activity. Whilst these neighbourhoods may have good levels of wealth and income prior to the COVID-19 outbreak, they may be suffering a large drop in income during the lockdown and have to cut back or dip into savings to cover the gap. The three measures that sit within this risk dimension are: • **Income Risk** is derived from a The Ark disaggregation of ONS Annual Survey of Hours and Earnings (ASHE) data • **Wealth Risk** is derived from a The Ark disaggregation of Inland Revenue counts of estates subject to Inheritance Tax • **Employment Risk** is derived from a The Ark imputation of Industry cross-tabbed with economic activity status (employed, self-employed, inactive) cross tabbed with hours worked (part time, full time). A subjective risk value of 1 to 10 is attached to each combination of Industry x Economic Activity x Hours worked to reflect what we think is the impact of the lockdown on different groups. Risk 1 is low and 10 is high, with people who are economically in-active given a value of 0. The weighted average of the risk value is calculated at Output Area level and converted to a ranked risk between 1 and 100.

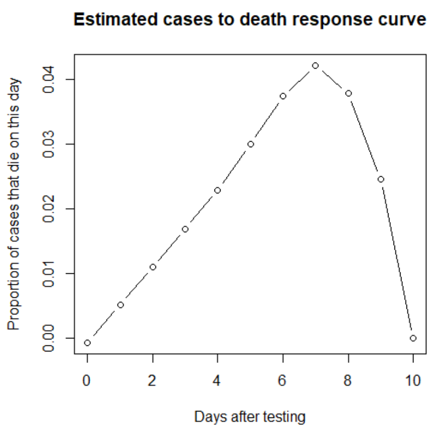
Risk ranking: Engagement	This is aimed at identifying those locations that may be less concerned and / or less well-informed about COVID-19 and its impacts. The risk is these neighbourhoods may pay less attention to the advice from Government resulting in higher infection rates. We have used out analysis of UK parliamentary petition data (pre the COVID-19 outbreak) to estimate these risks. For this analysis we have adjusted for local age profiles and Country to account for differences in participation rates caused by these two confounding factors. The two measures that sit within this risk dimension are: • **COVID-19 Engagement Risk** is derived from a The Ark disaggregation of a basket of petitions relating to health, environment and education factors that demonstrate a high-degree of empathy with vulnerable groups. Those locations with low levels of engagement with these particular petitions are viewed as being of relatively higher risk. • **Overall Engagement Risk** is derived from a The Ark disaggregation of all petitions. Those neighbourhoods that have a relatively low overall engagement in e-petitions may be less well informed on COVID-19 advice from Government and are viewed as being of relatively higher risk.
Risk ranking: COVID-19 Infection Rates	COVID-19 infection rates have been estimated by The Ark on a best efforts basis. The method used to do this is outlined later in this document. The two measures that sit within this risk dimension are: • **COVID-19 infection rate "as is".** This is an estimate of the cumulative infection rate defined as the proportion of the population aged over 10 that has been infected at some point with COVID-19 up to and including a particular date. This dataset starts in week 01 at 5th April 2020. The intention is to update this estimate periodically to show how cumulative infection rates are changing over time. There is a lot of regional variation in these estimates, with London weeks ahead of some other parts of England. This measure is only available for England. • **COVID-19 infection rate timeline adjusted.** This is a cumulative infection rate estimate adjusted for different timelines and is calculated on the assumption that every part of the UK started to become infected at exactly the same time point. The overall level of infection is scaled to be roughly the same as the average England level to aid comparisons with the "as is" estimates where appropriate. All parts of the UK are included in this measure with locations outside of England being scored up using the model derived from English data. This standardised measure will also be updated periodically to help track the progression of infection rates over time.
The method used to estimate UTLA COVID-19 infection rates	The analysis we have undertaken to investigate associations between COVID-19 daily infection rates and neighbourhood characteristics requires us to estimate COVID-19 infection rates at increasingly localised geographies, starting with calculating the national level average for England, then estimating the infection rate at UTLA level The national and UTLA infection rate estimates are obtained from week 01 at 5th April 2020 from the cumulative number of cases and deaths. We use publicly available data about COVID-19 death rates and time in hospital to find the relationship between the number of reported cases for each UTLA and the underlying infection rate in the general population. The calculation is split into a sequence of 3 steps as follows:

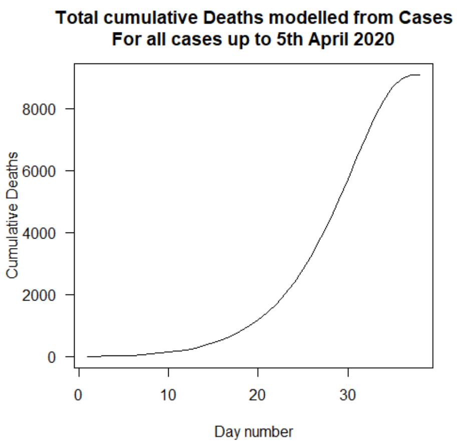
1. Account for the lag in deaths compared to cases.

A recent paper published mentioned on the BMJ website details the length of stay in hospital for a typical COVID-19 patient which indicates a stay of up to 17 days after admission[1]. It is logical to assume that deaths for cases primarily occur during the period after testing for COVID-19 when critically ill patients are likely to be admitted to hospital for observation and treatment.

To quantify this, we back solve between deaths and cases by assuming a death rate for cases that starts relatively low on the day the test is obtained, rises to a peak at some point of the hospital stay and falls to zero at some point before the hospital admission ends. This analysis provides us with a response curve that links cases to subsequent deaths. We have optimised the fit between cumulative cases and deaths by varying the length of time for the response curve, as well as the shape, position and height of the peak. We have used data up to 11/04/2020 for this analysis.

The results of this process are shown in the graphs below. The estimated best fit response curve has a time span of 10 days after testing with a peak death rate on day 7 (graph on left). This is broadly consistent with a hospital stay of up to 17 days allowing for the recovery and monitoring of some survivors after the death rate falls below the peak. The small negative value on the day of testing is interpreted as meaning that a few of the test results are recorded after death has occurred.





From this analysis we can obtain an estimate for the total number of deaths related to the cumulative total of cases at any particular date. In the case of England at 5th April 2020, the cumulative cases totalled 39,814 with cumulative deaths to this same date totalling 4,494. However, our estimate of the total number of deaths associated with these cases, taking account of the lag, is 9,071 (graph on right), which implies a death rate per confirmed case of c23%.

[1] This BMJ research news page gives information about COVID-19 https://www.bmj.com/content/369/bmj.m1327

2. Account for the under reporting of COVID-19 cases due to testing limitations

It is widely accepted that many cases of COVID-19 are being missed in the UK because of the limited amount of testing that is being done. To account for this, we use the widely reported global COVID-19 mortality rate of 0.66% for the population as a whole. We simply compare the estimated death rate (based on cases) to this global population mortality rate to estimate the number of people infected in the general population and the level of under-reporting in England.

As at 5th April 2020 we calculate there to be 1.37 million people infected at some point with COVID-19 in England, implying that current testing is picking up only about 1 in 35 cases across the country. To calculate the England infection rate we divide this total by the number of people aged over 10 which gives an estimate of circa 3% cumulative infection in the general population as at 5th April 2020. We exclude the population aged 10 and under on the basis that it is widely reported there are no significant deaths in this group and therefore it is assumed that the reported global death rate of 0.66% and the England case and death figures also excludes any significant counts from this age group

3. Obtaining UTLA infection rates from the England average infection rate

The approach used to split out the count of infections to UTLA level is basically to pro-rata the infections on the number of cases reported for each UTLA. The complicating factor in this, however, is that it is highly likely that infection rates vary significantly with age.

The published data for death rates by age group show that there is a rapid increase in death rates across age bands. For example, ages 10 to 29 are reported from the China data to have a death rate of 0.2% compared to the over 80s with a death rate of 14.8%[2]. Comparing these age banded death rates to the overall average death rate of 0.66% implies that younger age groups must have higher infection rates than older groups in order to get the death rate weighted by age band to equal the global average death rate.

There is logic to this age effect which is also worth considering. Higher risk groups (generally older) have more at stake and are therefore likely to practice social distancing and isolation diligently, encouraged by Government. This is why a unilateral lockdown of many care homes was one of the first actions taken by their managers. By contrast, the initial Government strategy of building herd immunity in the wider healthy UK population and the delayed closing of schools etc. would have disproportionately increased infection rates in younger families where death rates are relatively low.

To fit the observed data, we found using trial and error that infection rates needed to decrease with increasing age (to fit to observed deaths), offset by an increase in testing per head of population by increasing age (to fit to observed cases). The balance of these two competing age effects was to skew infection rates to UTLAs with younger populations when cases per head of population at the UTLA level are roughly the same. Again, there is logic to this in that it is very likely that the need for a medical intervention (prompting a test) is much lower for younger citizens (who generally have mild symptoms) compared to older citizens (who are much more likely to fall seriously ill). One of the checks of our overall method at this stage is to observe that the chosen trade-off results in the "right" overall age profile for tests that is in line with the hospital admission data. In particular that the peak in our modelled testing rates occurs at age bands between 50 and 70, broadly in line with the peak in hospital admissions by age in the UK.

Pro-rating by the UTLA level of cases per head, adjusting for the age mix enabled us to estimate average UTLA infection rates that tallied with the overall England infection rate value when aggregated back to national level.

[2] After the analysis presented here was finalised we learn from the BMJ that recent UK studies published in the Lancet show the China age figures for death rates are over-estimates, but the variation with age is similar. This is unlikely to affect our infection rate figures as long as the overall death rate of 0.66% remains unchanged, which at the time of writing seems to be the case.

Local Infection Rate estimation using Disaggregation	To obtain localised estimates of infection rates we apply our disaggregation method to obtain modelled estimates of infection rates at Output Area level. Once the Output Area estimates are obtained, we re-aggregate these local estimates back to intermediate geographies of interest, namely Ward, Parliamentary Constituency and CCG. Disaggregation involves building a regression model iteratively to apportion the calculated UTLA infection rates across neighbourhoods based on their characteristics. We have found that our disaggregation method gives reliable local estimates when the data available is of high quality and well distributed at higher geographies. Typically, we would disaggregate from a starting point of lower tier local authority or parliamentary constituency and would have more than 300 data points to work with. In the case of our COVID-19 disaggregation the data is not ideal. We start with only c150 data points at UTLA level in England. In addition, there are two significant complicating factors that we need to deal with as we undertake our disaggregation modelling. The first is that different parts of the country are at different time points on the infection curve, with London weeks ahead of other areas of the country. The second is that there are likely to be effects specific to students which include a significant movement of the student population, prompted by COVID-19. Universities have closed their off-line operations during the lockdown, prompting students to return home. This is potentially a mass migration of young people between parts of UK at a time when infection rates were rising exponentially. The net movements are unlikely to be uniformly spread across UTLAs, because there is a wide variation in the student proportions across UTLAs. To account for the timing issue, we use daily case growth rate estimates for each UTLA to calculate how long ago each was at a very low infection rate of 0.1%. This allows us to place all of the UTLAs along a timeline. Time variables are then used to adjust the target values used in the regression model to account for the timing effect separately from the variations related to neighbourhood characteristics. We add back the time adjustment once the neighbourhood model is built to calculate the “as is” infection rate estimates. We can only obtain the “as is” estimate for England UTLAs. We calculate a second “time adjusted” infection rate with the time variables set to the UTLA average for all UTLAs. This scored value can be calculated for all parts of the UK. The time-adjusted estimate is the best one we have for investigating the impact of local characteristics on infection rates on a uniform basis across the UK. To account for the student issue, we use ONS figures for the proportion of full-time students at their term address and at their home address summarised at UTLA level. These give us a measure of the size of the potential movement effect (from and to). They are included in the neighbourhood regression models for both the “as is” and “time adjusted” infection rate estimates, but estimates are calculated with these two variables set to an average UK value for all locations in both models. This should then provide a correction for the student population and their net movements across the country. Our local modelled estimates should be considered as providing guidance rather than precise estimates of infection rates. This is because of the limited data available, the potential problems with bias in the data collected, and the need for significant adjustments for timing and student movements.

Potential Uses of the Analysis	The disaggregated estimates at Output Area level are used to create our estimates of infection rates at intermediate geographies (Ward, Parliamentary and CCG). These estimates at intermediate geographies are calculated by re-aggregating the very local Output Area estimates and it is reasonable to assume that this averaging process should remove some of the "rough edges" from our modelled data, giving fit for purpose estimates that can be used to inform local decision making. For example, our analysis should allow a Member of Parliament to compare the level of risk in her constituency to others across the UK with reasonable confidence and then to work collaboratively with Councillors in her constituency to focus local efforts and resources in Wards where the need and risk is greatest. In another example, it should allow groups of MPs who are in high risk / low infection rate locations to lobby for approaches that relax the lockdown conditions very gradually to avoid overwhelming the NHS in their area. Whereas another group of MPs in low risk locations may wish to relax lockdown conditions more quickly to alleviate financial hardship safe in the knowledge that the number of COVID-19 cases is unlikely to "take off" in an uncontrolled way. It is therefore unlikely that a one-size fits all approach to relaxing the lockdown is going to be appropriate and our datasets are aimed at helping decision makers navigate the tricky balancing act that will be required to do this optimally. Having a view of both risk and infection rates side-by-side is of critical importance in finding this balance. Our data provides such a view and can be used to complement other sources of information as they become available, such as data from contact tracking apps. We will update our infection rate estimates periodically to help chart progress, particularly until the roll out of mass testing is underway. For large organisations with a sizeable workforce, our data could be used alongside sickness record data to provide a more robust assessment of infection rates that compares our top-down estimates against internal records on the self-isolation and COVID-19 infection rates of staff. For an organisation with 10,000 staff spread nationally, we would expect maybe 400 at the time of writing to have been infected with COVID-19. This number of cases should give a reasonable starting point for a comparison to our data, particularly if this is done using our most geographically detailed datasets. Once an organisation has confidence in how the level of infection varies across the areas it is operating in, it can make much better-informed operational decisions.

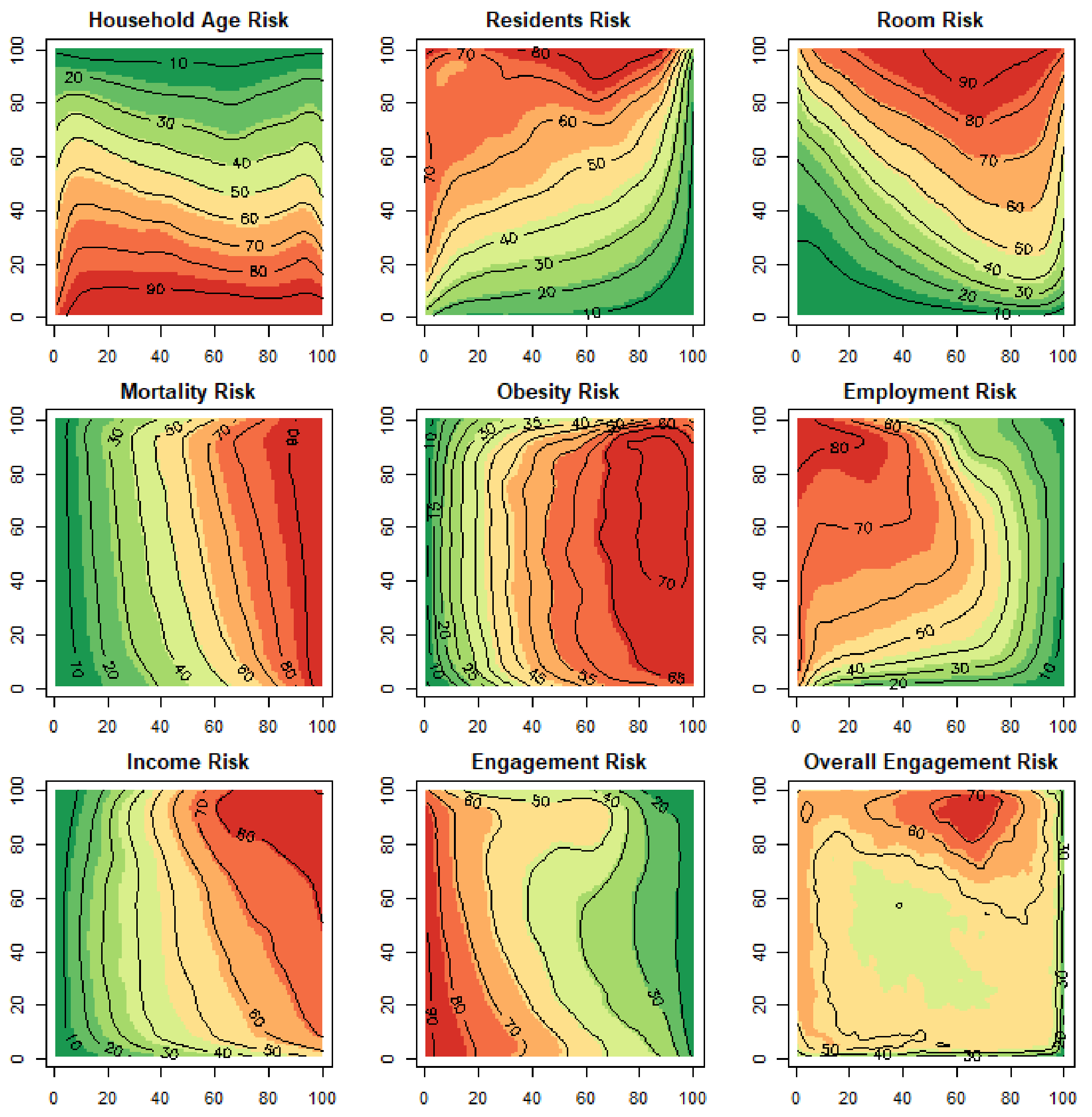
Assessment of The Ark modelled results

To provide some evidence that our estimates of COVID-19 infection rates are plausible we review how our infection rate estimates look at the most localised level (Output Area). To do this we start with a comparison of our different dimensions of risk using a COVID-19 Risk Map and follow on with an analysis of infection rates by the map and the ONS geo-demographic Output Area Classification (OAC) categories.

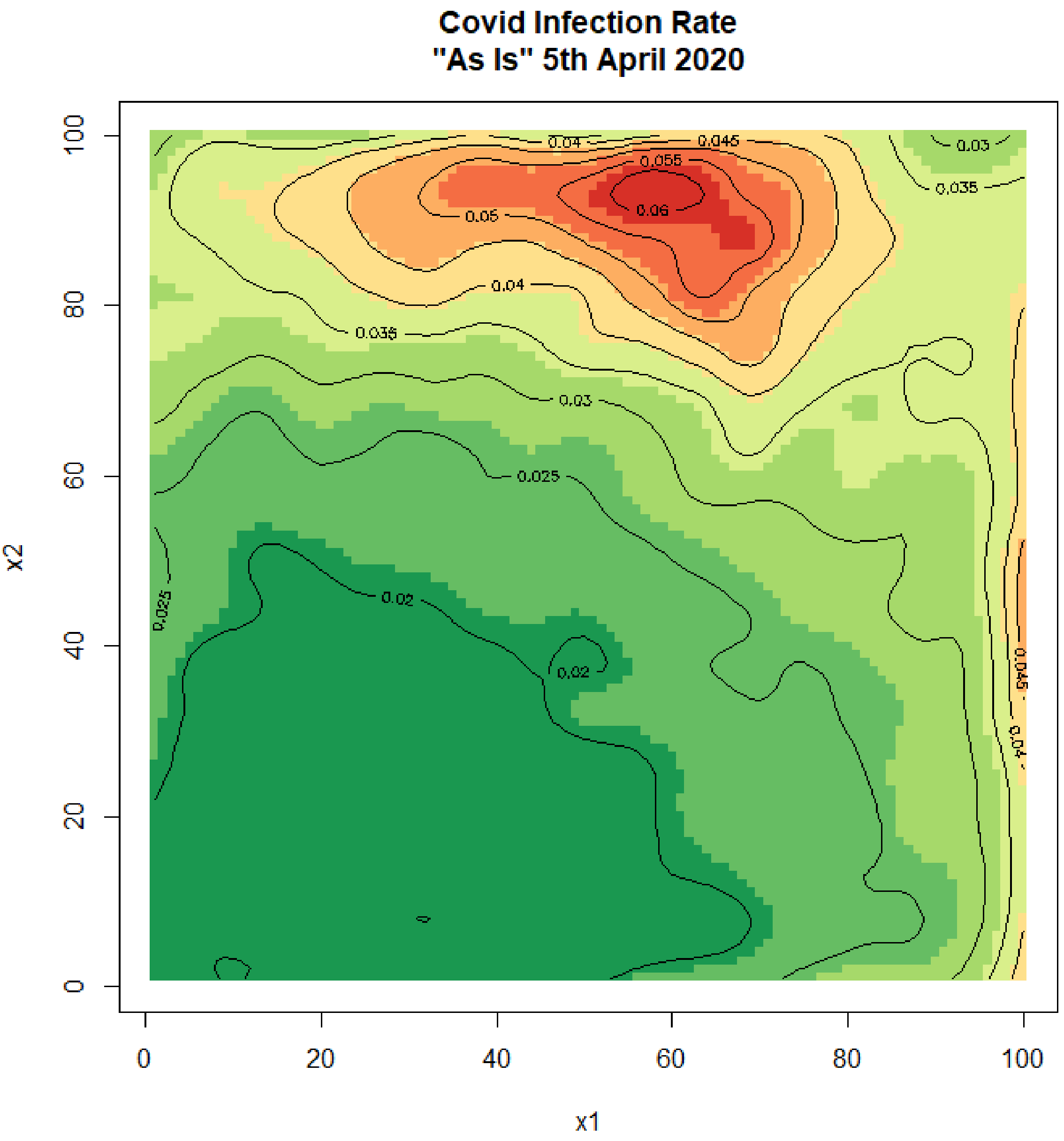
1. COVID-19 Risk Mapping

The COVID-19 Risk map provides a visualisation of COVID-19 risks across different dimensions.

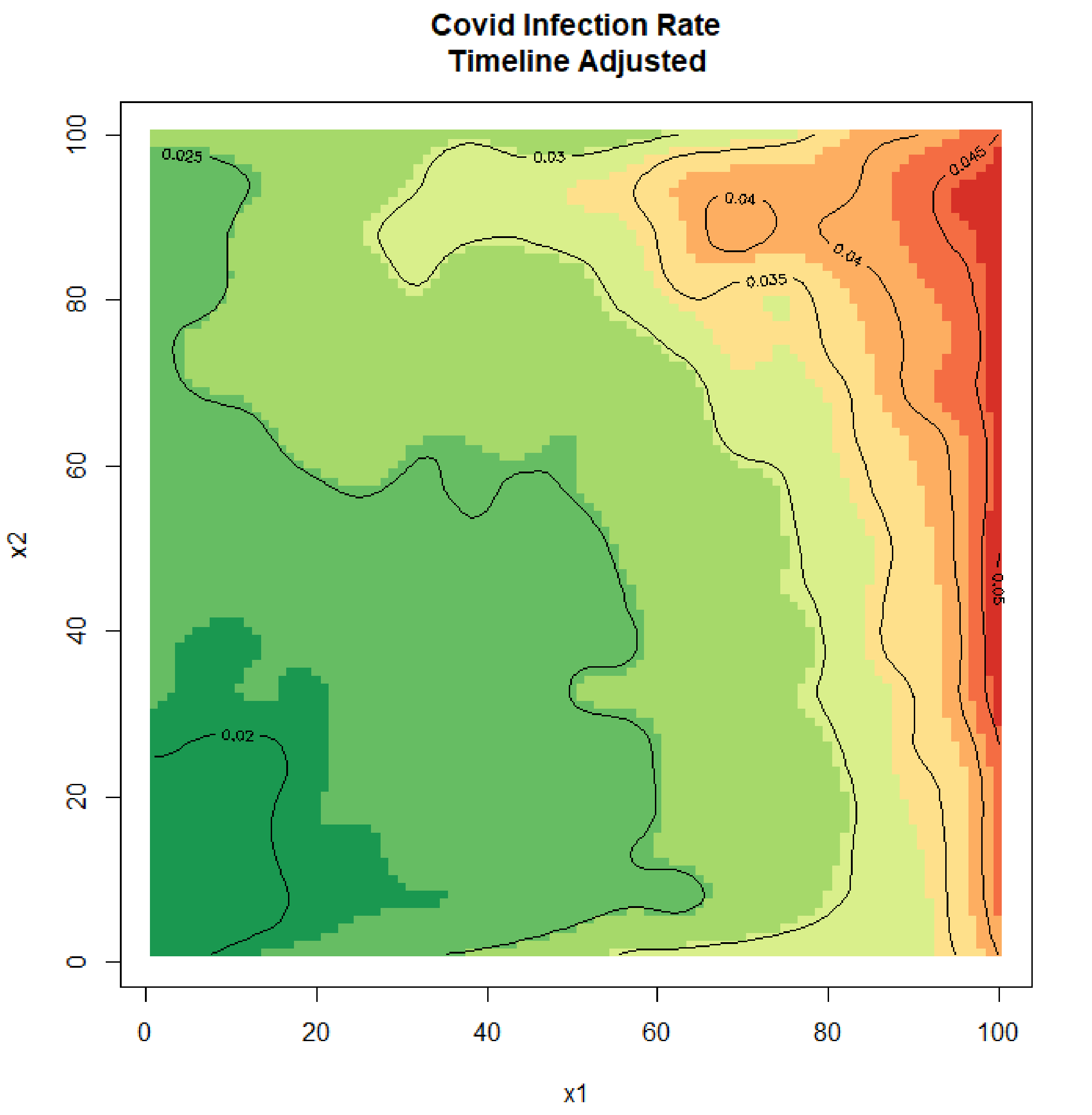
The map is created using a selection of our risk ranks to undertake a 2-dimensional factor analysis. The factor coefficients are then used to position individual Output Areas on a 100 x 100 grid with equal numbers of small-area neighbourhoods at each point on the grid. The average value for each risk index is then calculated and risk contours are plotted. Visual inspection allows us to identify different areas of the risk map associated with patterns of risk across multiple dimensions. As we see below, this gives us a useful framework for analysing COVID-19 infection rates. High risk is red and low risk is green.



We then overlay the estimated Output Area Infection rates on the same map to identify where we are seeing hot spots in relation to particular risks. The first analysis looks at the position “as is” on the 5th April 2020. In this map the hot spots largely reflect the characteristics of those areas in London that have been most impacted by COVID-19. This will be driven in part by the characteristics of these neighbourhoods and an element of “pot luck” as to where infections happened to start early or late on the timeline. The contours plot the infection rate, which is high (red) in the areas where Room, Resident and Overall Engagement risks are higher. This suggests over-crowding is a driver of infection rates. A second hotspot area on the edge of the map to the far right suggests that health risks may also be important.



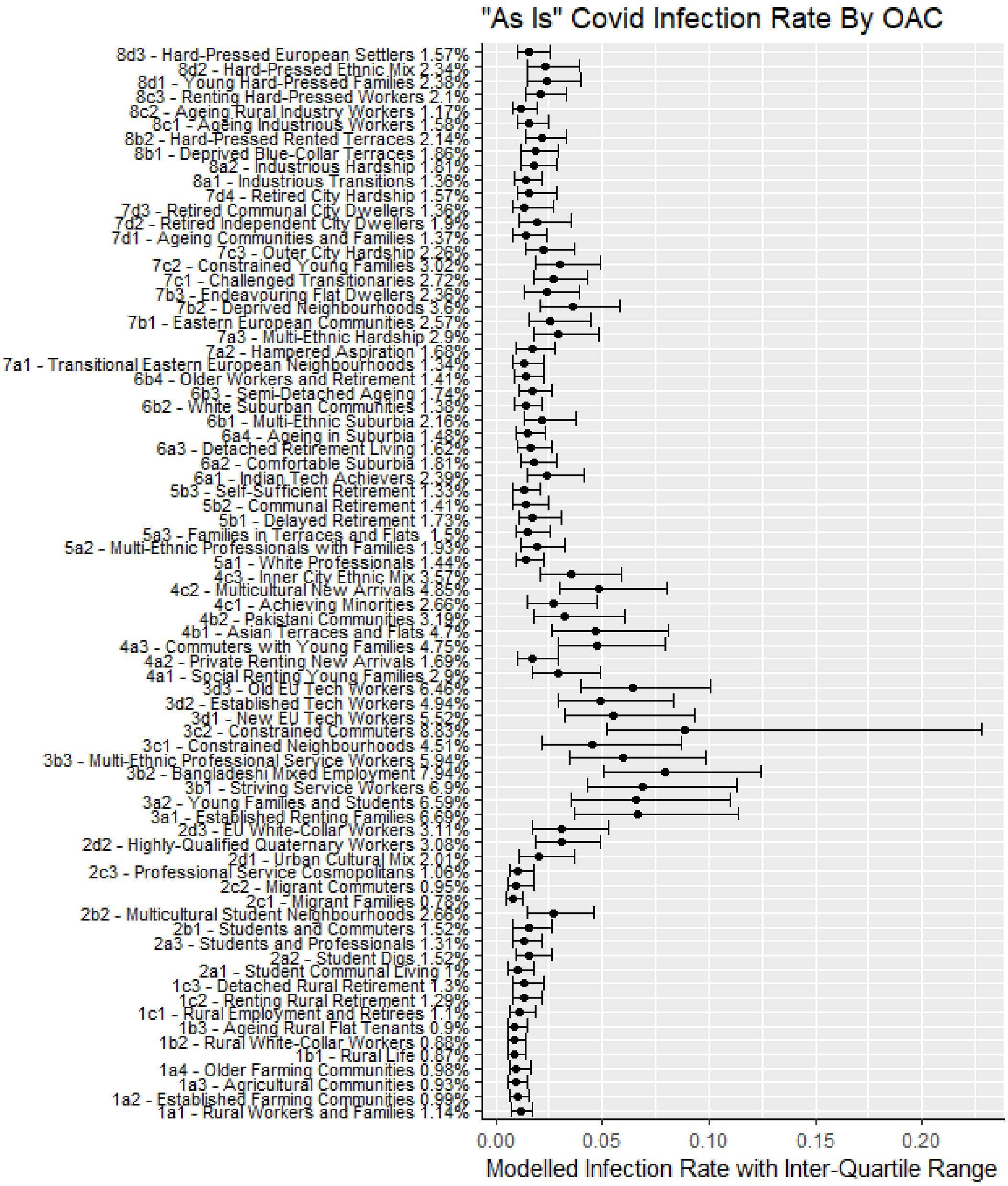
The second analysis is adjusted for time effects and estimates the infection rate for all parts of the UK assuming initial infection started at the same time point everywhere. The hotspot pattern is shifted significantly to the right towards areas of high mortality, health, wealth and income risks. The bulge in the pattern to the left at the top of the map sits where the Room and Resident and Overall Engagement risks are higher. This pattern suggests that a combination of over-crowding, morbidity and poverty factors are all associated with higher infection rates across the UK as a whole



2. Analysis by Output Area Classification (OAC) Categories

For this analysis we calculate the median and inter-quartile range (IQR) of the Output Area infection rate estimates for each OAC category. We review these values for both the "as is" and "timeline" adjusted models.

In the graph below we show estimated infection rates by OAC categories for the "as is" model. Super Groups 3 and 4 are over indexed. The sub groups that generally show higher infection rates compared to their parent super group (e.g. 2b2, 3c2, 4c2, 6a1, 7b2, 7c2) are found to have residents who live in more overcrowded conditions and / or use public transport more and / or are more likely to work in industries that have higher levels of contact with the general public (e.g. accommodation and food service).[3]

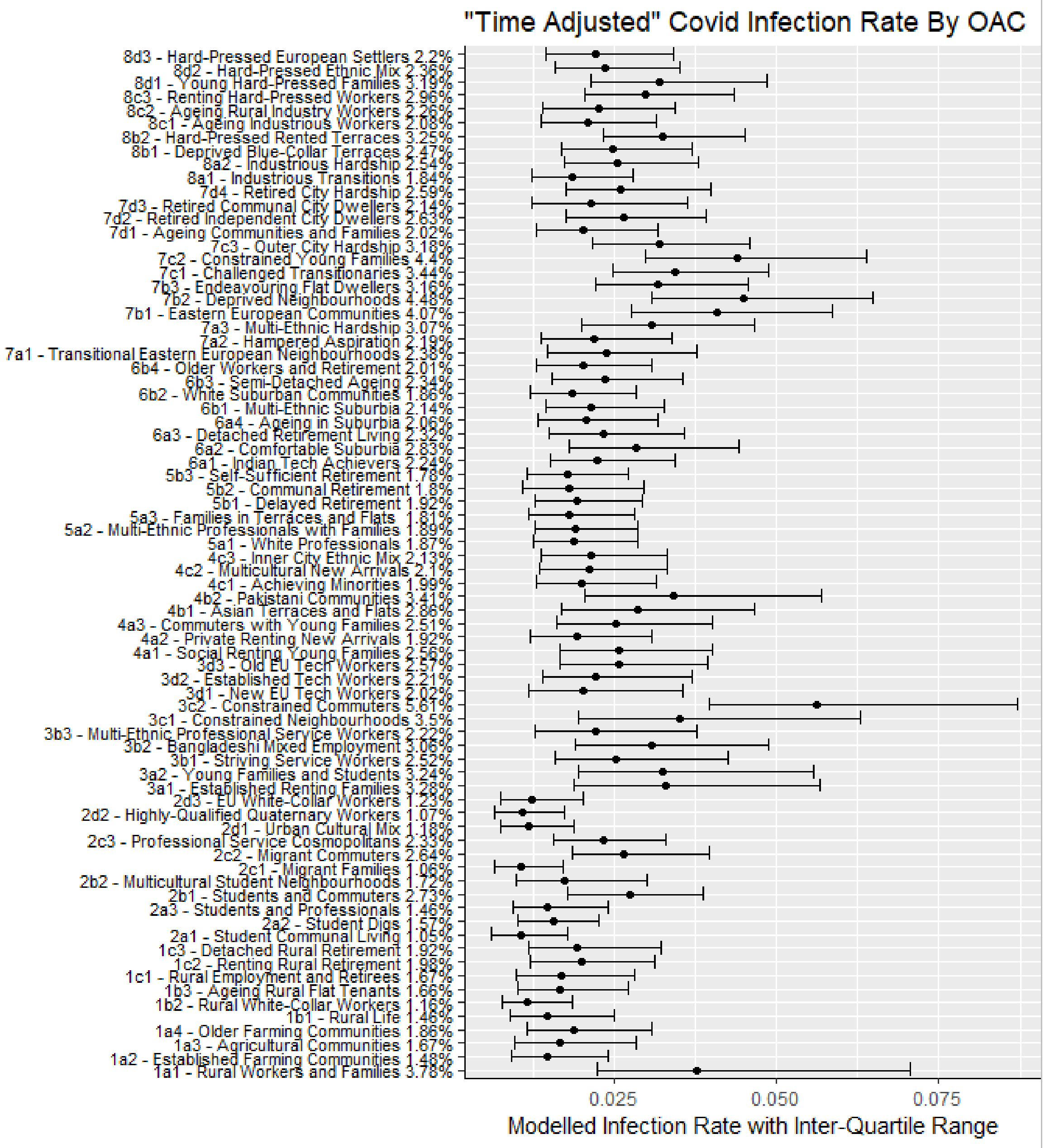


[3] More information on the OAC categories including pen portraits and radial plots can be found here. https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications

In the graph below we show the infection rate estimates by OAC categories for the “time adjusted” model. The dominance of London should not be a feature in this analysis, which attempts to give a balanced view across the whole of the UK.

The over-indexed Super Group 3 is the same as for the “as is” model, but with a smaller range. Super Group 7 is also generally higher than the others here, but Super Group 4 falls back into the pack with subgroup 4b2 remaining the highest. Super Group 2 remains low, but with commuter neighbourhoods (2b1 and 2c2) markedly higher than the rest within this Super Group.

Super Group 1 remains low with the notable exception of subgroup 1a1. We think the increase in category 1a1 infection rate is due to the inclusion of Northern Ireland in this analysis. This requires more investigation, but may be associated with relatively high levels of over-crowding due to larger families in rural Northern Ireland compared to the rest of the UK.



Obtaining the Data	Our products are available directly from The Ark or through one of our partnerships with leading data agencies. Follow the links on our website to get access to our data. The Ward level, CCG and OAC classification datasets can be downloaded in a single excel workbook from the The Ark website by registering your details. Alternatively, individual files of data can be obtained from our data agency partners. These data distributors can also supply more geographical detailed datasets on a commercial basis if required. Special rates are available for those users who can demonstrate that their use of our detailed data is only for non-commercial reasons that support the public good.
Data acknowledge-ments and attributions	The COVID-19 dataset contains data from other sources which have their own copyright notice as follows: Contains OS data © Crown copyright and database rights 2020 Contains Royal Mail data © Royal Mail copyright and database rights 2020 Contains National Statistics data © Crown copyright and database rights 2020 Contains Public Health England Data © Crown copyright and database rights 2020 The Ark retains copyright to the rest of the COVID-19 data derived from open source data © The Ark copyright and database rights 2020. Our main source of data for models is 2011 census data supplemented by a wide-range of more up to date data provided by National Records of Scotland (Crown Copyright, OGL), Northern Ireland Statistics and Research Agency (Crown Copyright, OGL), Office of National Statistics (Crown Copyright, OGL). Output Area mappings to other Geographies are taken from the ONSPD / NSPL files and other lookup files regularly published by ONS. These files contain National Statistics data © Crown copyright and database right 2020 **https://www.ons.gov.uk/methodology/geography/geographicalproducts/postcodeproducts** Other data sources used by us in feeder models for the COVID-19 dataset are Crown Copyright and used under Open Government Licence v.3.0, as follows: Inheritance Tax model uses data published by HMRC Earned Income model uses Annual Survey of Hours and Earnings (ASHE) data published by ONS Obesity and Smoker models use data published by Public Health England (PHE) and ONS Engagement risk models contain Parliamentary information licensed under the Open Parliament Licence v3.0 **https://www.parliament.uk/site-information/copyright/open-parliament-licence/**