Understanding community level influences on COVID-19 prevalence in England

New insights from comparison over time and space

Chaitanya Joshi[1,a], Arif Ali[1,+], Thomas ÓConnor[1,+], Li Chen[1,b], and Kaveh Jahanshahi[1,c]

[1]Data Science Campus, Office for National Statistics, UK
[a]chaitanya.joshi@ons.gov.uk
[+]these authors contributed equally to this work
[b]li.chen@ons.gov.uk
[c]kaveh.jahanshahi@ons.gov.uk

ABSTRACT

The COVID-19 pandemic has impacted communities far and wide and put tremendous pressure on healthcare systems of countries across the globe. Understanding and monitoring the major influences on COVID-19 prevalence is essential to inform policy making and device appropriate packages of non-pharmaceutical interventions (NPIs).
This study evaluates community level influences on COVID-19 incidence in England and their variations over time with specific focus on understanding the impact of working in so called high-risk industries such as care homes and warehouses. Analysis at community level allows accounting for interrelations between socioeconomic and demographic profile, land use, and mobility patterns including residents' self-selection and spatial sorting (where residents choose their residential locations based on their travel attitudes and preferences or social structure and inequality); this also helps understand the impact of policy interventions on distinct communities and areas given potential variations in their mobility, vaccination rates, behavioural responses, and health inequalities. Moreover, community level analysis can feed into more detailed epidemiological and individual models through tailoring and directing policy questions for further investigation.
We have assembled a large set of static (socioeconomic and demographic profile and land use characteristics) and dynamic (mobility indicators, COVID-19 cases and COVID-19 vaccination uptake in real time) data for small area statistical geographies (Lower Layer Super Output Areas, LSOA) in England making the dataset, arguably, the most comprehensive set assembled in the UK for community level analysis of COVID-19 infection. The data are integrated from a wider range of sources including telecommunications companies, test and trace data, national travel survey, Census and Mid-Year estimates.
To tackle methodological challenges specifically accounting for highly interrelated influences, we have augmented different statistical and machine learning techniques. We have adopted a two-stage modelling framework: a) Latent Cluster Analysis (LCA) to classify the country into distinct land use and travel patterns, and b) multivariate linear regression to evaluate influences at each distinct travel cluster. We have also segmented our data into different time periods based on changes in policies and evolvement in the course of pandemic (such as the emergence of a new variant of the virus). By segmenting and comparing influences across spaces and time, we examine more homogeneous behaviour and uniform distribution of infection risks which in turn increase the potential to make causal inferences and help understand variations across communities and over time.
Our findings suggest that there exist significant spatial variations in risk influences with some being more consistent and persistent over time. Specifically, the analysis of industrial sectors shows that communities of workers in care homes and warehouses and to a lesser extent textile and ready meal industries tend to carry a higher risk of infection across all spatial clusters and over the whole period we modelled in this study. This demonstrates the key role that workplace risk has to play in COVID-19 risk of outbreak after accounting for the characteristics of workers' residential area (including socioeconomic and demographic profile and land use features), vaccination rate, and mobility patterns.

1 Introduction and Background

The response of governments across the globe to COVID-19 pandemics has included non-pharmaceutical interventions (NPIs). In the UK, those include mobility restrictions, closures of some industrial sectors and schools, social distancing and mandatory face covering in public areas and public transport. Assessing the effectiveness of these policies requires thorough understanding of risk factors and behavioural responses which tends to vary over time and across communities and areas. In response, this paper focuses on understanding the spatial and temporal (from 4th October 2020-5th December 2021) variations of COVID-19 risk factors at community level.

Over the last couple of years, many studies have focused on individual and household level influences for COVID-19

infection (e.g. House et al, 2021; Williamson, E.J et al, 2020; Katikireddi et al, 2021; Jin J et al, 2021; Sze S et al, 2020). For instance, House et al (2021) uses the Office for National Statistics (ONS) COVID-19 Infection Survey (CIS)[1] to evaluate within and between household transmission and the associated factors ranging from socioeconomic and demographic characteristics to genome structures and vaccination records[2]. Williamson et al. reported one of the earliest studies outlining risk factors that are associated with COVID-19-related deaths in England (Williamson, E.J et al. 2020)[3]. Katikireddi et al (2021) highlight the interplay of individual demographic factors based around ethnicity which contributes to certain groups having unequal risk exposure to COVID-19 such as household sizes, disease vulnerability and occupation[4]. Raisi-Estabragh et al (2020) have investigated the heightened risk of COVID-19 to Black and Asian ethnicities using data from the UK Biobank (UKB) highlighting the importance of risk influences such as material deprivation and housing conditions[5]. Jin et al (2021) have devised a population risk calculator for COVID-19 mortality based on various sociodemographic factors and pre-existing conditions[6]. Sze S et al (2020) reviewed existing works on the association of ethnicity with vulnerability to COVID-19 infection and clinical outcomes at the individual level[7].

Going beyond inferring COVID-19 risk factors at individual level, there have been limited investigations mapping and identifying community risk factors for COVID-19 (Khunti 2020, Wang 2021)[8,9]. Examples are Khunti et al (2020) which pointed out potential association between ethnicity and outcome in COVID-19 in ethnically diverse communities such as in the UK[8] and Wang (2021) which explored the risk factors for nursing home COVID-19 death rates[9]. However, to the best of our knowledge, we are not aware of any prior investigation which comprehensively evaluates community level risks of coronavirus disease (COVID-19) including those of industrial sectors after controlling for areas' associated characteristics (including real time mobility patterns); this forms the motivation of our present investigation.

Whilst person level analyses mainly consider individual and household characteristics, the community level analysis focuses on the make-up of an area and can account for interactions between socioeconomic profile, mobility patterns, and land use features of neighbourhood and settlements. For example, the ONS individual level COVID-19 risk screening model[10] helps identify characteristics of people who are more likely to test positive for COVID-19 in specific periods of time. The model screens the different characteristics of people (including their occupation and associated industrial sector) sampled in CIS who have and have not been tested positive and uses a logistic regression model to assign risk to each of these characteristics. For example, the ONS screening model can report the relative risk of a person from a four-person household compared to those from a one-person household. On the other hand, a community level risk model has the potential to control for area characteristics; for instance, it might look at how the proportion of multi-person households within a Lower Layer Super Output Area (LSOA) [1] affects positivity after accounting for mobility and other area features.

Community level analysis can provide additional insights as pandemic tends to behave differently across communities and area types; for instance, the rate and number of infections is expected to be higher in dense urbanised areas when compared to more rural neighbourhoods. Also, under our approach, the spatial interactions such as those among socioeconomic characteristics and area features can be evaluated. For instance, we can control for the potential residential preferences and spatial sorting of ethnic minority groups (e.g. data reveals that the Asian ethnic group as a whole are more likely to live in larger families with children and more dense urbanised areas- refer to Table 4 and Figure 3). Moreover, our model allows policy makers to understand the impact of national interventions at local and community level; this helps evaluate potential inequalities across communities specifically for disadvantaged groups and monitor the areas and sectors which are most resistant to policy interventions.

Effective evaluating and monitoring of the pandemic and the associated policy interventions at community level would require geographically detailed datasets that not only cover an extensive range of static influences including socioeconomic and demographic profiles and land use features, but also provide a structured and temporally detailed understanding of behavioural responses to policy interventions such as those reflected in the trend of mobility indicators or vaccination. Furthermore, the empirical data would need to cover the dynamic of COVID-19 prevalence. This is a very tall order indeed, and the requirements are unlikely to be satisfied by the majority of known datasets.

In this context, we assembled and fused the required data from a wide range of data sources including Census 2011, 2019 Mid-Year population estimates, Inter-Department Business Register (IDBR) 2019 dataset on workplace and industrial sectors, the national travel survey (NTS) 2002 to 2015 to segment travel patterns and clusters, mobile phone data to extract real-time mobility indicators, uptake of first and second doses of COVID-19 vaccination and test and trace data for gathering dynamic information on COVID-19 incidents. The list of the variables is arguably the most comprehensive for LSOA level analysis of COVID-19 cases.

Methodologically, to account for highly interactive and interrelated influences, we augment machine learning and statistical techniques by (a) identifying distinct area types and travel patterns through latent clustering and (b) analysing and comparing influences on COVID-19 incidents across those more homogenous and distinct clusters. In other words, the adopted latent

[1]Please refer to https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography for more information on definitions of CENSUS geography including LSOA.

clustering approach allows separating out the impact of socioeconomic and demographic characteristics, including working in certain industrial sectors, from that from the built form and travel patterns of the area where people reside (e.g. residing in high dense populated areas). In addition to spatial clustering, we have also split the time period into certain time tranches[2] defined by policy interventions, vaccination status and the dominant virus variants; this allows separating out interrelated influences and homogenising the data further for making causal inferences. Further information on our approach is provided in Section 3.

2 Data sources

The risk model utilised a mixture of static variables, which we have assumed do not change over time, and dynamic datasets, which do change over time. This section covers the sources and geographical granularity of data and discuss how those are used in our analyses. A list of features and associated datasets used in the LSOA COVID-19 risk model is shown in Table 1. Further details on the list of variables and the break down of levels for static variables are provided in Table A.1.

Table 1. List of features and associated datasets used in the LSOA COVID-19 risk model.

Static and dynamic datasets used in the model			
Input feature at residential end	**Datasets used/ published department**	**Geography**[2]	**Temporal**
Travel Clusters extracted from our earlier work[11]	Individual level National Travel Survey (NTS)- 2002 to 2015 / DfT	LSOA	Static
Ethnicity, Method of travel to work, Family structure (with and without children), Journey to work (JTW)	Census 2011, 2019 Mid-Year Population Estimate / ONS	LSOA	Static
High risk industries' workers	Inter-department Business Registry (IDBR) 2019 / ONS	LSOA	Static
Workers and Visitors Footfall	Deimos footfall dataset / Telefonica (O_2)	MSOA (dis-aggregated to LSOA using worker proportion as weights e.g. if 20% of MSOA's workers work in an LSOA, 20% of the MSOA footfall will be allocated to that LSOA)	Dynamic
Vaccination uptake-first and second doses administered	National Immunisation Management Service (NIMS) vaccination dataset / NHSD	LSOA	Dynamic
COVID-19 Cases (Y Target variable)	Second Generation Surveillance System (SGSS) positive tests dataset / PHE.	LSOA	Dynamic

2.1 Static Features

The main static features were derived from Census 2011[12–14], which albeit being 10 years old, still provides a rich source of data and detailed snapshot on socioeconomic and demographic profiles of LSOAs [3]. Nevertheless, where possible, data was supplanted by more recent versions of data such as 2019 population estimates. More detailed description of each variable used in our model is provided in Table A.1.

We also used the Inter-Departmental Business Registry (IDBR) 2019[15] to extract information on the high-risk industries we have analysed in this paper namely meat and fish processing, textiles, care homes [4], warehousing and ready meals. These group of industries are selected based on discussions we had with UK Health Security Agency (UKHSA) and Health and Safety

[2]Geography entails ONS census geography boundaries for the UK. More information can be found here: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

[3]Socioeconomic patterns and land use features such as population density or accessibility to amenities tend to be relatively stable. It can take decades for changes in population profile (such as tendency to live in less populated area for younger population due to increase appetite to work from home) to take place, therefore, Census 2011 can still capture well the land use patterns and geographical distribution of population by their socioeconomic and demographic patterns.

[4]Workers who carry out work consisting of residential nursing activities, Residential care activities for learning disabilities, mental health and substance abuse, residential care activities for the elderly and disabled, and other residential care activities

Executive (HSE) and review of high-risk industries in other parts of the word; the analysis, however, can be expanded to include other industries.

The IDBR is a comprehensive register of UK businesses used by government for statistical purposes and provides the main sampling frame for surveys of businesses carried out by the ONS and other government departments. The two main sources of input for IDBR are Value Added Tax (VAT) and Pay As You Earn (PAYE) records from HMRC. Additional information comes from Companies House, Dun and Bradstreet and ONS business surveys. The IDBR covers around 2.7 million businesses in all sectors of the economy, but since the main two tax sources have thresholds, very small businesses operating below these will, in most cases, not be included. This is unlikely to be a major problem for our analysis as the high-risk industries we have modelled are unlikely to include many of the small size businesses below the tax threshold. The IDBR 2019 data was aggregated to a workplace zone (WPZ) geography[16] and each row in the aggregated table provides the Standard Industrial Classification (SIC) code and name, industrial classification group code, WPZ, and the number of individual employed at the corresponding SIC[17]. Table 2 lists the SIC codes of the high-risk industries used in this paper.

Table 2. SIC codes for a list of high-risk industries

High risk industry	SIC CODE	SIC NAME
Meat and fish processing	10.1	Processing and preserving of meat and production of meat products
Meat and fish processing	10.2	Processing and preserving of fish, crustaceans and molluscs
Textiles	13.1	Preparation and spinning of textile fibres
Textiles	13.2	Weaving of textiles
Textiles	13.3	Finishing of textiles
Textiles	13.9	Manufacture of other textiles
Textiles	14.1	Manufacture of wearing apparel, except fur apparel
Textiles	14.2	Manufacture of articles of fur
Textiles	14.3	Manufacture of knitted and crocheted apparel
Textiles	15.1	Tanning and dressing of leather; manufacture of luggage, handbags, saddlery and harness; dressing and dyeing of fur
Textiles	15.2	Manufacture of footwear
Care	87.1	Residential nursing care activities
Care	87.2	Residential care activities for learning disabilities, mental health and substance abuse
Care	87.3	Residential care activities for the elderly and disabled
Care	87.9	Other residential care activities
Warehousing	52.1	Warehousing and storage
Ready meals	10.85	Manufacture of prepared meals and dishes

Workplace Zones are an output geography, produced using workplace data from the 2011 Census for England and Wales. They are designed to supplement the Output Area (OA) geographies which were introduced with the 2001 Census, and have been constructed from OAs, or sub-divisions of these called postcode-level building-blocks (PCBBs). While OAs are designed to contain consistent numbers of persons based on where they live, WPZs are designed to contain consistent numbers of workers, based on where people work.

The Census 2011 Journey to work dataset provides the location of usual residents living at OA and working at WPZ[5]. We merged the Census 2011 travel-to-work dataset and IDBR 2019 dataset to produce the number of residents living at the OA level by industrial sectors in which they work. OA data is then aggregated to LSOA to create LSOA level of working population by industry type needed for our analysis. Finally, for the purpose of modelling, the number of high-risk industries' workers is normalised by LSOA area and expressed in area adjusted units (per Hectare).

2.2 Dynamic Features

One of the novel aspects of our analysis was to incorporate dynamic mobility data and augment that with static socioeconomic and demographic and land use features. For this analysis, we used work and visiting footfall data from Telefonica (we call this Deimos data). For our model, we tested different mobility indicators- both their impact on COVID-19 cases and correlations between the indicators for validation. Deimos footfall of worker and visitor data proved to be the most suitable for our geographically detailed requirements. Deimos footfall provides information on the number of devices counted at the Middle

[5]The dataset can be downloaded from the census official website (https://www.nomisweb.co.uk/output/census/2011/wf02ew_oa.zip).

Layer Super Output Areas (MSOA) overlayed with behavioural insights broken by different age bands, gender, and travel purposes (visitor, worker and resident) on an hourly basis. The data that we can access is anonymised and aggregated, never allows for identification or mapping of individuals and no personal information can be identified. We aggregate the data from hourly to daily counts and further average for each modelled time tranche (see Table 3), and then dis-aggregate the counts from MSOA to LSOA geographical level, using the total number of workers at workplace (from IDBR data) normalised by area as weights.

The second dynamic dataset was the vaccination data reported by the National Immunisation Management Service (NIMS). The dataset contained anonymised individual level information including the vaccine dose the person received (their first or second), the date, and their age. For our model, we consider total number of administrated doses at the end of each successive time tranche for each individual LSOA and then normalised by the population of each individual LSOA (using 2019 population Mid-year estimate). There are some known issues with NIMS data specifically when one wishes to aggregate and estimate proportion of those vaccinated. NIMS may over-estimate denominators in some age groups, for example because people are registered with the NHS but may have moved overseas[18]. Using 2019 population Mid-year estimate can also erroneously lead to greater than 100% vaccine coverage for some age groups. To minimise some of the biases, we focused on the rate of administration of second dose after the first dose (the difference between first and second doses within given time tranches). We construct a single feature to collectively account for the speed of vaccination between the first and the second dose within each time tranche and each LSOA defined as follows:

$$\Delta V(2,1)(t_k) = \sum_{t=t_o}^{t=t_k} V(2)(t) - \sum_{t=t_o}^{t=t_k} V(1)(t) \tag{1}$$

where $V(1)(t)$ and $V(2)(t)$ are LSOA aggregated cumulative proportion of population administered with the first and second doses of COVID-19 vaccination correspondingly within each time tranche, t and $\Delta V(2,1)(t_k)$ represents the difference between the proportion of the fully vaccinated and partially vaccinated population for each successive time tranche (refer to Table 3 for the definition of time tranches). In addition to the better chance of cancelling out the biases through incorporating the difference in the cumulative administrated doses, the constructed feature (rate of being fully vaccinated) is most relevant to controlling COVID-19 incidents, our target variable. Other investigations have suggested that one dose of vaccine cannot fully immunise people against the infection and that vaccination protection against the risk of catching COVID-19 infection slides over time[19,20].

Finally, the dataset on confirmed COVID-19 positive test results was provided to us by Public Health England (PHE) and include the Second Generation Surveillance System (SGSS). SGSS contains all positive specimens for any notifiable disease in England, including all pillar 2[6] positives which includes wider population tests outside of NHS labs and hospitals such as through regional test sites, home testing kits and mobile testing sites[7]. These are subsequently transformed into case records as appropriate for a given disease. This dataset holds records on an individual level of positive test results while providing demographic information of the individuals such as age and gender. Individuals are anonymised, however, geographic information of the individual's residence is provided at a LSOA level. For the purposes of modelling, this dataset was aggregated to a LSOA geography and normalised by area (expressed as per square km) for each individual LSOA; this data served as the dependent variable for our model. It is worth highlighting that although LSOA aggregated COVID-19 positive test results can be of the most reliable indicator of the state of the epidemic, the dataset has its own limitations[21]. For instance, it has been shown that symptomatic testing is likely to underrepresent younger population. Other testing biases reported in the literature include accessibility, reporting lags, and the ethical aspect upon receiving a positive result.

3 Method of analysis

Figure 1 shows our modelling framework, we have adopted a two stage process: latent cluster analysis (LCA) to capture distinct travel and land use clusters across the country (left hand side graph in Figure 1), and a multigroup multivariate regression to estimate influences on COVID-19 prevalence within each identified travel cluster (the right hand side graph in Figure 1). We have also employed Exploratory Factor Analysis (EFA) to explore the potential correlations among features and combine the interrelated ones to construct new input variables where necessary. This is further explained in Section 4.1.

The LCA is based on our earlier work[11] where we analysed a wide range of built form indicators (namely area type, population density, walk time to bus stop and rail station, and bus frequency) and socioeconomic characteristics from the individual level National Travel Survey (NTS) data and identified five distinct travel clusters within which exist more homogenous land use patterns and travel attitudes. Based on their characteristics, we labelled the clusters as: L1 - Metropolitan

[6] Under pillar 2 testing route, swab testing is conducted for the wider population as set out in the UK government guidance.

[7] PHE has confirmed that data included home tests through lateral flow devices (LFDs) alongside lab results.

core dwellers (mainly Inner and Central London), L2 - outer metropolitan dwellers (mainly outer London and Metropolitan areas), L3 - suburban dwellers, L4- exurban dwellers and L5- rural dwellers (Figure 2(a) presents the geographical distribution of travel clusters).

We developed separate models for each travel cluster (multigroup modelling approach) to account for heterogeneities across geographical areas. First, through multigroup approach, we can account for potential self-selection and spatial sorting effects reflected in the tendency of socioeconomic and demographic groups to reside in residential areas based on their land use, travel preferences, social structure, or social inequalities (e.g. as shown in Section 4.1, Asian/ Asian British ethnicity group has a higher tendency to live in metropolitan cities). Through segmenting by travel clusters, we dismantle the effect of area types from ethnicity in evaluating their associated COVID-19 risks. Second, different travel clusters have potentially very different associated risks of infection; for instance, the COVID-19 risk factors in dense urbanised areas with larger levels of mobility and more diverse socioeconomic and demographic profile of residents (e.g. London) are likely to be very different from those in small/medium urban or rural areas.

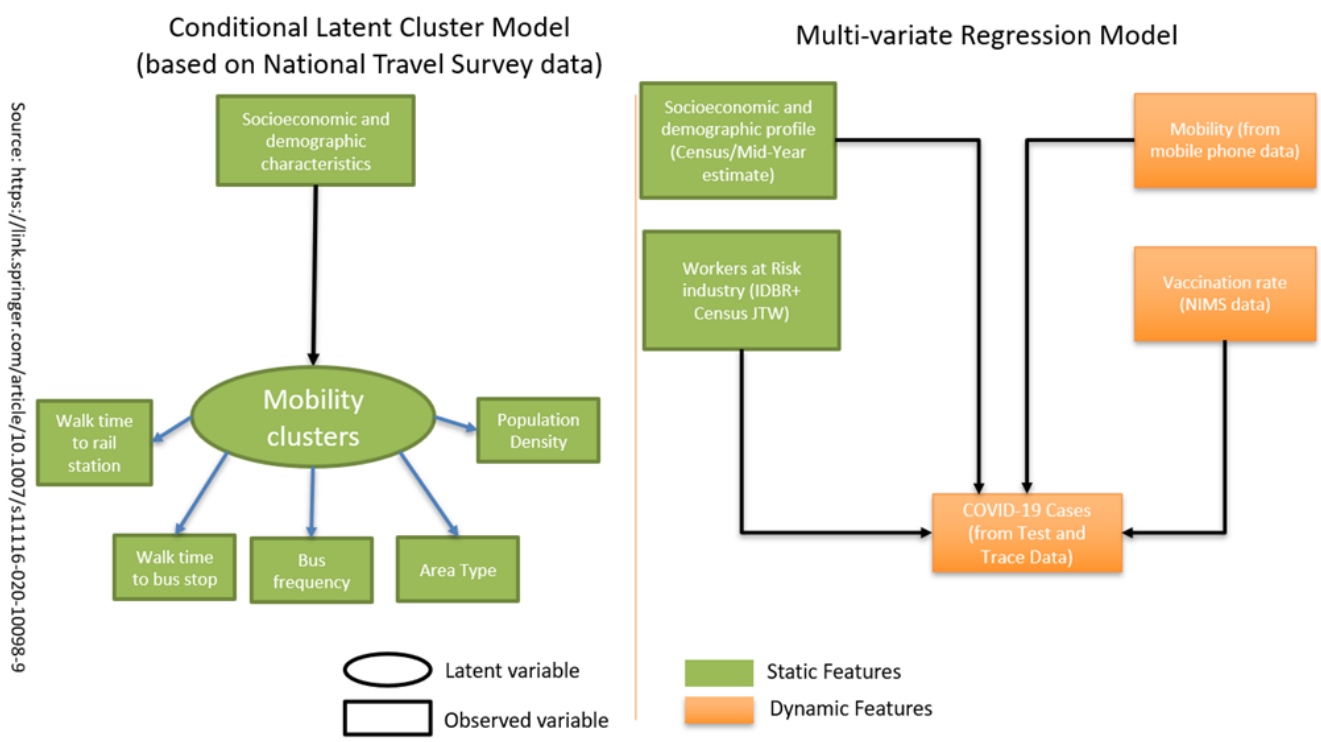


Figure 1. Conceptual model

For the rest of this section, we first provide the mathematical specification of the conditional LCA with built form indicators and socioeconomic and demographic characteristics (readers are referred to our earlier work[11] for more comprehensive information on how clusters are defined and capture distinct travel patterns). We then provide formulation for exploratory factor analysis and multivariate linear regression which are used in our multigroup modelling framework.

3.1 Conditional Latent Cluster Analysis

Conditional LCA involves allocating individuals to distinct clusters in the way that it maximises similarity within clusters and the differences between clusters based on individuals' residential built environment characteristics and conditional on their socioeconomic and demographic profile.

To formulate, let χ_{ij} be the jth indicator variable (e.g.. population density, area type, etc) of the travel cluster, C_i for individual i. As all our indicators are ordered categorical variables, we can formulate the link function by defining an underlying continuous variable, χ^*_{ij} such that

$$\mathscr{X}_{ij} = s|C_i = k \Leftrightarrow = \tau_{kj,s} < \mathscr{X}^*_{ij} < \tau_{kj,s+1} \tag{2}$$

where C_i, our travel cluster variable, takes values 1,...,k and τ are a set of threshold parameters. Conditional on regressors X

(i.e. socioeconomic characteristics [8] in our case) we can then present the link function as:

$$\mathscr{X}_{ij}^{*}|C_i = k, x_i = v_{k,j} + K_{kj}X_i + \varepsilon_{ij}. \tag{3}$$

The normal distribution assumption for ε_{ij} is equivalent to a probit regression for categorical variable χ_{ij} on X_i with the following probability function:

$$\Pr(\mathscr{X}_{ij} = s|c_i = k) = \Phi[\tau_{kj,s+1} - v_{k,j} - K_{kj}X_i] - \Phi[\tau_{kj,s} - v_{k,j} - K_{kj}X_i]. \tag{4}$$

Finally, the travel cluster membership probability conditional on X is given by multinomial logistic regression with the following formula:

$$\Pr(C_i = k|X_i) = \frac{\exp(\alpha_k + \gamma_k X_i)}{\sum_{s=1}^{k} \exp(\alpha_s + \gamma_s X_i)}. \tag{5}$$

3.2 Segmenting by time periods

In addition to splitting by travel clusters (spatial segmentation), we also split the time series into segments to represent different policy interventions (NPI regimes) and assess the variations in influences over time[2]. This also helps account for the heterogeneity from external influences (e.g. the new variant of virus). Each time tranche corresponds to a period defined by a specific external influence and NPI and allows us to model our data more homogeneously. The distinct tranches under exploration are shown in Table 3.

Table 3. Distinct time tranches alongside the main external influence dominating a specific tranche.

Tranche	Period	External influences during different stages of the pandemic
1	2020-05-03 to 2020-08-30	Low prevalence; schools closed; Alpha and Delta variants not yet emerged; no vaccine available.
2	2020-09-06 to 2020-11-08	High prevalence; schools open; negligible Alpha variant; Delta variant not yet emerged; no vaccine available.
3	2020-11-15 to 2020-12-27	High prevalence; schools open; Alpha variant becomes dominant; Delta variant not yet emerged; negligible vaccine coverage.
4	2021-01-03 to 2021-02-14	High prevalence; schools closed (except for pre-school); Alpha variant dominant; Delta variant not emerged yet; over 10 million first vaccine doses by the end of the time period.
5	2021-02-21 to 2021-04-25	Low prevalence; schools open; Delta variant negligible; over 35 million first and 15 million second vaccine doses by the end of time period.
6	2021-05-02 to 2021-07-11	High prevalence; schools open; Delta variant becomes dominant; over 45 million first and 35 million second doses administered by the end of the time period.
7	2021-07-18 to 2021-12-05	Lifting of almost all lockdown restrictions in England and before the Omicron variant became dominant.

3.3 Exploratory Factor Analysis

We employed exploratory factor analysis (EFA) on our input features in order to explore and account for potential spatial interactions among predictors (e.g. the likelihood of workers in textile and ready meat industries to live near each other). EFA is a statistical method[22] to reduce dimensionality of our features through estimating the minimum number of unobserved factors that represent the observed variables. In other words, EFA describes variability among observed correlated variables in terms of a potentially lower number of unobserved variables called factors. In our case, we use EFA to evaluate the interaction between our features and decide how to account for those as inputs to our model. To formulate, we consider each observable variable

[8] The comprehensive list of socioeconomic and demographic features used from the NTS data are provided in Jahanshahi and Ying, 2021[11]

(factor indicator) X_i as a linear function of independent factors and error terms which can be written as

$$X_i = \zeta_{i0} + \sum_k \zeta_{ik} F_k + e_i \tag{6}$$

In the equation above ζ_{i0} is the bias term and the error terms e_i, serve to indicate that the hypothesised relationships are not exact. In the vocabulary of factor analysis, the parameters ζ_{ik} is referred to as factor loading of variable X_i on factor F_k. It can be shown that the variance of X_i consists of two parts

$$\text{var}(X_i) = \underbrace{\zeta_{i1}^2 + \zeta_{i2}^2 \ldots}_{\text{communality}} + \sigma_i^2 \tag{7}$$

The first part, the communality of the variable, is the part that is explained by the common factors $F_k (k = 1, 2, \ldots N)$. The second part, the specific variance, is the part of the variance of X_i that is not accounted for by the common factors. One of the prime goals in exploratory factor analysis is to determine a set of loadings which bring the estimate of the total communality as close as possible to the total of the observed variances.

To perform EFAs on our dataset, we use Python's FactorAnalyzer package[23] and use varimax rotation on the dataset. We choose a loading threshold[24] of 0.35 to obtain the features contributing most to each individual factor. Factor analysis allowed us to gain invaluable insights into the dataset and create new features to account for spatial interactions in inferring major risk factors of COVID-19 infection at the community level. Details of the results of the factor analysis are discussed in Section 4.1.

3.4 Multigroup Multivariate linear regression at each time tranche

For each travel cluster and time tranche, we fit a multivariate linear regression model[22]. One can use mean square error (MSE) which is simply the sum of squared errors between our estimated and the true observations to quantify goodness of fit. The Least Squares regression model is the model that minimises the squared distance between the model and the observed data estimated by the following cost function:

$$J(\beta) = \frac{1}{2N} \sum_{i=1}^{N} ((\beta_0 + \beta_1 X_1^i + \beta_2 X_2^i + \ldots + \beta_m X_m^i) - \tilde{Y}_i)^2 \tag{8}$$

In above, β_0 denotes the intercept (in our case this represents the infection risk for the reference population for each time tranche and travel cluster) and β_i denotes the coefficient for the ith feature. The positive sign of β reflects positive correlation of the target variable and the respective predictor; the reverse directional effect is true for the negative sign. Linear regression outputs are relatively easy to interpret and under the assumption of independent and identically distributed random variables, the OLS (Ordinary Least Squares) estimators of the linear regression model are unbiased[22] and can yield useful insights into the behaviour of outcome variables.

We tested linear regression with and without regularisation and found that Linear regression approach with and without regularisation provided qualitatively similar results. Including regularisation, although it helps avoid overfitting in forecasting, produces somewhat biased estimation as it adds an extra term to the cost function. Therefore, we decided to use linear regression without regularisation to report unbiased significant risk predictors and associated p-values. Linear regression with added regularisation was used to predict area adjusted COVID-19 cases on unseen test data where unbiased estimates of the coefficients is not of primary interest.

4 Model outputs and findings

This section presents the main findings from our analysis including the major risk factors and their evolvement over time. First, we highlight the findings from exploratory factor analysis (EFA) for static predictors; this helps group the input features where they are highly correlated to account for their spatial interactions; second, we present the distribution of input features across the five distinct travel clusters; this further affirm the use of multi-group approach (i.e. fitting separate model for each cluster) to account for spatial sorting and self-selection effects. Finally, we report the major influences over time and across travel clusters and evaluate the role of workplaces' industrial sectors after controlling for a wide range of static and dynamic features.

4.1 Exploratory factor analysis and spatial interactions of input features

Table 4 shows the factor loadings from EFA; we have only reported those with significant size- i.e. an absolute loading cut-off of 0.35. Factors in essence represent common communality across features and the associated factor loadings (which are ranged from -1 to 1) show the extent to which each feature has contributed to each factor. This means the features with large absolute

factor loadings for each factor tend to correlate with each other and the associated factor can represent their interrelations. For instance, in our analysis below, we can see relatively large factor loadings for the first factor for black, mixed and other ethnic groups and for commuting with public transport (all above 0.7). This shows strong spatial correlation amongst these features suggesting higher probability for these ethnic groups to live in places with relatively good public transport access for commuting (i.e. mainly London and metropolitan areas). A complete list of spatially correlated features inferred for each individual Factor is shown in Table 5.

Table 4. Results of exploratory factor analysis of the static features.

	Factor_1	Factor_2	Factor_3	Factor_4
census_2011_asian_british	0.447709	**	**	0.464808
non_motorised_to_work	**	**	**	**
public_transport_to_work	0.850881	**	**	**
work_mainly_from_home	**	-0.353423	-0.486848	**
families_with_no_dependent_children	-0.385816	**	-0.720684	**
care	**	0.688389	**	**
meat_and_fish_processing	**	**	**	**
ready_meals	**	**	**	0.420955
textiles	**	**	**	0.503811
warehousing	**	0.397486	**	**
census_2011_mixed_multiple_ethnic_groups	0.839686	**	**	**
census_2011_other_ethnic_group	0.711826	**	**	**
census_2011_black_african_caribbean_black_british	0.783018	**	**	**

Table 5. Descriptions of the different Factors obtained after EFA.

Factors	Description
Factor_1	Strong spatial correlation amongst populations groups from Black (African, Caribbean and Black British), Mixed and Other ethnicities. Alongside strong (spatial) correlation with regions where the use of public transport to work by the residents is higher.
Factor_2	Spatial correlation amongst residents working in care homes, and warehouses.
Factor_3	Spatial correlation amongst regions with larger concentration of families with no children and larger proportion of the population working from home.
Factor_4	Spatial correlation amongst residents working in textiles and ready meals.

Based on the above inference from EFA as reported in Table 5, we make the following assumptions for our modelling stage:

- We use the proportion of the working population using public transport as the representative feature for Factor_1 and drop other highly correlated features comprising Factor_1 from our feature list. This is to avoid multicollinearity issues.

- We define a new feature capturing the total resident population working in care homes, and warehousing industries.
- We define a new feature capturing the total resident population working in ready meals and textiles industries.
- We incorporate residents working in meat and fish processing as independent features in our analysis.
- For each LSOA, we also include the proportions of families with no children, population working from home, population using non-motorised transport to work and British-Asian ethnic groups as independent predictors in our analysis.

The alternative approach would have been incorporating latent factors into our model instead of choosing one of the correlated features or combine those with each other; however, the latter that we have adopted made our findings more interpretable for the purpose of this paper.

4.2 Descriptive analysis of spatial distribution of input features

In this section, we present the geographical distribution of input features across the five identified travel clusters. Figure 2 compares the geographical distribution of travel clusters (Figure 2 (a)) with that of COVID-19 cases (Figure 2 (b)) over the period of analysis (i.e. from 4th October 2020- 5th December 2021). It can be observed that travel clusters are already accounting for a large proportion of variations in COVID-19 cases.

Figure 3 and Figure 4 show the distribution across spatial clusters of the static and dynamic features and suggests clear distinctions in socioeconomic and demographic as well as mobility patterns across travel clusters. This shows not only can travel clusters capture some of the spatial interactions but also it can help account for spatial sorting and self-selection of socioeconomic and demographic groups. In other words, through segmenting by travel clusters, our multigroup analysis allow comparing like with like.

In particular, we can infer the following from Figure 3:

- More dense urbanised areas have the highest proportion of people from an Asian background
- Use of public transport and non-motorised transport to work is also highest amongst the dense urbanised areas.
- Proportion of the working population more likely to work from home is highest amongst more rural areas.
- Proportion of households with no children is comparable across urban and rural areas.
- Resident population working in high-risk industries is highest amongst the most densely populated urban areas except for workers in meat and fish processing who have slightly higher tendency to live in medium and smaller urban areas but not much in rural areas.

With respect to dynamic features, Figure 4 shows that more dense urbanised areas also have the highest average footfall of workers and visitors for each distinct time tranche (refer to Table 3 for time tranche definitions). In addition, Figure 5 shows the cumulative proportion of population administered with first and second doses of COVID-19 vaccinations for each travel cluster and successive time tranches [9]; this suggests that vaccination uptake and rates is higher in more rural areas compared to more dense urbanised areas. This probably is due to the difference in age profiles of residents in urban and rural areas with the latter tend to have older population who got priority in the vaccination roll out.

The above analyses of spatial distribution of static and dynamic predictors show the importance of adopting a multi-group approach; analysing more homogenous data at each travel clusters better captures linear impacts when interrelations with geography is controlled. Through accounting for residential location characteristics and associated travel patterns, we can move towards making causal inference after controlling for externalities from self-selection and spatial sorting. For instance, without segmenting by travel clusters, in analysing those who work in meat and fish processing industries, we would not be able to distinguish the impact of living in dense urbanised areas from working in meat and fish processing as those who work in this industrial sector have also higher probability of living in dense urbanised areas. Segmenting by travel clusters, on the other hand, allows separating these impacts and compare the risk of working in meat and fish processing across different land use settings.

[9] The source of vaccination data (NIMS) and population at LSOA (MYE) are different and hence the proportion of those vaccinated (vaccination normalised by population) we used for this paper might be subject to some level of biases. However, we believe that the spatial distribution of those across LSOAs (relative values), which is the major input to our model, should be more robust and suitable for our analysis. We also use the changes from first dose uptake to the second as our input feature which can further help cancelling out some of the potential inherent biases in data.

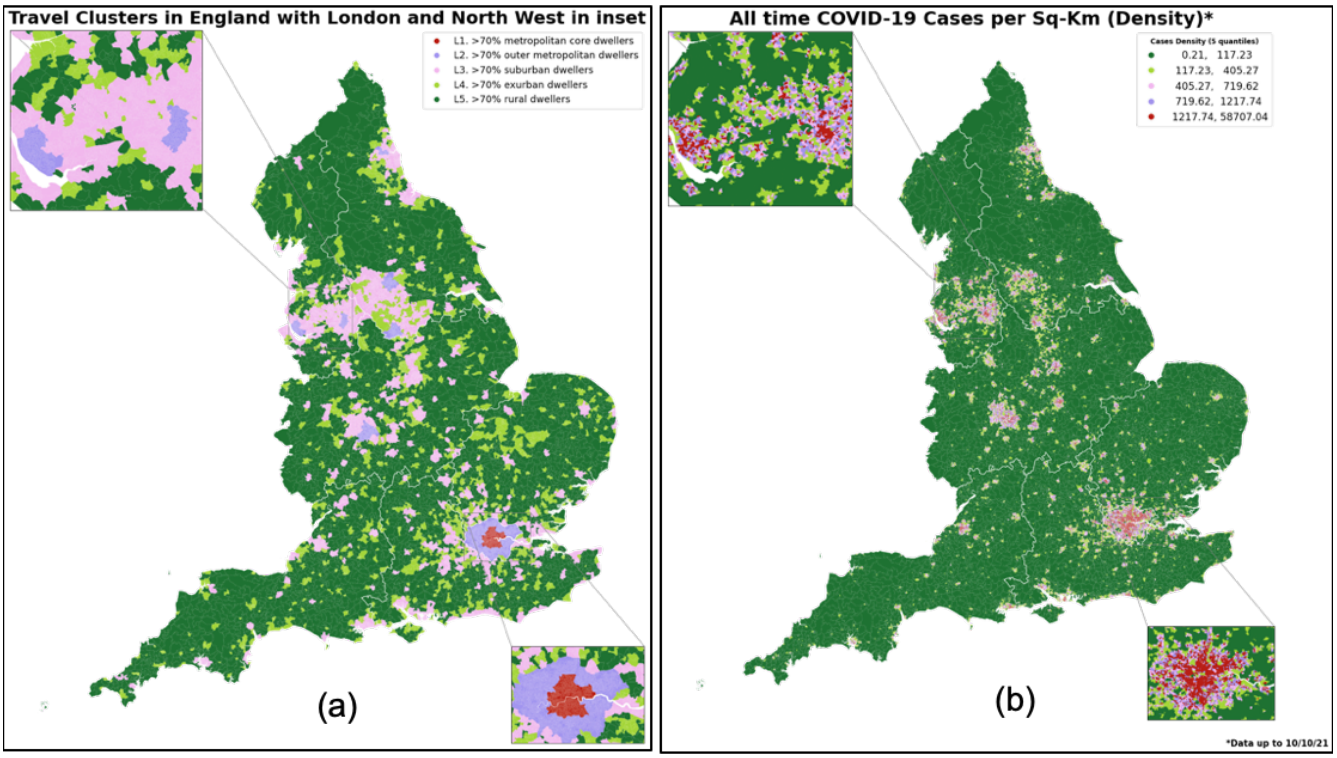


Figure 2. (a) Spatial distribution of latent travel clusters (b) spatial distribution of density of COVID-19 cases over the period of the analysis.

4.3 Identifying Risk factors

Table 6 presents the findings from multi-group linear regression for each travel cluster and time tranche. Non-standardised coefficients, measured in their original scales, of static and dynamic influences in addition to their level of significance are reported. Non-standardised coefficients show the scale of change in a dependent variable for one unit of change in an independent variable keeping all other independent variables constant. For instance, the value of 31.6 in Table 6 below shows that in the second time tranche for travel cluster 1, when all other influences remain constant, one unit increase in the proportion of Asian and Asian British ethnicity group results in 31.6 unit increase in the number of positive COVID-19 cases per square km. Non-standardised coefficients obtained using a multi-group approach allows us to assess relative risk of infection for different travel clusters. For example, it can be concluded from Table 6 that the relative risk of infection in areas with a large proportion of workers in high-risk industries is highest in most dense urbanised areas. However, not all features in our model are measured on the same scale and we need to use the standardised version of their coefficients to compare different features with each other. In terms of standardised coefficients, a change of one standard deviation in the predictor is associated with a change in the standard deviations of the dependent variable with the magnitude of the corresponding standardised coefficient. Standardised risk influences driving transmissions during the most recent time tranche (covering the period since the lifting of lockdown restrictions and before the Omicron variant became dominant in England- between the 18th July 2021 and the 5th December 2021) is shown in Figure 6. Standardised coefficients for other travel clusters and time tranche are presented in Table B.1. This allows us to additionally compare relative importance of different (now unitless) predictors within a given travel cluster and time tranche.

From Figure 6, it is relatively straightforward to assess risk factors for each travel cluster for the latest period reported in this paper (i.e. between 18th of July 2021 to 5th of December 2021). The most profound finding is that areas with a higher proportion of residents working in high-risk industries are among those most at the risk of infection. This includes both residents working in care homes and warehouses as well as those working in ready meals and textile industries. Amongst all the risk influences, areas with residents working in high-risk industries were the dominant risk factor in both the dense urbanised and rural areas (see light blue bars in Figure 6). This is also the case across all other time tranches (refer to Table B.1 in Appendix B which shows standardised coefficients across all time tranches). The fact that the relative risk of infection in high-risk industries is significant in all travel clusters suggests that part of the risk stems from workplace or job requirements. We can also make the following observations:

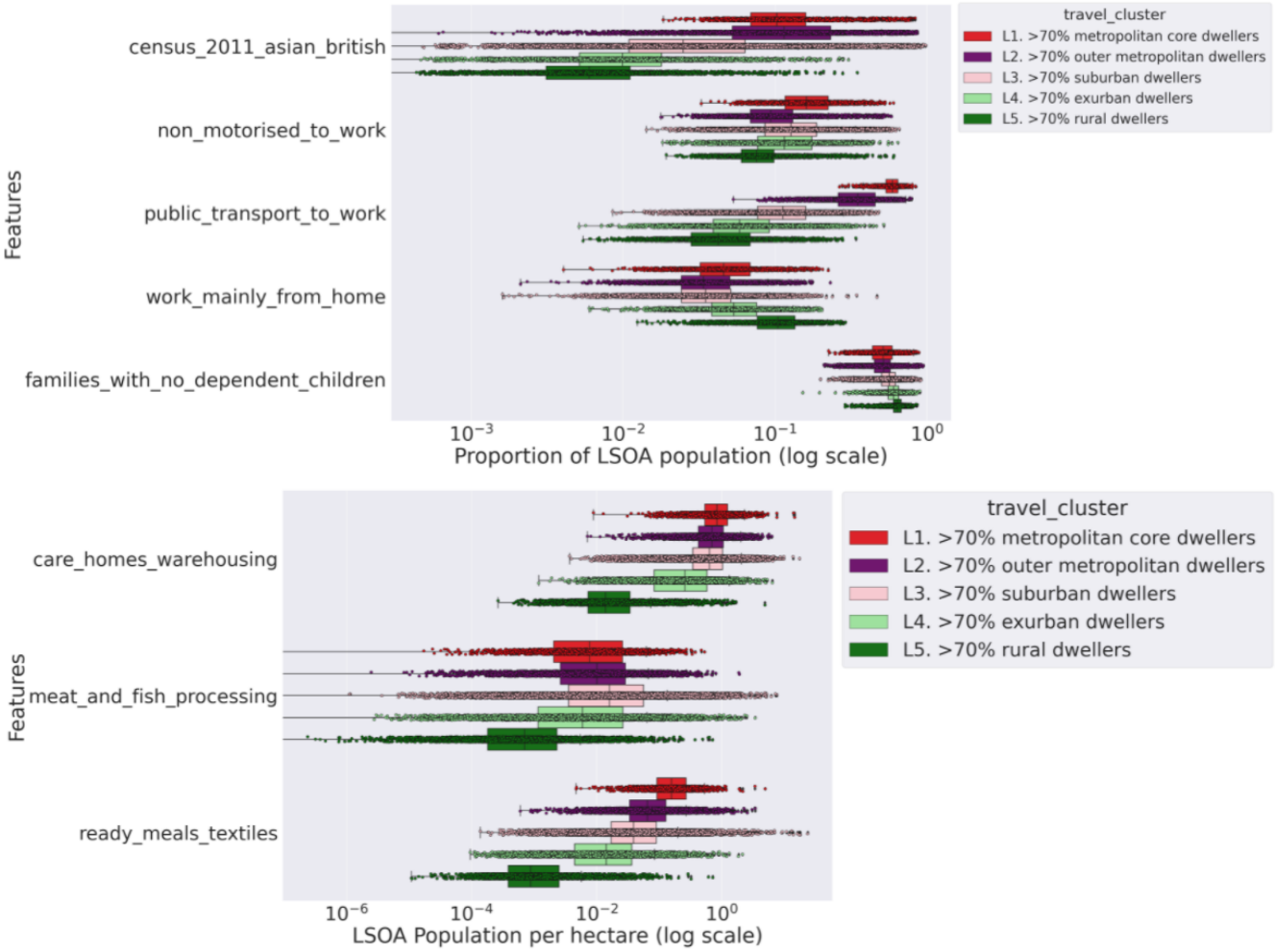


Figure 3. Distribution of the static predictors used in the analysis across distinct travel clusters.

- Areas with smaller family sizes with no dependent children are less at risk of infections compared to those with a higher proportion of larger families with children. This is found to be true for suburban, exurban and rural dwellers for the most recent time tranche. An exception to this were metropolitan dwellers (inner and central London), where areas with a greater proportion of smaller family sizes with no dependent children were found to be at an increased risk of infections for the most recent time tranches (6,7) potentially associated with role of mobilities and greater mixing in London amongst its younger population.

- Areas with a higher proportion of public transport users and non-motorised commuters are more at risk. The fact that this is the case in different travel clusters (both densely and sparsely populated areas) means the risk is more likely associated with public transport and not only limited to land use features of the areas. This risk also tends to increase in most recent time tranches (as mobility increases and restrictions lifted) as can be evidenced from Appendix C.

- Areas with higher tendency to work from home are associated with lower risk of infection for more rural areas. For central and inner London (and for outer London and suburban dwellers since the lifting of lockdown restrictions), a higher proportion of the working population working from home is not significantly associated with reduced risk of infection (specifically in most recent time tranches aligned with ease in restrictions). This can be associated with a complex interplay between other purposes and forms of mobilities in more dense urbanised areas, change in human behaviour post vaccination roll-out and greater mixing following lifting of lockdown restrictions.

- As expected, in all travel clusters (land use settlements), increase in visiting and working footfall is significantly associated with higher risk of infection. This is found to be true for each individual time tranche explored in this study.

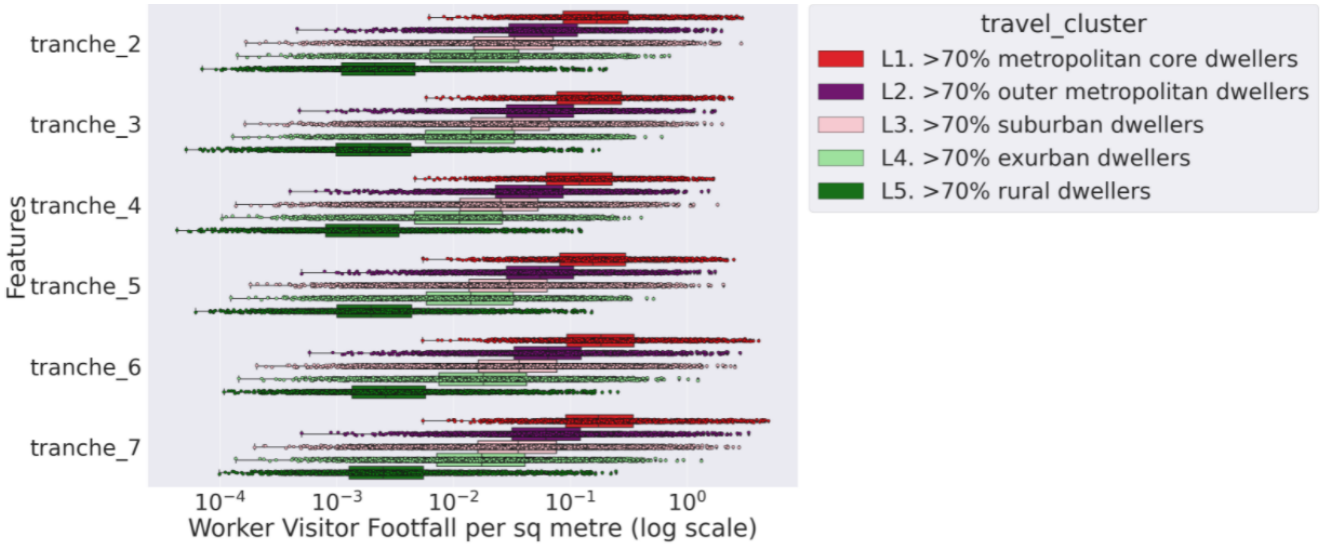


Figure 4. Distribution of the worker and visitor footfall across distinct travel clusters for distinct time tranches studied under this investigation **.

** Note: the data sharing agreement with our Mobility footfall data provider does not include the period under tranche 1.

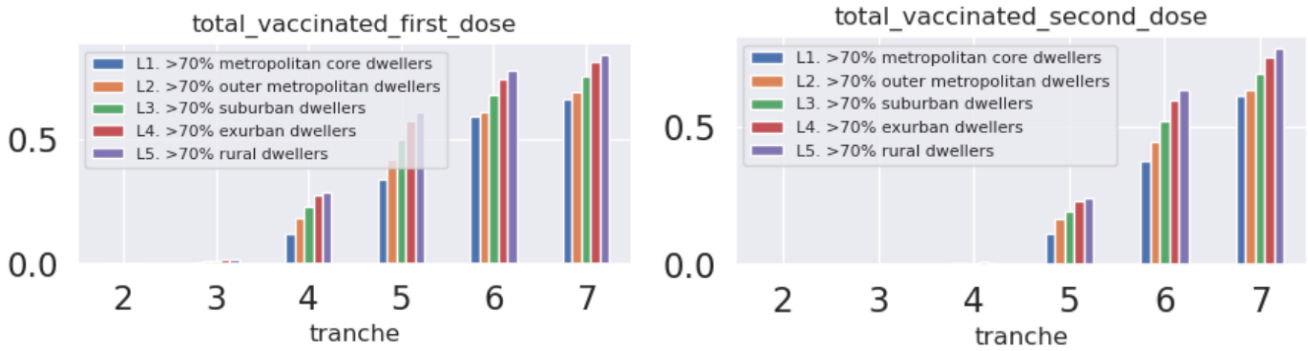


Figure 5. Proportion of population administered with first and second doses of COVID-19 vaccinations for each travel cluster.

- Areas with higher rate of administering second dose of COVID-19 vaccinations (relative to the first dose of COVID-19 infection) are associated with lower risk of infection. Since the lifting of lockdown restrictions in England, this has been found to be statistically significant for the majority of population living in suburban, exurban and rural dwellers (travel clusters L3, L4, L5) which have the highest proportion of fully vaccinated populations in England by the date of this study.

4.4 Variability of Risk predictors

This section evaluates the extent to which the identified risk factors have been stable over time. We can check the coefficient variability through $k-$fold cross-validation which is a form of data perturbation. $k-$fold divides the data into k non-overlapping parts (called folds), hold out one of these folds and use the remaining folds $(k-1)$ to train a model[25]. This process is repeated across all the folds until all the data has been used. To reduce potential biases, one can also repeat the $k-$fold cross-validation procedure multiple times and report the results across all folds from all runs.

Figure 7 reports the regression coefficients obtained for all folds from all runs. If estimated coefficients vary significantly when changing the input dataset their robustness is not guaranteed and they should probably be interpreted with caution. In our case, in addition to checking the robustness of results for specific time tranches (refer to Figure 7a which shows the stability of both static and dynamic influences for the latest time tranche), we also use this technique to evaluate the variations in static influences across the whole study period within each travel cluster (refer to Figure 7b[10]). The latter shows the extent to which

[10]Since the dynamic influences (vaccination and footfall) tend to vary over time, the analysis of the stability of coefficients over all time tranches is done

Table 6. Non-Standardised risk predictors influencing infections in England covering the period 4th October 2020- 5th December 2021 $***p < 0.01, **p < 0.05, *p < 0.1$.

travel_cluster	tranche	const	asian	non_motorised	public_transport	wfh	fam_no_dep_chdrn	care_homes_warehousing	meat_and_fish	ready_meals_textiles	worker_visitor_footfall	vaccine_2_mins_1
L1	2	-18.55*	31.64***	62.88***	20.88	71.15**	-20.42**	7.03***	-8.7	52.53***	10.2***	17180.0***
	3	-41.83***	121.74***	101.57***	55.97**	32.48	-41.88***	24.27***	25.08	115.31***	14.28***	-268.61*
	4	-77.0**	162.22***	146.73***	104.81***	66.98	-160.12***	65.38***	-222.37***	252.33***	55.55***	-268.47***
	5	-9.26**	14.9***	20.5***	12.19***	19.38**	-4.97*	5.02***	-7.12	3.49**	1.51*	-0.55
	6	-57.03***	0.62	70.92***	51.03***	81.88***	12.43***	9.84***	-1.58	28.65***	0.2	-54.82***
	7	-68.24***	15.1**	105.65***	74.12***	165.14***	20.27**	25.63***	-110.54***	99.04***	1.7	64.34
L2	2	-26.58***	14.86***	227.16***	-21.02***	-90.35***	32.24***	16.37***	33.75***	-1.01	0.72	-3654.09
	3	-3.22*	27.08***	-11.49***	46.48***	-25.41**	-12.67***	20.18***	23.59***	2.68*	4.23*	-42.09
	4	-1.76	53.55***	-28.57***	77.64***	-91.37***	-36.83***	40.9***	13.44**	27.99***	26.2***	-5.5
	5	6.15***	7.47***	14.61***	-3.73***	-8.03**	-9.77***	4.26***	12.83***	2.38***	-0.7	2.09
	6	-15.65***	-2.0	67.25***	-0.15	-5.71	2.43	8.99***	38.7***	-1.32	-5.35***	-68.32***
	7	-6.4	-1.55	40.75***	22.64***	56.24***	-2.75	21.73***	25.78***	3.93	6.13**	-24.88
L3	2	-8.86***	38.2***	15.58***	28.89***	-55.31***	12.89***	6.33***	6.27***	6.18***	17.59***	683.68
	3	2.34***	28.31***	3.63***	30.11***	-10.94***	-7.63***	9.92***	2.88***	6.65***	5.71***	-30.56***
	4	1.91*	33.42***	13.89***	60.25***	-31.78***	-13.64***	18.07***	-0.84*	6.25***	8.57***	-7.57***
	5	2.29***	12.39***	0.59	5.21***	-16.46***	-2.88***	2.9***	2.19***	2.6***	-0.46	1.22**
	6	-7.02***	7.94***	9.51***	21.45***	-2.13	2.89***	5.6***	1.92***	1.86***	2.8***	-18.99***
	7	5.2***	-12.12***	26.4***	32.27***	18.71***	-13.19***	16.6***	5.23***	3.97***	4.42***	-28.64***
L4	2	3.4***	11.21***	-6.38***	-1.91*	-18.92***	-1.09	3.69***	0.77*	7.45***	10.11***	-273.03
	3	4.69***	9.72**	-4.06***	25.89***	-28.34***	-5.21***	6.17***	-0.11	7.2***	11.75***	-8.92**
	4	6.67***	27.29***	0.43	47.98***	-43.79***	-15.75***	11.72***	6.98***	9.7***	30.97***	-13.32***
	5	1.39***	4.44***	-0.91***	1.33***	-7.54***	-1.08***	1.6***	1.44***	2.91***	4.86***	0.26
	6	2.42***	13.27***	-2.25***	2.96***	-7.68***	-2.58***	2.98***	1.7***	4.66***	6.14***	-0.04
	7	10.5***	-2.65	3.3**	7.73***	-16.23***	-14.83***	14.49***	4.17***	14.81***	21.35***	-29.77***
L5	2	0.19	2.82***	-0.3	1.82***	-2.72***	0.17	2.94***	4.82***	11.67***	8.29***	-177.49
	3	1.33***	5.07***	-0.89**	3.91***	-3.05***	-1.63***	7.91***	-2.66**	8.22***	16.05***	-1.33
	4	2.36***	7.94***	-0.75	5.06***	-4.54***	-3.52***	10.55***	2.84**	13.88***	35.12***	-1.29**
	5	0.26**	1.46***	-0.25	0.52**	-0.3	-0.41**	1.21***	7.42***	10.18***	1.67*	-0.06
	6	0.56***	0.3	-0.28*	0.96***	-1.69***	-0.64***	2.92***	0.73*	0.88*	4.74***	-0.63***
	7	0.63*	-1.64	0.1	1.36*	-4.75***	-1.96***	15.63***	9.27***	19.27***	29.19***	-24.91***

NPIs, virus variants and other external factors might have affected the direction of impact and the relative importance of static influences. For instance, we can evaluate whether in-risk industries have had different direction of impacts on risk of infection in different time periods or whether their relative importance compared to other influences has changed over time, say due to adoption of certain policies.

Figure 7a shows the outcome of stability analysis of both the static and dynamic influences for the last time tranche 7. Tranche 7 covers the time period of the lifting of lockdown restrictions in England (increased mobility) and also when the cumulative coverage of the proportion of fully vaccinated population is highest across all travel clusters. The results suggest fairly stable and robust patterns of influences aligned with the findings presented in Figure 6.

Figure 7b shows the results of this cross validation exercise for the static risk influences across all time periods combined. It is interesting to note that static influences are relatively robust to temporal variations when modelled at geographically aggregate level and segmented by travel clusters. For instance, high-risk industries have stayed highly significant influence across all time periods modelled and travel clusters with positive impact on infection risk[11]. This suggests that policy interventions, although might have controlled the total level of infections across all communities, have not shown much influences on the relative risks specifically for the vulnerable communities working in high-risk industries.

In summary, we make the following observations from stability analysis at all modelled time periods:

- Large density of residents working in high-risk industries are a positive risk factor of infections in all land use settings.
- Larger density of smaller households with no dependent children are a negative risk factor for infections.
- Larger density of public transport and non-motorised users for commuting is a positive risk factor of infections for most dense urbanised areas.

only for static features.

[11] The exception is the influence of meat and fish processing in Central and Inner London cluster which can be explained by the small sample size of workers in meat and fish processing in Central London.

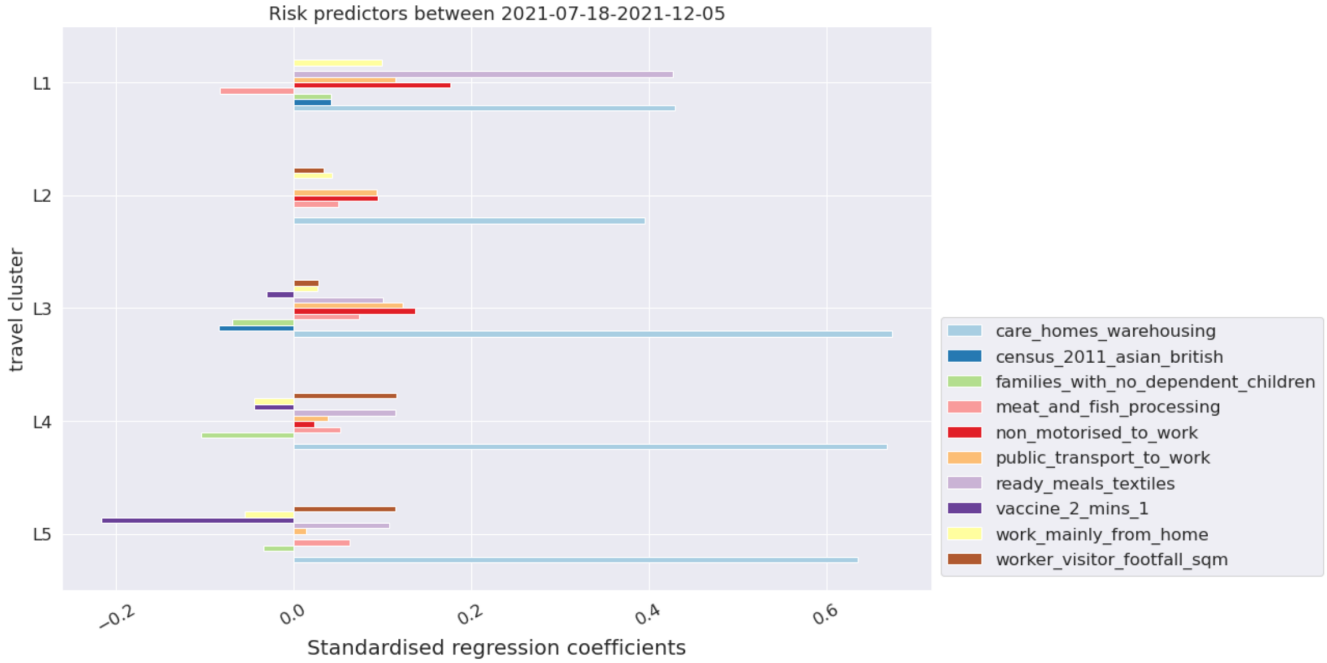


Figure 6. Significant risk predictors influencing infections in England since the lifting of lockdown restrictions for the period 18th July 2021-5th December 2021. The coefficients are standardised and $*p < 0.1$.

- Residential areas with high density of those who tend to work from home are negatively associated with the risk of infections except for travel clusters L1 (Inner and Central London).

4.5 Model performance

We use R^2 score to evaluate the performance of our model following the training on the static and dynamic features for each tranche. The R^2 score explains the dispersion of errors of a given dataset and can be used to measure the discrepancy between a model and actual data[22]. Scores close to 1.0 are highly desired, indicating better squares of standard deviations of errors.

Figure 8 shows the R^2 score for the fitted estimator for each time tranche of training and for every travel cluster. It is worth mentioning that the performance of our fitted estimators can be potentially improved by including additional features in our modelling including the antibodies datasets which have not been available to us. It can be seen from Figure 8 that fitted estimators are better at capturing the spatial-temporal distribution of cases for less densely populated and rural areas (highest average R^2 score).

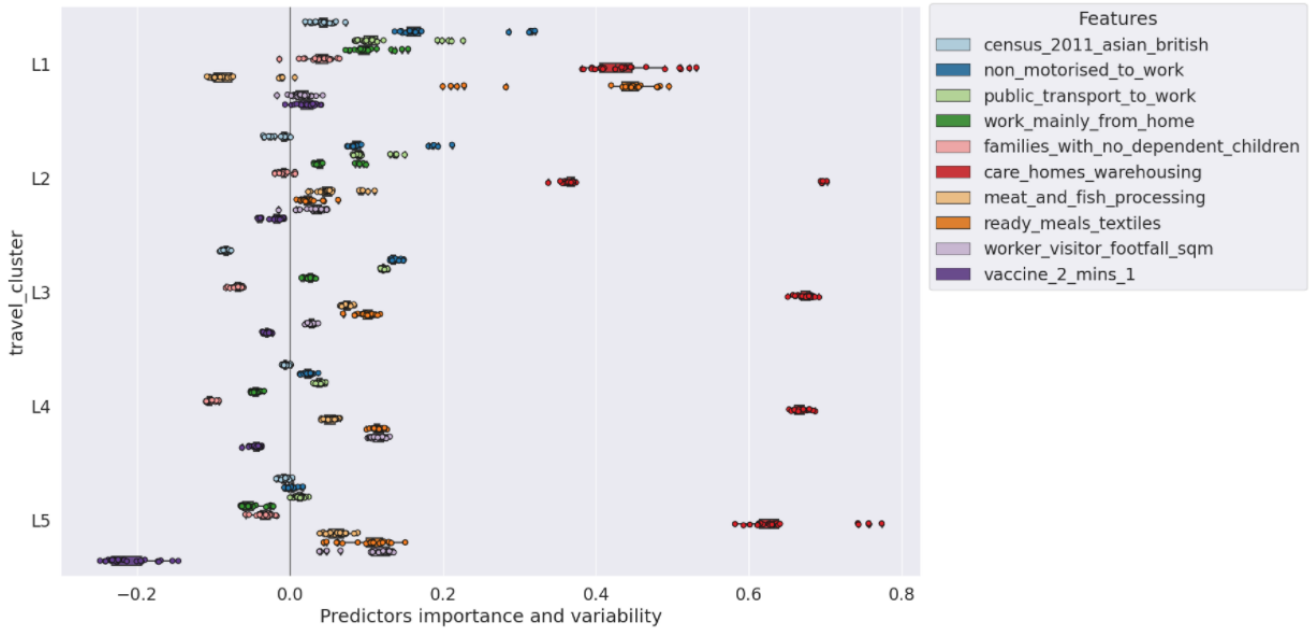
5 Limitation, mitigation, and future plan

The analyses above provide some in-depth insights into the most important COVID-19 risk parameters at a granular spatial level and their spatiotemporal variations. To the best of our knowledge, this is the first study of its kind where a range of complementary datasets from various sources including vaccination and telecoms data have been used to understand the influences on COVID-19 risk.

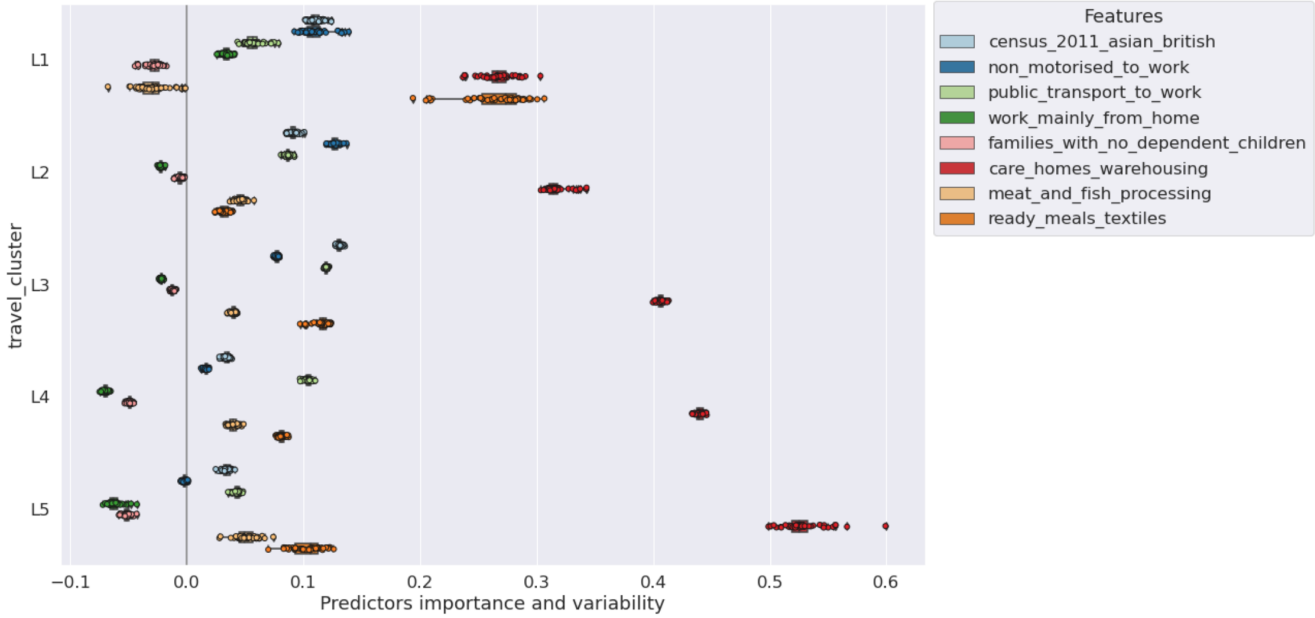
Like all studies of this type, however, we had some limitations, the potential effects of which are discussed in this section. Here, we also discuss some of the potential avenues of further explorations towards expansion and implementation of the current work as a comprehensive framework of an early warning system at LSOA level for the UK.

One of the major limitations of our study is the potential ascertainment bias[26] specifically in self-reporting datasets such as the test and trace. Our target variable can suffer from ascertainment bias despite wider testing available since the summer 2020. Recent studies[27] have reported on the idea that using randomised testing schemes, such as the REACT study in the UK, can be used to debias fine-scale targeted testing data in order to provide accurate localised estimates of the number of infectious individuals.

Our modelling approach, however, could help accounting for some of these biases; first, using aggregated data means we focus on spatial (LSOA level) variations in cases rather than individual level information. This would make analysis less sensitive to biases arising from asymptomatic diseases which are likely to be underreported. In analysing spatial distribution of



(a) Static and dynamic risk predictors variability for different travel clusters for the period 18th July 2021-5th December 2021 covering the period following the lifting of lockdown restrictions in England.



(b) Static risk predictors variability for different travel clusters for the period 4th October 2020-5th December 2021 covering all time tranches.

Figure 7. Variability of risk predictors used in our model.

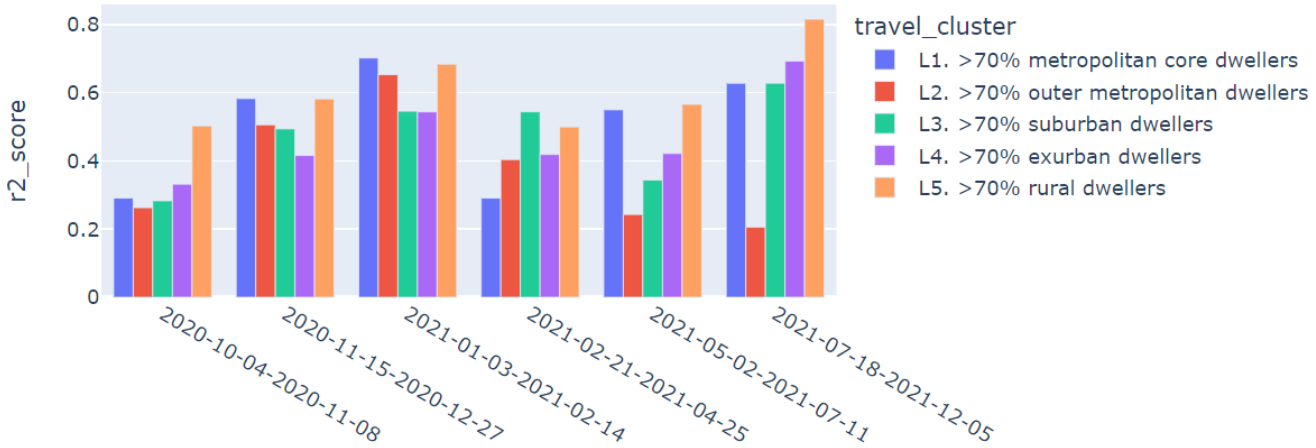


Figure 8. Goodness of fit captured through R^2 score using the multi-group regression analysis for different travel clusters encapsulating different stages of the pandemic.

disease, one can assume high spatial correlation between asymptomatic COVID-19 cases and symptomatic COVID-19 ones making the spatial distribution of the latter a good representative of that for the former. Second, multi-group analysis adopted in this paper can help make the biases more randomised. We expect that the under or over representation of COVID-19 cases are correlated with land use features, area types and associated socioeconomic and demographic patterns; our segmentation by travel clusters can account for this as biases within each travel cluster are more likely to be random and equally distrusted across areas. That is also true for vaccination and mobility features. Despite above methodological considerations, a detailed analysis to correct for potential biases in our input dataset can be an interesting avenue for future exploration.

One potential extension of our current framework could be to incorporate additional dynamic data including the level of antibodies present in the population as well as additional segmentation by different virus variants in circulation. Risk incidents modelled here is also set to be expanded to include the risk of hospitalisation and mortality in addition to that of infection. The latter can be modelled as a set of conditional dependencies to postulate the probability of mortality conditional on hospitalisation and the probability of hospitalisation conditional on catching the disease (cases). One way to account for conditional dependencies (as discussed further below) is through a structural equation modelling framework which helps assess both direct and indirect influences.

Furthermore, to improve the model interpretability, which is the target of the current paper, we have adopted a multivariate linear regression and accounted for potential nonlinearity through further segmentation by travel clusters and time and consideration of interactive terms. The model predictability[12], however, can be potentially improved (specifically after incorporation of further dynamic data discussed above) by introducing more advanced statistical models such as Structural Equation Model (SEM)[28] to capture systematically interrelations between parameters (including conditional dependencies between cases, hospitalisation and mortality) and two-way fixed effects regression[29], which can help causal inferences as well as short-term forecasting. The structure and coding pipeline of the latter is currently developed and can be tested, as potential future avenue of exploration, if needed and the access to the additional dynamic data is secured. Two-way fixed effect involves two modelling steps: a) modelling COVID-19 prevalence only on the static features and b) estimating the impact of changes in dynamic features on those in residuals from step (a) (i.e. the remaining effects after removing those from static features). Estimating the residuals by dynamic features can be incorporated under a linear or non-linear approximation including through use of machine learning techniques such as LSTM [13]. This approach can facilitate short term and near real time risk inferences and forecasting.

6 Summary of key findings

The key findings of this work are highlighted below.

First, our work has shed light on assessing the impact of the pandemic on some of the most vulnerable sections of the population including those who work in high-risk industries, with relatively higher risk of exposure to infection. Our results, after controlling for real time mobility, vaccination and socioeconomic and demographic profiles, show that areas with a

[12] See Appendix D for a brief analysis on how our current formalism can be potentially used as an early warning detection framework for risk mitigation.
[13] Long short-term memory

larger proportion of residents working in care homes and warehouses and to a lesser extent ready meals and textile sectors are prone to higher risk of infection across all travel clusters and all time periods. Similar influences across all residential area clusters suggests the potential association with workplace risk and regulations which can be further examined in a more detailed (individual level) workplace outbreak analysis.

Second, the findings underline the critical importance of geographical variations in influences on COVID-19 prevalence (after controlling for mobility and vaccination rate). For instance, for the most recent time tranche covering the period of lifting of lockdown restrictions in England, areas with a bigger proportion of small families and fewer density of children are prone to lower risk of infection in medium and smaller urban and rural areas; this, however, is not a significant risk factor in central and inner London and metropolitan cities. This is also the case for areas comprising of a larger proportion of those who can work from home; while work from home is shown to reduce the infection risk in less populated and smaller cities, this is not the case in metropolitan core dwellers where people might be more active in a diverse set of activities apart from work.

Finally, except rural settlements, areas with residents who are more dependent on the use of public transport for commuting have also been identified with greater risk of infections across all travel clusters. Although the risk is lower in the fifth tranche of time when vaccination has started to take effect and the Delta variant has not yet become dominant, use of public transport has been one of the main risk factors in smaller and bigger urban areas alike.

Given the above, our spatially aggregated model is well suited to tailor the research questions for further investigation at more detailed and granular level (e.g. through epidemiological models at individual and household levels). For instance, following our finding about the critical importance of certain industries' workplace risk after controlling for the land use characteristics and mobility patterns at the residential area, further analysis can be tailored to evaluate safety measures and regulations for high-risk workplaces. Continuous evaluation of community level risk can identify new threats and risk patterns at an early stage when there is a better chance to respond.

7 Conclusion

In this work we have developed a LSOA level COVID-19 risk model by bringing together a variety of granular rich datasets at LSOA level, in order to provide risk estimates of COVID-19 incidence for neighbourhoods. Using LSOA level indicators encompassing population demographic, information on higher risk industries, housing conditions, urban/rural area classification, vaccination rate, and real time mobility patterns, we developed one of the most comprehensive dataset to date to model major COVID-19 risk factors at aggregate geographical level in England. To fuse learning from this detailed dataset and control for exogeneities arising from the highly interrelated influences, we adopted machine learning and econometric techniques in our analysis pipeline on Google Cloud Platform. We used Latent Cluster Analysis (LCA) to identify distinct travel clusters within which exist more homogeneous travel patterns and attitudes. Training the model for each built form (travel) cluster separately, we then applied multivariate regressions to gauge the static and dynamic influences and infer the major risk factors on COVID-19 incidents. Our model can be updated regularly and run in real time to uncover the most recent risk factors driving the COVID-19 cases and associated variations in geographical patterns of risk. For the purpose of this paper, however, we run the model for seven distinct time periods which can best reflect the virus variants and policy interventions. Our comprehensive identification of the risk factors affecting COVID-19 transmission may be useful to policymakers to aid them devise effective strategies for population groups most at risk and observe in real time the impact of global and federated policy interventions in mitigating the COVID-19 risk.

Acknowledgment

We would like to thank Dr Thanasis Anthopoulos, Senior Data Scientist at the Data Science Campus for reviewing and quality assurance of the models' code, Dr Louisa Nolan, Deputy Director at the Data Science Campus, and Dr Owen Daniel, Lead Data Scientist at the Data Science Campus for reviewing the manuscript and their very constructive and comprehensive comments. We also like to thank Alistair Calder and Chris Gale from the ONS Geospatial Analysis team for their support in data preparation of a wide range of static indicators tested or used in the model. Alongside, we want to thank colleagues from the Health and Pandemic Insights Team at ONS for their feedback and suggestions on the manuscript.

References

1. Steel, K. & Yapp, R. Office for national statistics. coronavirus (covid-19) infection survey: methods and further information. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation (2022).

2. House, T. & et. al. Inferring risks of coronavirus transmission from community household data. *arXiv e-prints:2104.04605* DOI: https://arxiv.org/abs/2104.04605 (2021).

3. Williamson, E. J. & et. al. Factors associated with covid-19-related death using opensafely. *Nature* **584**, 430–436, DOI: https://doi.org/10.1038/s41586-020-2521-4 (2020).

4. Katikireddi, S. V. & et. al. Unequal impact of the covid-19 crisis on minority ethnic groups: a framework for understanding and addressing inequalities. *J Epidemiol Community Heal.* **75**, 970–974, DOI: https://doi.org/10.1136/jech-2020-216061 (2021).

5. Raisi-Estabragh, Z. & et. al. Greater risk of severe covid-19 in black, asian and minority ethnic populations is not explained by cardiometabolic, socioeconomic or behavioural factors, or by 25 (oh)-vitamin d status: study of 1326 cases from the uk biobank. *J. Public Heal.* **42**, 451–460, DOI: https://doi.org/10.1093/pubmed/fdaa095 (2020).

6. Jin, J. & et. al. Individual and community-level risk for covid-19 mortality in the united states. *Nat. medicine* **27**, 264–269, DOI: https://doi.org/10.1038/s41591-020-01191-8 (2021).

7. Sze, S. & et. al. Ethnicity and clinical outcomes in covid-19: a systematic review and meta-analysis. *Nat. medicine* **29**, 100630, DOI: https://doi.org/10.1016/j.eclinm.2020.100630 (2020).

8. Khunti, K. & et. al. Is ethnicity linked to incidence or outcomes of covid-19? *BMJ* **369**, DOI: https://doi.org/10.1136/bmj.m1548 (2020).

9. Wang, X. Mapping and identifying community risk factors for covid-19 nursing home deaths. *Heal. Serv. Res.* **56**, 85–85, DOI: https://doi.org/10.1111/1475-6773.13841 (2021).

10. Yapp, R. & et. al. Office for national statistics. coronavirus (covid-19) infection survey technical article: analysis of populations in the uk by risk of testing positive for covid-19. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveytechnicalarticle/analysisofpopulationsintheukbyriskoftestingpositiveforcovid19september2021# (2021).

11. Jahanshahi, K. & Jin, Y. Identification and mapping of spatial variations in travel choices through combining structural equation modelling and latent class analysis: findings for great britain. *Transportation* **48**, 1329–1359, DOI: https://link.springer.com/article/10.1007/s11116-020-10098-9 (2021).

12. Office for national statistics. open geography portal. https://geoportal.statistics.gov.uk/.

13. English indices of deprivation, department for levelling up, housing and communities and ministry of housing, communities & local government. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (2019).

14. Full ons 2011 census data for different categories can be found on the ons website. https://www.ons.gov.uk/.

15. Office for national statistics. inter-departmental business register (idbr). https://www.ons.gov.uk/aboutus/whatwedo/paidservices/interdepartmentalbusinessregisteridbr.

16. Office for national statistics. classification of workplace zones for the uk methodology and variables. https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011workplacebasedareaclassification/classificationofworkplacezonesfortheukmethodologyandvariables (2011).

17. Office for national statistics. uk sic 2007. https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007 (2022).

18. Uk health security agency. covid-19 vaccine surveillance report. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1063023/Vaccine-surveillance-report-week-12.pdf (2022).

19. Kozlov, M. Waning covid-19 super-immunity raises questions about omicron. *Nature* **48**, 1329–1359, DOI: https://www.nature.com/articles/d41586-021-03674-1 (2021).

20. Goldberg, Y. & et. al. Protection and waning of natural and hybrid covid-19 immunity. *MedRxiv* **48**, 1329–1359, DOI: https://doi.org/10.1101/2021.12.04.21267114 (2021).

21. Pagel, C. & et. al. Tackling the pandemic with (biased) data. *Science* **374**, 403–404, DOI: https://www.science.org/doi/10.1126/science.abi6602 (2021).

22. Johnson, R. & Wichern, D. *Applied Multivariate Statistical Analysis (6th ed)* (Pearson, Ch. 9, 2007).

23. More details of the package can be found here. https://factor-analyzer.readthedocs.io/en/latest/index.html.

24. What thresholds should i use for factor loading cut-offs? https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/thresholds.

25. Geron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (OŔeilly Media, 2019).

26. E.J., G. & et. al. Estimation of ascertainment bias and its effect on power in clinical trials with time-to-event outcomes. *Stat Med.* **40**, 1306–1320, DOI: https://doi.org/10.1002/sim.8842 (2021).

27. G., N. & et. al. Local prevalence of transmissible sars-cov-2 infection: an integrative causal model for debiasing fine-scale targeted testing data. *medRxiv* DOI: https://doi.org/10.1101/2021.05.17.21256818 (2021).

28. Kline, R. *Principles and Practice of Structural Equation Modelling (2nd Edition ed.)* (New York: The Guilford Press, 2005).

29. Imai, K. & Kim, I. On the use of two-way fixed effects regression models for causal inference with panel data. *Polit. Analysis* **29**, 405–415, DOI: https://doi.org/10.1017/pan.2020.33 (2021).

A Appendix A

Table A.1. Data dictionary outlining all the static features used in the LSOA level risk model and the respective data sources.

Variable	Description	Components	Source
Ethnicity	This feature provides information on the percentage of people belonging to a specific ethnic group of the usual resident population within a lower super output area (LSOA).	White,Asian-British, Black-British, Mixed ethnicity and other ethnicities.	Census 2011.
Age	This feature provides information on the percentage of residents who belong to an age group within a LSOA and is used to estimate the LSOA population.	0-12;13-17;18-29;30-39;40-49; 50-54;55-59;60-64;65-69;70-74;75-79;80+.	2019 Mid-Year Population estimates.
Method of travel to work	This feature provides information on the percentage of LSOA residents who use different modes of transport to travel to work.	Public transport (e.g. buses, underground rail, metro, trams and taxi); Private Transport (Cars, motorbikes etc.); Non-Motorised transport (walking or cycling to work).	Census 2011.
Families (with and without children)	This feature provides 2011 estimates that classify families in households in England and Wales by the number of dependent children in the family. A dependent child is any person aged 0 to 15 in a household (whether or not in a family) or a person aged 16 to 18 who is in full-time education and living in a family with his or her parent(s) or grandparent(s). It does not include any people aged 16 to 18 who have a spouse, partner or child living in the household.	Family with dependent children; Family with no dependent children	Census 2011.

B Appendix B

Table B.1 below shows the standardised coefficients across all time tranches. Please note that one should compare coefficients at each row as they are standardised for each travel cluster and time tranche to help compare relative importance of each influences within every travel cluster and time tranche.

Table B.1. Standardised risk predictors influencing infections in England covering the period 4th October 2020- 5th December 2021 $***p < 0.01, **p < 0.05, *p < 0.1$.

travel_cluster	tranche	asian	non_motorised	public_transport	wfh	fam_no_dep_chdrn	care_homes_warehousing	meat_and_fish	ready_meals_textiles	worker_visitor_footfall	vaccine_2_mins_1
L1	2	0.13***	0.16***	0.05	0.06**	-0.06**	0.17***	-0.01	0.34***	0.1***	0.06***
	3	0.26***	0.13***	0.07**	0.02	-0.07***	0.32***	0.01	0.38***	0.06***	-0.03*
	4	0.18***	0.1***	0.06***	0.02	-0.13***	0.43***	-0.07***	0.43***	0.09***	-0.08***
	5	0.2***	0.17***	0.09***	0.06**	-0.05*	0.41***	-0.03	0.07**	0.04*	-0.0
	6	0.0	0.28***	0.19***	0.12***	0.06***	0.4***	-0.0	0.3***	0.0	-0.17***
	7	0.04**	0.18***	0.11***	0.1***	0.04**	0.43***	-0.08***	0.43***	0.02	0.02
L2	2	0.06***	0.4***	-0.06***	-0.05***	0.08***	0.22***	0.05***	-0.0	0.0	-0.01
	3	0.2***	-0.04***	0.26***	-0.03**	-0.05***	0.49***	0.06***	0.02*	0.02*	-0.02
	4	0.21***	-0.05***	0.23***	-0.05***	-0.08***	0.54***	0.02**	0.11***	0.06***	-0.01
	5	0.2***	0.17***	-0.08***	-0.03**	-0.15***	0.39***	0.13***	0.07***	-0.01	0.02
	6	-0.02	0.25***	-0.0	-0.01	0.01	0.26***	0.12***	-0.01	-0.04***	-0.19***
	7	-0.01	0.1***	0.09***	0.04***	-0.01	0.4***	0.05***	0.02	0.03**	-0.02
L3	2	0.24***	0.07***	0.1***	-0.07***	0.06***	0.24***	0.08***	0.14***	0.09***	0.0
	3	0.23***	0.02***	0.13***	-0.02***	-0.05***	0.47***	0.05***	0.19***	0.03***	-0.03***
	4	0.18***	0.06***	0.18***	-0.04***	-0.05***	0.57***	-0.01*	0.12***	0.02***	-0.03***
	5	0.29***	0.01	0.07***	-0.08***	-0.05***	0.4***	0.11***	0.22***	-0.01	0.01**
	6	0.1***	0.09***	0.15***	-0.01	0.03***	0.42***	0.05***	0.09***	0.03***	-0.1***
	7	-0.08***	0.14***	0.12***	0.03***	-0.07***	0.67***	0.07***	0.1***	0.03***	-0.03***
L4	2	0.05***	-0.11***	-0.02*	-0.13***	-0.02	0.42***	0.02*	0.14***	0.11***	-0.01
	3	0.03**	-0.05***	0.21***	-0.13***	-0.06***	0.47***	-0.0	0.09***	0.07***	-0.02**
	4	0.05***	0.0	0.24***	-0.12***	-0.11***	0.53***	0.09***	0.07***	0.09***	-0.08***
	5	0.05***	-0.04***	0.04***	-0.12***	-0.05***	0.44***	0.11***	0.13***	0.1***	0.01
	6	0.09***	-0.05***	0.05***	-0.07***	-0.06***	0.47***	0.07***	0.13***	0.12***	-0.0
	7	-0.01	0.02**	0.04***	-0.04***	-0.1***	0.67***	0.05***	0.12***	0.12***	-0.04***
L5	2	0.05***	-0.01	0.06***	-0.11***	0.01	0.4***	0.11***	0.22***	0.09***	-0.02
	3	0.05***	-0.02**	0.08***	-0.07***	-0.05***	0.61***	-0.03**	0.09***	0.09***	-0.01
	4	0.06***	-0.01	0.07***	-0.07***	-0.09***	0.6***	0.03**	0.11***	0.11***	-0.03**
	5	0.04***	-0.02	0.03**	-0.02	-0.04**	0.27***	0.27***	0.31***	0.02*	-0.01
	6	0.01	-0.02*	0.05***	-0.1***	-0.06***	0.59***	0.02*	0.02*	0.1***	-0.04***
	7	-0.01	0.0	0.01*	-0.05***	-0.03***	0.64***	0.06***	0.11***	0.11***	-0.22***

C Appendix C

This appendix provides variations in influences over time across travel clusters; as the risk influences are expressed as non-standardised coefficients (with units), influences within each travel cluster are not comparable against each other. However, the evolution of a risk influence over time within each travel cluster can still be inferred as shown in Figure C.1.

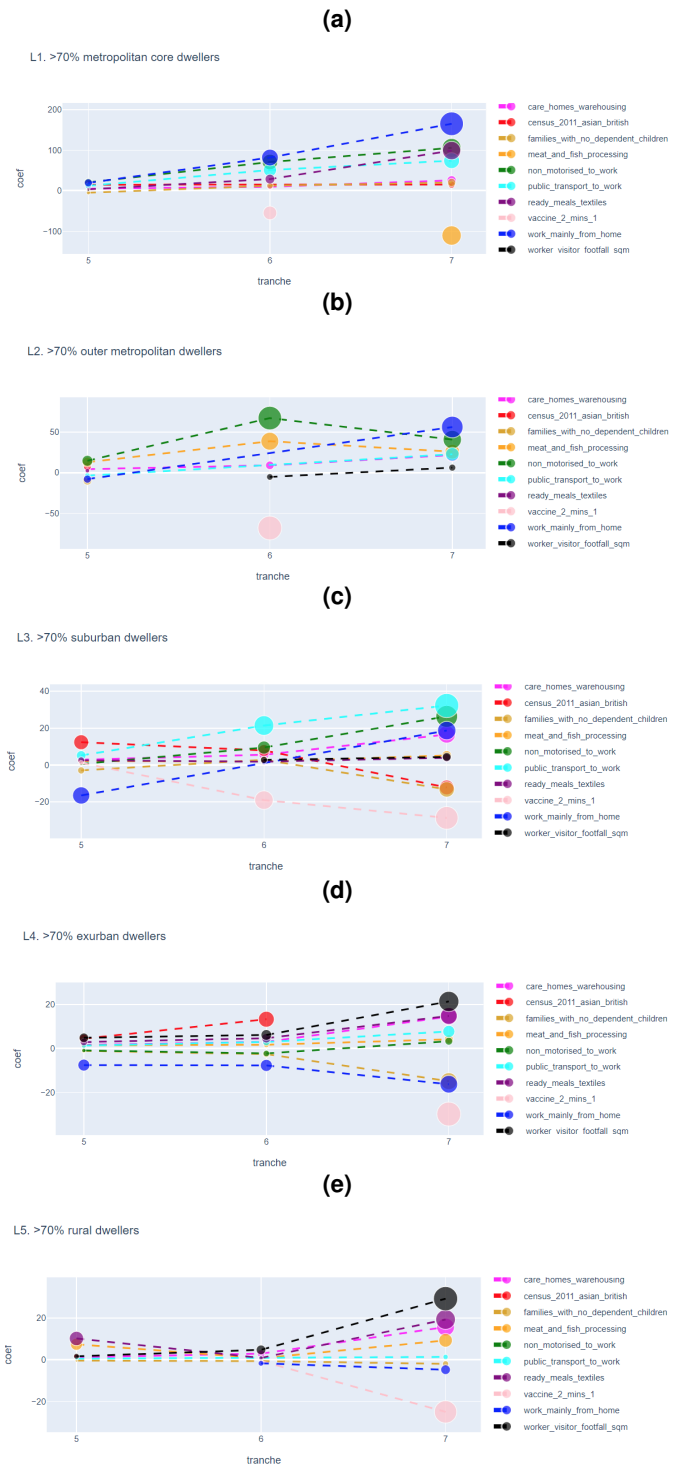


Figure C.1. Evolution of standardised regression coefficients for static and dynamic predictors over the various tranches of interest.

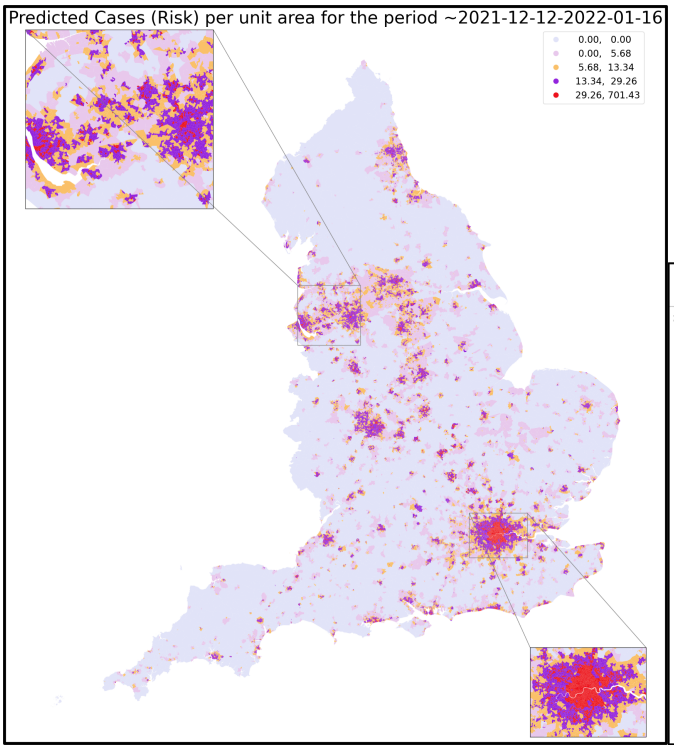
D Appendix D: Prediction of Risk: an early warning detection framework

In the previous section, we aggregated the static and dynamic predictors and inferred the most significant risk factors driving infections for each individual travel cluster. To further leverage the usefulness of the model, we use the trained estimator to predict the future number of cases on unseen test dataset.

For this we use a "sliding window" training approach where we can make use of an estimator fitted on a prior time tranche

to predict the number of cases for the next time tranche. It is worth pointing out that the size of the time-window of the test tranche is governed by the time span of the dynamic predictors alone, and can be somewhat arbitrary. The test tranche can span over a single week or longer if desired. If the test dataset covers a time period significantly different from the train data, the predictions from the model are expected to be different from the actual number of cases. For example, the Omicron variant was not dominant in England for the period covered under tranche 7, and estimator trained on the dataset for tranche 7 is expected to perform poorly in terms of predictions for the successive tranche covering the period of enhanced transmissions driven by the Omicron variant. Deviation from the models' predictions can be an important indicator in flagging areas of potential interest. For example, LSOAs with significant positive deviation (underestimation by the model) can be brought to the attention of decision makers as emerging hotspots where the growth of the infection curve cannot be accounted for despite controlling for a range of static and dynamic predictors investigated in this study.

To illustrate the above idea further, we fit an estimator for each travel cluster using the dynamic and static predictors aggregated between the period 18th July 2021-5th December 2021 (tranche 7). Following the training, we use the estimators to predict future cases per unit area. We also apply regularisation to tackle the issue of potential overfitting and feature selection in the training step. The predicted number of cases per unit area are shown in a thematic map in Figure D.1. As evident, the highest number of infections (per unit area) are expected to occur in the most dense urbanised area. Of the top 20 LSOAs with the highest predicted number of cases, the majority of the LSOAs are in London signalling some of the highest infection rates prevalent in most dense urbanised areas.



Date	travel_cluster	MSOA11NM	RGN19NM	Predicted_cases_test
2021-12-12-2022-01-16	L1. >70% metropolitan core dwellers	Tower Hamlets 031	London	701.0
		Kensington and Chelsea 021	London	568.0
		Haringey 020	London	444.0
		Islington 011	London	341.0
	L3. >70% suburban dwellers	North Tyneside 012	North East	288.0
	L1. >70% metropolitan core dwellers	Tower Hamlets 012	London	287.0
		Hackney 021	London	242.0
	L3. >70% suburban dwellers	Swindon 015	South West	225.0
	L1. >70% metropolitan core dwellers	Tower Hamlets 025	London	217.0
		Lewisham 002	London	213.0
		Camden 016	London	197.0
	L3. >70% suburban dwellers	Leicester 018	East Midlands	193.0
	L1. >70% metropolitan core dwellers	Hackney 016	London	175.0
		Tower Hamlets 009	London	173.0
		Hackney 007	London	172.0
	L3. >70% suburban dwellers	Northampton 022	East Midlands	170.0
	L1. >70% metropolitan core dwellers	Newham 015	London	169.0
	L3. >70% suburban dwellers	Northampton 023	East Midlands	167.0
	L1. >70% metropolitan core dwellers	Westminster 014	London	166.0
		Westminster 022	London	166.0

Figure D.1. Predicted mean number of cases per unit area on the latest test data encompassing the period 12th December 2021-16th January 2022. Also shown are the top-20 LSOAs with the highest predicted cases per unit area for the corresponding time period *-*.

- The model was trained on the dynamic and static predictors aggregated between the period 18th July 2021-5th December 2021. The training dataset covers the period of no national level restrictions in England except the introduction of facial covering in certain public places and on public transport wef 30th November 2021. The Omicron variant was dominant in the period covered by the test data.