Spatio-temporal characteristics of the SARS-CoV-2 *Omicron* variant spread at fine geographical scales, and comparison to *Delta*

A.J. Wood[1], A.R. Sanchez[1], P.R. Bessell[1], and R.R. Kao*[1,2]

[1]Roslin Institute, University of Edinburgh
[2]Royal (Dick) School of Veterinary Studies, University of Edinburgh

December 7, 2022

Abstract

To better understand the drivers of spread of an infectious disease such as COVID-19, it is crucial to identify the most likely individuals to become infected. The continued circulation of SARS-CoV-2 remains a global concern, however with falls in testing only a small fraction of infections are now being recorded. Comprehensive data from earlier periods may then prove a unique resource for probing finer patterns of infection.

To this end we analyse the fine spatio-temporal distribution of over 450,000 cases of COVID-19 in Scotland in waves of the B.1.1.529 *Omicron* and B.1.617.2 *Delta* lineages, from May 2021 to January 2022. We use random forest regression on case numbers, informed by measures of geography, sociodemographics, testing and vaccination. We then identify indicators for higher case numbers, showing that despite marked differences in the velocity of the outbreaks, the risk factors are remarkably similar.

We show how finer variation is only adequately explained through the use of multiple explanatory variables, implying that case heterogeneity resulted from a complex interplay of individual behaviour, immunity, and willingness to test. This analysis also provides evidence that the case distribution may be biased relative to that of *all* infections, particularly with respect to local deprivation.

1 Introduction

The B.1.1.529 *Omicron* lineage of SARS-CoV-2 was first reported to the World Health Organization by South Africa on 24th November 2021, and it was immediately designated as a variant of concern (VOC) [1]. Early, rapid spread of this variant in South Africa's Gauteng province was linked to evidence of an elevated risk of reinfection, displaying many mutations with evidence of an enhanced capability to evade prior immune protection [2, 3]. The Omicron variant became the dominant variant globally in a matter of weeks after its first reporting, and as of November 2022 remains dominant, albeit with further sub-mutations [4].

While there is now strong evidence that Omicron infection typically leads to less severe illness compared to earlier SARS-CoV-2 variants [5, 6, 7], the rate of its initial spread resulted in the imposition of many measures for its control, and the volume of infection has led to sustained pressure on healthcare systems since its introduction. The continued high volume of circulation of the disease remains a matter of considerable international concern, and predictable trends in incidence or variant emergence are yet to be established.

*Corresponding author: rowland.kao@ed.ac.uk

The first COVID-19 cases in Scotland consistent with the Omicron variant (BA.1 sub-lineage) were detected on or around 15th November 2021, preceding a rapid rise in cases from about 2,500 per day, to a peaking of over 15,000 cases per day by the beginning of January 2022. Limited non-pharmaceutical interventions (NPIs) were introduced to curb the spread, however these mostly took the form of guidance [8, 9], rather than stricter legal restrictions and closures of business imposed in earlier stages of the epidemic. The spread was rapid in terms of case volume and the breadth of affected regions and demographics, reaching remote rural communities as well as urban areas.

Continued randomised community testing indicates that, as of November 2022, COVID-19 prevalence remains high [10]. However, the winding down of free community testing (both in Scotland and globally) means that these earlier periods may become the only time in which such comprehensive testing data are collected. Our aim in this work is to better understand the patterns of cases in these earlier COVID-19 waves, with respect to a broad set of factors associated with COVID-19 infection, and severe disease.

We study the first 227,286 cases in Scotland identified as "Omicron-type", between 15th November 2021 and 6th January 2022, after which changes in legal requirements for testing make it difficult to consistently compare the available data with earlier spread. We compare to the first cases identified as *Delta*-type (B.1.617.2), spanning an earlier period between 1st May and 7th September 2021.

Throughout the epidemic, COVID-19 incidence has been highly spatially heterogeneous [11, 12, 13], a situation exacerbated by the voluntary nature of the testing programme, with previous evidence suggesting that test seeking behaviour may depend on a number of different demographic factors [14, 15]. Our work aims to explain variation in test outcomes, probing at an unprecedented neighbourhood-level resolution. We fit a machine learning model to cases across all of Scotland informed by several variables, including population structure, vaccination coverage, testing behaviour and local-level demography. We then identify factors that are broadly indicative of a subpopulation having more or fewer cases. Finally, fitting the same model to the set of Delta cases we show that, while the velocity of the Omicron wave far exceeded that of Delta, the risk factors for cases remained broadly consistent.

2 Data and methods

2.1 Data

COVID-19 testing data are obtained from Public Health Scotland's *electronic Data Research and Innovation Service* (eDRIS) system. The data include individual tests by type (polymerase chain reaction (PCR) or rapid lateral flow device (LFD)), test result (positive, negative, void, inconclusive), test date, S-gene test result if known, age, sex, and residing data zone (*DZ*, a census area typically comprising 500–1,000 individuals). De-identified IDs link repeat tests by the same individual. These data — in particular the DZ, specifying location to within an area as small as 0.1km^2 in densely populated areas — allow us to study the spread of COVID-19 on a fine spatio-temporal scale. Data on severe outcomes relating to COVID-19 (hospitalisation, ICU admission and mortality) are also provided by eDRIS. We use the version of the eDRIS data from July 14th 2022.

The period 15th November 2021 – 6th January 2022 covers the first outbreak and peak of the Omicron VOC. Beforehand, the Delta VOC (S-gene positive) was dominant. From 15th November 2021, S-gene dropout cases consistently rise, and all subsequent "dropout" cases are assumed Omicron. Remaining S-gene positive cases are presumed to be Delta, consistent with broader nationwide sequence data [4]. We note that the Omicron sub-variant BA.2/B.1.1.529.2 later replaced BA.1/B.1.1.529.1, becoming dominant in Scotland from around 25th February 2022. This variant, like Delta, has an S-gene positive test signature. However by the end of the period studied the BA.2 variant was only being identified in fewer than 1% of fully sequenced cases in the UK [16], and here we assume all remaining S-gene positive cases to be Delta.

Prior to January 6th 2022 in Scotland, positive LFD tests (typically taken at home) required PCR confirmation. Approximately 90% of cases in this period have a definitive S-gene result. A policy change then dropped this PCR requirement [17], after which cases with S-gene results fell to about 50% by February 2022 (per eDRIS data).

2.2 Random forest methods and choice of predictors

We fit the time-aggregated case distribution via random forest regression in *R*, using the *randomForest* package [18]. This allows us to fit the distribution without specifying any explicit relation between cases and other variables. Random forest methods are suitable here due to their ability to capture interactions between different (but often correlated) variables which may be highly non-linear in nature, as well as random forests' known robustness to overfitting, given appropriate hyperparameter selection [19].

We group individuals with a given age range, sex, DZ and when (if at all) they last tested positive, with each of these *cohorts* usually comprising 0–50 individuals. For each cohorts, we aggregate the total number of relevant cases from the eDRIS data. For Omicron, we seek S-gene dropout cases between 15th November 2021 and 6th January 2022, and for the earlier Delta outbreak, S-gene positive cases between 1st May and 7th September 2021 (choosing this limit so as to have an similar number of cases in each set). We then fit the case distribution using a set of explanatory variables relating to prior COVID-19 testing, vaccination, age, sex, and indices associated with deprivation. A detailed description of all variables used is given in Supplementary Material B.2.

We take two metrics for variable importance from the *RandomForest* function output: the node purity, and the loss of model accuracy on random permutation of a given predictor's values. We also calculate the *accumulated local effects* (ALEs) of each predictor. ALEs describe how the model fit value changes, in response to changing one predictor variable in isolation, averaged over many different entries in the data set [20]. In this context, they provide an indicator as to whether a variable value is associated with fewer or more cases.

Hyperparameters are detailed in Supplementary Material B.3.

3 Results

3.1 Time evolution and early patterns of spread

We identified 385,558 cases between November 15th 2021 and January 6th from the eDRIS data, of which 227,286 were S-gene dropout, therefore likely Omicron. From 1st May to 7th September 2021 we identified 269,838 cases, of which 229,073 were likely Delta (S-gene positive).

The start date for each of these periods is the first date from which there are consistent rises in cases that are likely the new VOC. Per regression on $\log_{10}$(cases) (Fig. A.1), Omicron-type cases (hereafter referred to as Omicron) had a doubling time of 2.9 days over the first 28 days, compared to 6.2 days for Delta. Omicron also outpaced Delta by spatial spread, measured by new DZs having reported at least one case over time (also Fig. A.1). At the scale of Scotland's 32 local authorities (LAs), the time-dependent reproduction number R_t (methodology detailed in Supplementary Material B.1) consistently rises for Omicron, peaking at above 2 for nearly all LAs 28 days in to the outbreak, and only consistently falling below 1 after 50 days. Reproduction numbers for Delta are less consistent between LAs; while the number generally remains above 1 for most LAs in the period, there is no coherent peak at the start of the wave.

On the introduction of Omicron, the Delta variant was still circulating in high volume, though the subsequent spread of Omicron coincided with Delta declining. In the intermediate period with both variants circulating in high volume, the *age* distributions by variant differed. Taking the mid-points of the five-year age brackets, the mean ages of the Delta-type cases was 3.9 years lower than the Omicron-type cases (31.8 years compared to 35.7 years) (Student's $t = -52.2$, $p < 0.001$). This was the case from relatively early on when Omicron accounted for at least 5% of

cases (Fig. A.3A). However, the median ages are equal (both 32.5 years), as in the Omicron-type cases there is a trough in those aged 0–14, with fewer than 50% of cases in this age group Omicron, but then a peak in the 20–29 age group (Fig. A.3B).

3.2 Detailed case distribution analysis

Fig. 1 shows the case distributions by age, sex, prior cases, deprivation and health board. Omicron case rates were highest in younger adults, peaking at 90 cases/1,000 in ages 20–24. There was no marked difference in rates between men and women. Case rates were much lower amongst those that had tested positive for COVID-19 previously.

Fitting our models, Fig. 2 shows performance at DZ level, comparing observed cases to fit cases. Points deviating from the diagonal indicate DZs where the model was less accurate. Our full model explains 70% (fit: 71%, test: 62%) of local variation in the case distribution (R-squared for case numbers, aggregated at a DZ level). We also simultaneously fit three "univariate" models, under equivalent parameters to the full model, but informed with fewer data. An equivalent model informed by population and population density alone explained 59% (fit: 60%, test: 55%) of variation. A model informed by only population/deprivation rank explained 53% (fit: 53%, test: 51%), and one informed by only population/age explained 48% (fit: 48%, test: 51%). Fig. 2 shows further deviation of the data-fit slopes away from the diagonal for these models.

To probe *spatial* variation in cases not explained by the model, we measure the Moran's I autocorrelation [21] on the residuals, relating to their physical location (see also Fig. A.4). We compare local DZ-aggregated residuals over physical distances (from 1–100km), as well as network distance (number of nearest neighbours apart). Fig. 3 shows for the full model, autocorrelation of residuals within 1km (representative of correlation between close neighbours) is 0.35, falling to 0.15 at 5km (representative of correlation across distinct neighbourhoods within towns/cities), and 0.05 at 50km. The single-variable models exhibit much higher residual clustering, with the density-only model performing best, but persisting over larger distances. This is consistent with the ability of those variables to explain spatial variation; while there is little clustering in age and deprivation beyond short neighbourhood-level distances, there is much stronger clustering in density (e.g., city centres).

Returning to the data, we also see variation over Scotland's 14 health boards, most notably in Orkney, Shetland and the Western Isles (all island communities), as well as a fall with increasing rurality (map representation in Fig. A.4). The trend with respect to multiple deprivation rank is bimodal, with higher rates in both the most deprived, and least deprived communities.

Fig. 1 shows these trends are also well captured by the model fit. It is important that the trends with deprivation, rurality and location have been captured, as the model was not explicitly informed by these, rather by adjacent measures (e.g., for rurality, population density).

3.2.1 Accumulated local effects

The accumulated local effects (ALEs) help identify *risk factors* (variable values that the model associates with more cases). For the range of explanatory variables in Fig. 4, the ALE is the average cohort-level "contribution" to the final fit value, over the full data set, given that predictor value. If the value is greater than zero, the fit cases on average increases given that predictor value.

Population has a trivial ALE, with the effect on the fit cases increasing near-linearly with cohort population. The ALEs for the other high-resolution predictors (*age*, *sex*, *prior case status*) also mirror the distributions in Fig. 1.

Low *local S-gene coverage* (see Supplementary Material B.2 for precise description) is protective. At the time S-gene coverage was 90% overall across mainland Scotland (per eDRIS data), but significantly lower in the LAs of Orkney Islands, Shetland Islands and Na h-Eileanan Siar (74%, 20% and 23% respectively). Increasing *local outbreak duration* (the time between January 6th and the first day an Omicron case was detected locally) is also indicative of higher cases.

Higher *mean household size* emerges as a risk factor, as is consistent with evidence of high secondary attack rates for SARS-CoV-2 [22, 23], and increased risk of inter-household transmission

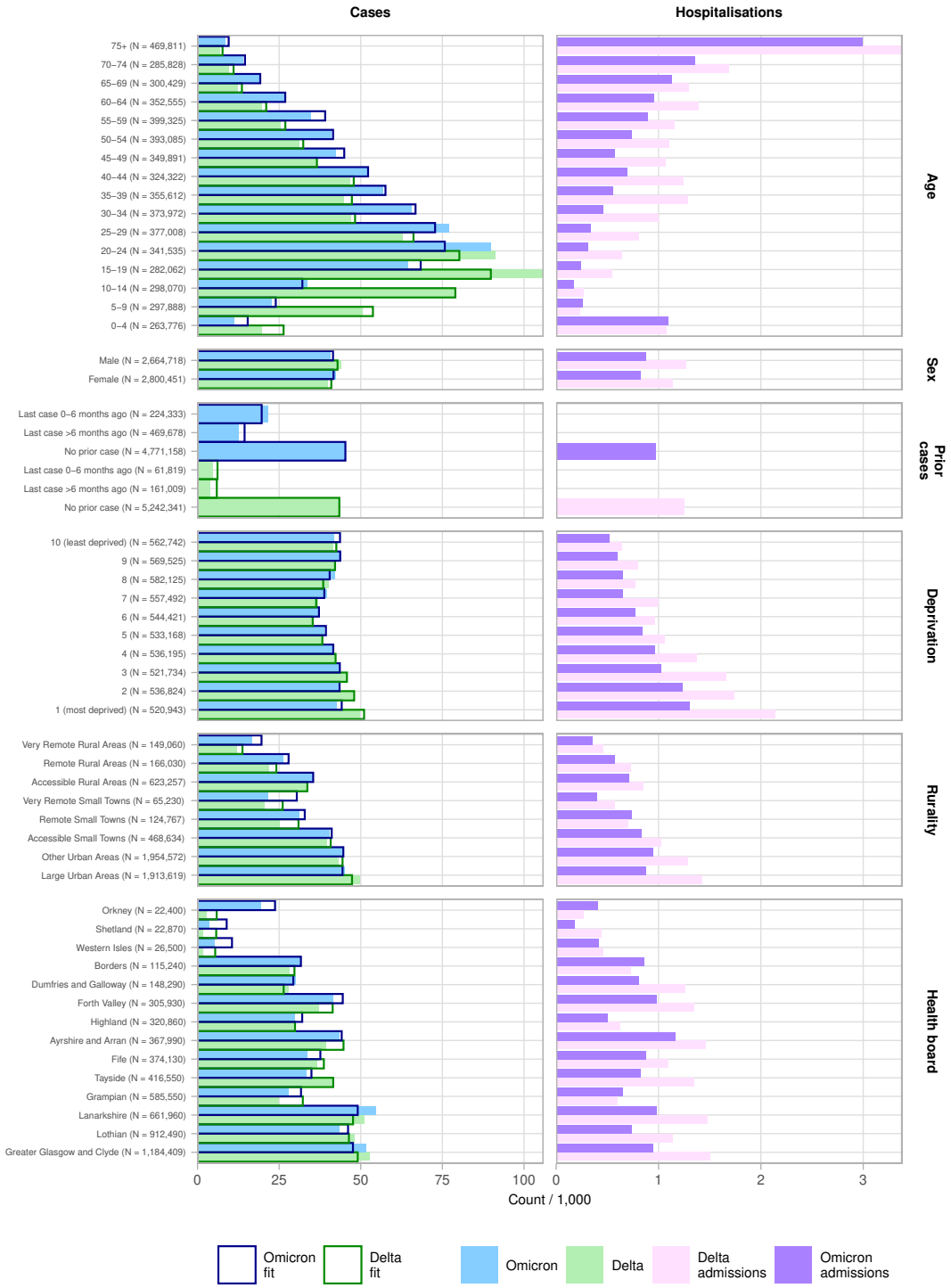

Figure 1: Cases (left) and hospitalisation admissions (right) by age, prior cases, deprivation, rurality and location. Note that hospital admissions count all COVID-19 admissions in that period, and may be related to cases excluded from these analyses due to the lack of an S-gene result, or may not have an associated case at all.

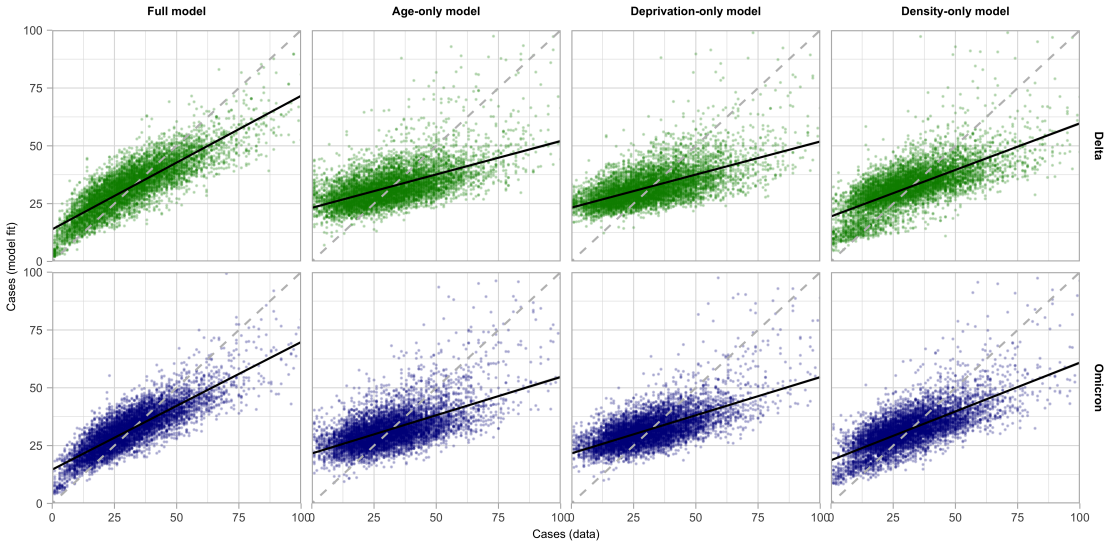

Figure 2: Performance of different models in terms of modelled cases (y-axis) in comparison with the observed cases (x-axis), for both Omicron and Delta. Each point represents a DZ. Also included are comparisons to models informed with only population, and one of either age, overall deprivation rank, or population density.

relative to other close contacts outside of the home [24].

The fit shows that the ALE for third/booster vaccination uptake is generally protective for increasing uptake. The protective value at *zero* uptake is most likely an artefact arising from differences in eligibility. During the wave, eligibility was not even across ages, with many younger groups either not yet eligible, or not eligible at all (Supplementary Material B.4). Zero-uptake cohorts then mainly include ages 0–14; individuals that reported fewer cases generally, but also may also have been protected for other reasons associated with young age.

The ALE for frequency of *negative* LFD test reporting shows a positive relationship, but exhibits a tailing-off with an increasing number of tests per person. Looking in more detail, on aggregating PCR positive test results (any S-gene result) by LFD tests reported in each period, individuals that had reported no LFD results had the lowest reporting rate of PCR positives (6.2%, 4.1% for Omicron and Delta respectively). Those that reported 1–2 LFD results had the highest PCR positive rate (15.1%, 10.5%), but individuals reporting *more* than 3 LFD results had a *lower* rate of reporting again (8.1%, 5.4%).

Indicators associated with deprivation appear less influential. However, modest deprivation-level variation in case rates (Fig. 1) is well captured, and deprivation indicators emerge as important in the model fit (Fig. C.6). There are correlations between different measures of deprivation, but these measures also correlate strongly with other variables in the data, such as age (with more deprived DZs skewing younger), negative LFD reporting (with more deprived DZs generally reporting fewer tests), and vaccine uptake (with more deprived DZs generally having lower uptake). It is plausible, then, that the fit is more strongly influenced with a combination of several correlated indicators of deprivation, rather than one in isolation.

Now comparing our results for the Omicron wave to earlier Delta cases from 1st May to 7th September 2021, the spatial distribution (Fig. A.4) shows high case rates again within urban areas, as well as a small number of localised outbreaks in rural communities. Fig. 1 shows cases skewed slightly younger, with the highest rates within ages 15–19. Across deprivation deciles, as with Omicron, the distribution is bimodal, with higher rates in both the most and least deprived DZs.

The Delta model was fit using the same parameters as Omicron (Section 3.2), with an exception that third/booster dose uptake was replaced with second dose uptake (third/booster doses were

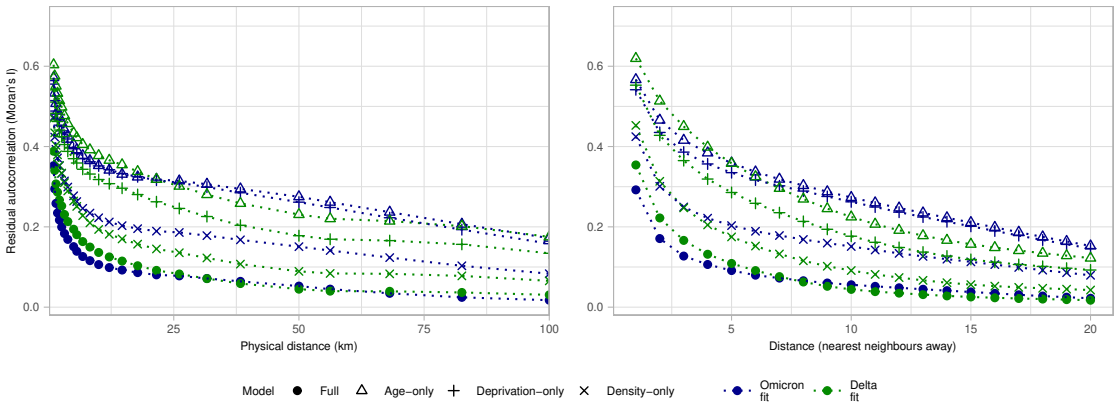

Figure 3: Residual clustering as measured by the Moran's I statistic, at different physical (left) and network-based distances (right). Higher values represent higher autocorrelation between residuals within a given locus. The point locations of each DZ are the population-weighted centroids. Nearest neighbours are defined as DZs that share a boundary.

only administered later). Model performance was similar, explaining 72% (fit: 73%, test: 61%) of DZ-level variation. Fig. 4 shows the same broad risk factors for Omicron also emerged for Delta. Excluding non-eligible cohorts with zero uptake, higher second dose uptake emerged protective. A notable difference from Omicron is that cohorts with higher black and minority ethnic populations were associated with higher case rates, more sharply than seen for Omicron.

3.3 Comparison of distribution of cases to hospital admissions in the same period

Included alongside the case distribution in Fig. 1 is the distribution of COVID-19 related hospital admissions within the same periods. As our case distributions only take a subset of cases in the period (excluding those with either a likely different, or unknown variant), these should not be used to infer a raw case-to-hospitalisation rate, as not all cases in the period are accounted for.

For both periods, admissions were highly skewed towards older ages, consistent with earlier stages of the epidemic [25, 26]. By associating hospital admissions and particular COVID-19 cases with a matching DZ, age range and sex, up to 21 days prior to the admission, we find nearly all admissions had no associated prior case, in the past.

Admissions in both periods had a strong deprivation skew; for Omicron, the lower half of DZs ranked by multiple deprivation reported 63% of admissions, and for Delta, the proportion was 64%. This pattern is not seen in cases, which we discuss further in Section 4.

4 Discussion

The detection of the *Omicron* variant in Scotland in November 2021 preceded an explosive wave of cases, consistent with the enhanced ability of the variant to evade prior immunity, and the lack of NPIs imposed compared to prior waves. In this work we have probed the spatio-temporal patterns of this wave and an earlier wave of the Delta variant, at the resolution of neighbourhoods. These waves bookend the period in which free community COVID-19 testing was widely available, after which rates of case ascertainment fall sharply.

Our spatial distribution model shows that variations in case counts in these waves of infection are associated with broad factors such as age structure, but also with factors relating to testing, vaccination, geography and demography, all of which vary on finer spatial scales. We have shown

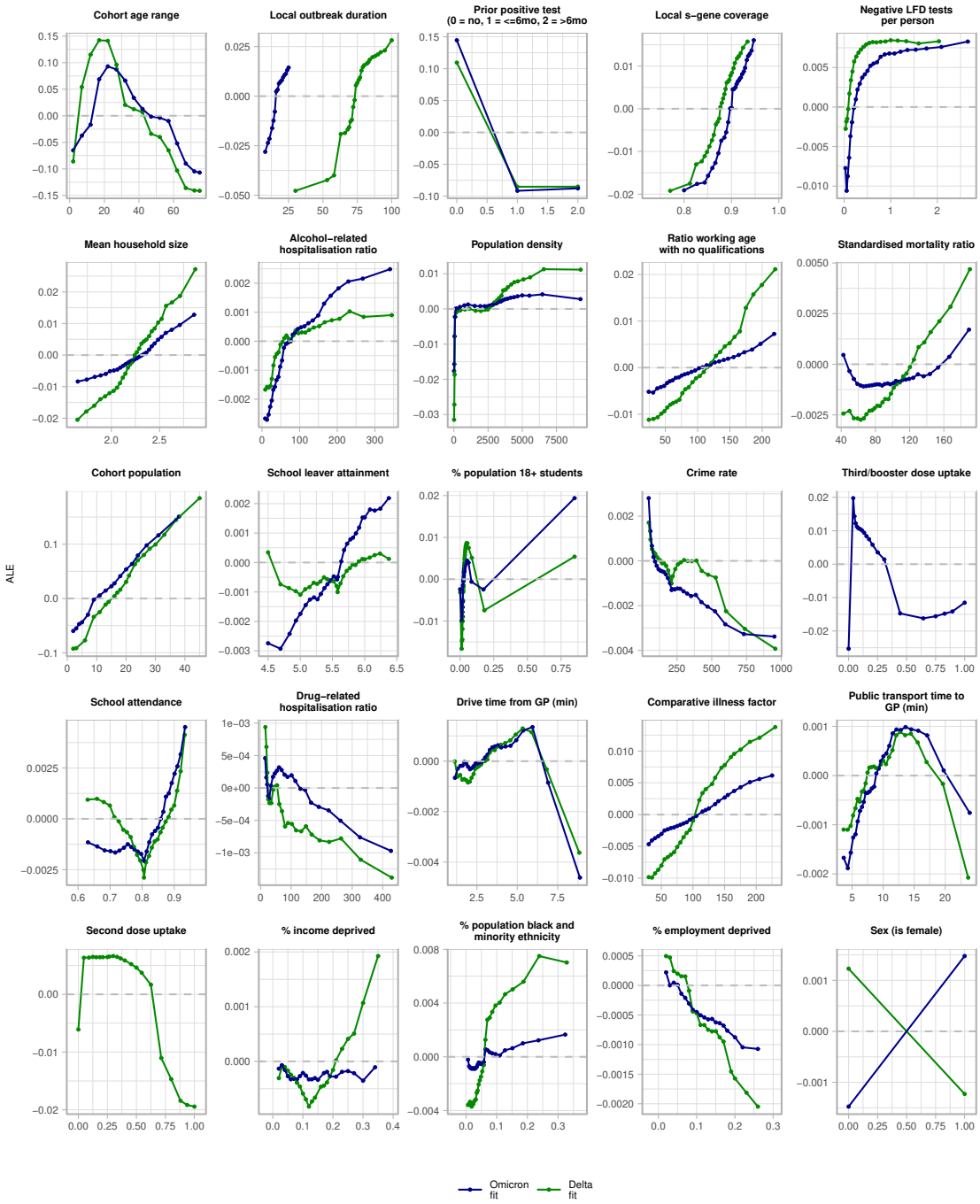

Figure 4: Accumulated local effects across all explanatory variables. For each variable, the x-axis represents the range of values of that variable in the data, and the y-axis is the univariate ALE for that variable value. The overall magnitude of the ALE represents the relative size of the effect.

that the full spatial distribution of COVID-19 cases is only reasonably explained after considering a combination of all of these factors.

At the individual level, the causative drivers of COVID-19 transmission are close-proximity interactions, and individual-level variation in virus susceptibility and transmissibility. Given individual interactions, immunity from either vaccination or prior infection then affects the probability of infection, if exposed. Finally, the propensity of individuals to test, and report the results of those tests, affects whether an infection is identified and registered. Evidence from random community testing suggests that only a minority of infected persons do so [27, 28]. Our distribution model uses explicit data on COVID-19 vaccination and tests, as well as data on geography and demography to serve as proxies for these various drivers. While it is beyond the scope of this work to fully disentangle the relative impacts of all these factors, our analyses provide some important insights.

Living in a deprived community has been identified as a risk factor for poor COVID-19 outcomes [29, 30, 31], however our analysis does not show such an inequality in case volume. Rather, a clear skew is only seen in admissions, with approximately two thirds reported from the lower half of neighbourhoods ranked by deprivation. We suggest this skew arises from a higher risk of hospitalisation given *infection* with higher deprivation, compounded by potential variation in rates of case *ascertainment*. In the period studied, LFD tests were abundantly available in Scotland, free-of-charge, with the public encouraged to test and report results regularly regardless of symptoms or outcome. The reporting of LFD negatives may then relate to propensity for testing overall, and higher reporting of asymptomatic or mild infections that would not result in hospital admission. Rates of LFD reporting were consistently lower in more deprived areas, and LFD positivity higher (e.g., [32], and also observed in a wider study in Liverpool, England [33]).

The widespread deployment of effective vaccines against COVID-19 fundamentally changed the nature of the epidemic in Scotland, with the individual risk of infection reduced, as well as disease severity. Our models generally associate higher uptake with fewer cases amongst those eligible at the local level, however the vaccinations and cases data are not linked, thus we do not know the specific vaccination status of those reporting cases. Had they been linked, we may have captured an even stronger protective effect.

While the Omicron wave was much more rapid than earlier waves, we found that the risk factors for testing positive were broadly similar to those for testing positive for Delta, when analysing a similar number of cases from an earlier wave. This is despite substantial differences in time of year, imposed NPIs, and overall outbreak trajectories, with the Omicron outbreak being more rapid and more consistent across different local authorities, as compared to Delta.

Our model is by design not informed with position data, nor by network characteristics (e.g., data on nearest neighbours). It is highly likely that the DZ-level case rates are influenced not just by the characteristics of that area, but also of surrounding communities, and communities connected to it by other means (e.g., transport links). We attempted to account for some of this with the *local outbreak duration* parameter, describing when cases had begun circulating locally, but one could also consider informing a model with measures relating to neighbours of DZs, or census data detailing patterns of travel between different DZs. Residual correlation persists locally, but falls significantly over length scales of 10km, or network separations of 5–10 nearest neighbours. We also acknowledge our model fits cases aggregated over time. We do not therefore account for within-period changes in factors such as immunity and public health policy.

The COVID-19 data collated by eDRIS are remarkable in terms of volume and resolution, and has allowed us to isolate cases down to as few as 1–50 individuals. However, regardless of resolution, cases only partially represent the full underlying pattern of infection, and variation in testing frequency and differing trends in hospital admissions suggests the distribution of infections may be very different to that of reported cases. By incorporating trends on cases, testing behaviour, and data on severe outcomes more closely linked to infection (hospitalisation, ICU admission and mortality), it may be possible to build a much more comprehensive picture of how infections were distributed amongst the population.

5 Acknowledgements

We thank Public Health Scotland's electronic Data Research and Innovation Service (eDRIS) for the provision of COVID-19 testing, vaccination and severe outcomes data.

6 Author contributions

R.R.K. conceived the project. A.J.W. wrote the model code and performed the spatial analysis. A.R.S. and P.R.B. performed the temporal analysis. A.J.W. and R.R.K. wrote the manuscript.

7 Competing interests

The authors declare no competing interests.

8 Code availability

Analysis code is available at https://git.ecdf.ed.ac.uk/awood310/scotland-covid-case-distribution-random-f

9 Data availability

The COVID-19 testing, vaccination and hospitalisation data utilised in this work are not publicly available. They are provided to the authors for academic research by Public Health Scotland's electronic Data Research and Innovation Service, under a data sharing agreement (*Spatial and Network Analysis of SARS-CoV-2 Sequences to Inform COVID-19 Control in Scotland*).

Data relating to deprivation are derived from the 2020 Scottish Index of Multiple Deprivation, and are publicly available [34]. Fine-scale population statistics are drawn from publicly-available population estimates as of mid-2020 [35]. The population of students and those belonging to a minority ethnicity are drawn from public Scottish census data (tables KS501SC, LC2101SC respectively).

10 Funding statement

This work has been funded by the ESRC grant ES/W001489/1: *Real-time monitoring and predictive modelling of the impact of human behaviour and vaccine characteristics on COVID-19 vaccination in Scotland*.

References

[1] Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 2022;603(7902):679-86.

[2] Tian D, Sun Y, Xu H, Ye Q. The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. Journal of Medical Virology. 2022;94(6):2376-83.

[3] Dejnirattisai W, Huo J, Zhou D, Zahradník J, Supasa P, Liu C, et al. SARS-CoV-2 Omicron-B. 1.1. 529 leads to widespread escape from neutralizing antibody responses. Cell. 2022.

[4] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.

[5] Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. The Lancet. 2022.

[6] Veneti L, Bøås H, Kristoffersen AB, Stålcrantz J, Bragstad K, Hungnes O, et al. Reduced risk of hospitalisation among reported COVID-19 cases infected with the SARS-CoV-2 Omicron BA. 1 variant compared with the Delta variant, Norway, December 2021 to January 2022. Eurosurveillance. 2022;27(4):2200077.

[7] Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, et al. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B. 1.1. 529) and delta (B. 1.617. 2) variants in England: a cohort study. The Lancet. 2022;399(10332):1303-12.

[8] Nightclubs to Close;. The Scottish Government. url: https://www.gov.scot/news/nightclubs-to-close/.

[9] Further measures needed to reduce contacts;. The Scottish Government. url: https://www.gov.scot/news/further-measures-needed-to-reduce-contacts/.

[10] Coronavirus (COVID-19) Infection Survey: Scotland Dataset;. Office for National Statistics. url: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/covid19infectionsurveyscotland.

[11] Lee D, Robertson C, Marques D. Quantifying the small-area spatio-temporal dynamics of the Covid-19 pandemic in Scotland during a period with limited testing capacity. Spatial Statistics. 2021:100508.

[12] Banks CJ, Colman E, Doherty T, Tearne O, Arnold M, Atkins KE, et al. SCoVMod–a spatially explicit mobility and deprivation adjusted model of first wave COVID-19 transmission dynamics. Wellcome Open Research. 2022;7(161):161.

[13] Lone NI, McPeake J, Stewart NI, Blayney MC, Seem RC, Donaldson L, et al. Influence of socioeconomic deprivation on interventions and outcomes for patients admitted with COVID-19 to critical care units in Scotland: a national cohort study. The Lancet Regional Health-Europe. 2021;1:100005.

[14] Bevan I, Stage Baxter M, Stagg HR, Street A. Knowledge, attitudes, and behavior related to COVID-19 testing: a rapid scoping review. Diagnostics. 2021;11(9):1685.

[15] Dayton L, Song W, Kaloustian I, Eschliman EL, Strickland JC, Latkin C. A longitudinal study of COVID-19 disclosure stigma and COVID-19 testing hesitancy in the United States. Public Health. 2022;212:14-21. Available from: https://www.sciencedirect.com/science/article/pii/S0033350622002323.

[16] SARS-CoV-2 variants of concern and variants under investigation in England: technical briefing 35;. The UK Health Security Agency. url: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1050999/Technical-Briefing-35-28January2022.pdf.

[17] Self-Isolation and testing changes;. The Scottish Government. url: https://www.gov.scot/news/self-isolation-and-testing-changes/.

[18] Liaw A, Wiener M, et al. Classification and regression by randomForest. R news. 2002;2(3):18-22.

[19] Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

[20] Apley D, Apley MD. Package 'ALEPlot'. 2018.

[21] Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1/2):17-23.

[22] Jalali N, Brustad HK, Frigessi A, MacDonald EA, Meijerink H, Feruglio SL, et al. Increased household transmission and immune escape of the SARS-CoV-2 Omicron variant compared to the Delta variant: evidence from Norwegian contact tracing and vaccination data. medRxiv. 2022.

[23] Fonager J, Bennedbæk M, Bager P, Wohlfahrt J, Ellegaard KM, Ingham AC, et al. Molecular epidemiology of the SARS-CoV-2 variant Omicron BA. 2 sub-lineage in Denmark, 29 November 2021 to 2 January 2022. Eurosurveillance. 2022;27(10):2200181.

[24] Dupraz J, Butty A, Duperrex O, Estoppey S, Faivre V, Thabard J, et al. Prevalence of SARS-CoV-2 in household members and other close contacts of COVID-19 cases: a serologic study in canton of Vaud, Switzerland. In: Open forum infectious diseases. vol. 8. Oxford University Press US; 2021. p. ofab149.

[25] COVID-19 statistical report, as at 15 November 2021;. Public Health Scotland. url: https://www.publichealthscotland.scot/publications/covid-19-statistical-report/covid-19-statistical-report-17-november-2021/.

[26] Coronavirus (COVID-19) weekly insights: latest health indicators in England, 12 February 2021;. Office for National Statistics. url: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19weeklyinsights/latesthealthindicatorsinengland12february2021.

[27] Colman E, Puspitarani GA, Enright J, Kao RR. Ascertainment rate of SARS-CoV-2 infections from healthcare and community testing in the UK. Journal of Theoretical Biology. 2022:111333. Available from: https://www.sciencedirect.com/science/article/pii/S0022519322003241.

[28] Nightingale ES, Abbott S, Russell TW. The local burden of disease during the first wave of the COVID-19 epidemic in England: estimation using different data sources from changing surveillance practices. BMC public health. 2022;22(1):1-14.

[29] Blundell R, Costa Dias M, Joyce R, Xu X. COVID-19 and Inequalities. Fiscal studies. 2020;41(2):291-319.

[30] Bambra C, Riordan R, Ford J, Matthews F. The COVID-19 pandemic and health inequalities. J Epidemiol Community Health. 2020;74(11):964-8.

[31] Baena-Díez JM, Barroso M, Cordeiro-Coelho SI, Díaz JL, Grau M. Impact of COVID-19 outbreak by income: hitting hardest the most deprived. Journal of Public Health. 2020;42(4):698-703.

[32] Coronavirus (COVID-19): modelling the epidemic (issue no.93);. The Scottish Government. url: https://www.gov.scot/publications/coronavirus-covid-19-modelling-epidemic-issue-no-93/documents/.

[33] Green MA, García-Fiñana M, Barr B, Burnside G, Cheyne CP, Hughes D, et al. Evaluating social and spatial inequalities of large scale rapid lateral flow SARS-CoV-2 antigen testing in COVID-19 management: An observational study of Liverpool, UK (November 2020 to January 2021). The Lancet Regional Health-Europe. 2021;6:100107.

[34] Scottish Index of Multiple Deprivation 2020;. The Scottish Government. url: https://www.gov.scot/publications/scottish-index-of-multiple-deprivation-2020v2-indicator-data/.

[35] Mid-2020 Small Area Population Estimates for 2011 Data Zones;. National Records of Scotland. url: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/small-area-population-estimates-2011-data-zone-based/mid-2020/.

[36] Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of epidemiology. 2004;160(6):509-16.

[37] Coronavirus (COVID-19): vaccine deployment plan 2021;. The Scottish Government. url: https://www.gov.scot/publications/coronavirus-covid-19-vaccine-deployment-plan-2021/.

[38] Major milestone in vaccination programme;. The Scottish Government. url: https://www.gov.scot/news/major-milestone-in-vaccination-programme/.

[39] Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19 vaccine effectiveness against the omicron (B. 1.1. 529) variant. New England Journal of Medicine. 2022.

[40] Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature. 2022;602(7898):654-6.

[41] Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, et al. Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. The Lancet. 2021;397(10285):1646-57.

[42] Scotland's autumn/ winter vaccination strategy 2021;. The Cabinet Secretary for Health and Social Care. url: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2021/09/scotlands-autumn-winter-vaccination-strategy-20212/documents/scotlands-autumn-winter-vaccination-strategy/scotlands-autumn-winter-vaccination-strategy/govscot%3Adocument/scotlands-autumn-winter-vaccination-strategy.pdf.

[43] Sheikh A, Kerr S, Woolhouse M, McMenamin J, Robertson C. Severity of Omicron variant of concern and vaccine effectiveness against symptomatic disease: national cohort with nested test negative design study in Scotland. 2021.

Supplementary material

Spatio-temporal characteristics of the SARS-CoV-2 Omicron *variant spread at fine geographical scales, and comparison to* Delta

A.J. Wood, A.R. Sanchez, P.R. Bessell, R.R. Kao

A Summary plots of Omicron and Delta waves

Supplementary plots of the time evolution and spatial distribution of Omicron and Delta cases are presented in Figures A.1–A.4.

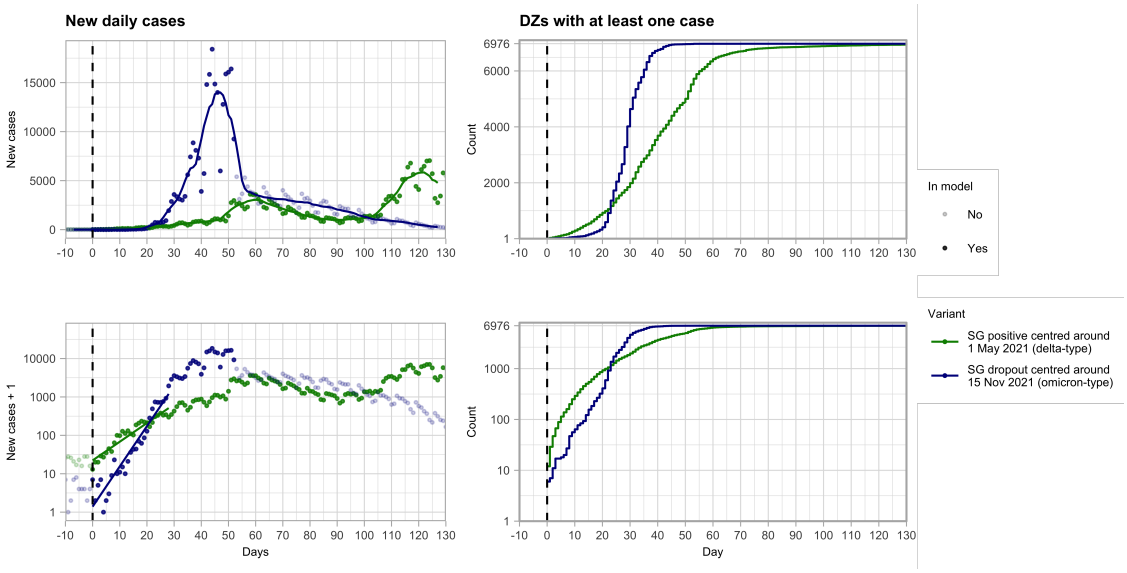

Figure A.1: Timeseries of the initial outbreaks of the Delta and Omicron variants.

B Methodology details

B.1 Estimation of reproduction numbers

The time-dependent reproduction number R_i is the average number of forward infections caused by a person infected on day t_i. Define n_j as the number of new infections on day t_j. These new infections came from individuals infected on days on, or prior to t_j. Define A_{ij} as the number of new infections on day t_j *specifically* from those infected on day $t_i \leq t_j$:

$$A_{ij} = \frac{(n_i - \delta_{ij})P(t_j - t_i)}{\sum_{i' \leq j}(n_{i'} - \delta_{i'j})P(t_j - t_{i'})} n_j \ .$$

$P(\Delta t)$ is the probability of an individual passing on the infection, Δt days after being infected. The presence of the Kronecker delta excludes the possibility of infected individuals infecting themselves. The reproduction number R_i is then the average total of infections generated over all subsequent days [36]:

$$\begin{aligned} R_i &= \frac{1}{n_i} \sum_{j \geq i} A_{ij} \\ &= \frac{1}{n_i} \sum_{j \geq i} \frac{n_j (n_i - \delta_{ij})P(t_j - t_i)}{\sum_{i' \leq j}(n_{i'} - \delta_{i'j})P(t_j - t_{i'})} \ . \end{aligned}$$

We take $P(\Delta t)$ to be

$$P(\Delta t) \sim \mathrm{e}^{-\lambda \Delta t}$$

with λ^{-1} the mean infectious period. Individuals are equally infectious throughout the entire infection. In our calculations we estimate $1/\lambda = 6.26$ days, using the posterior mean duration

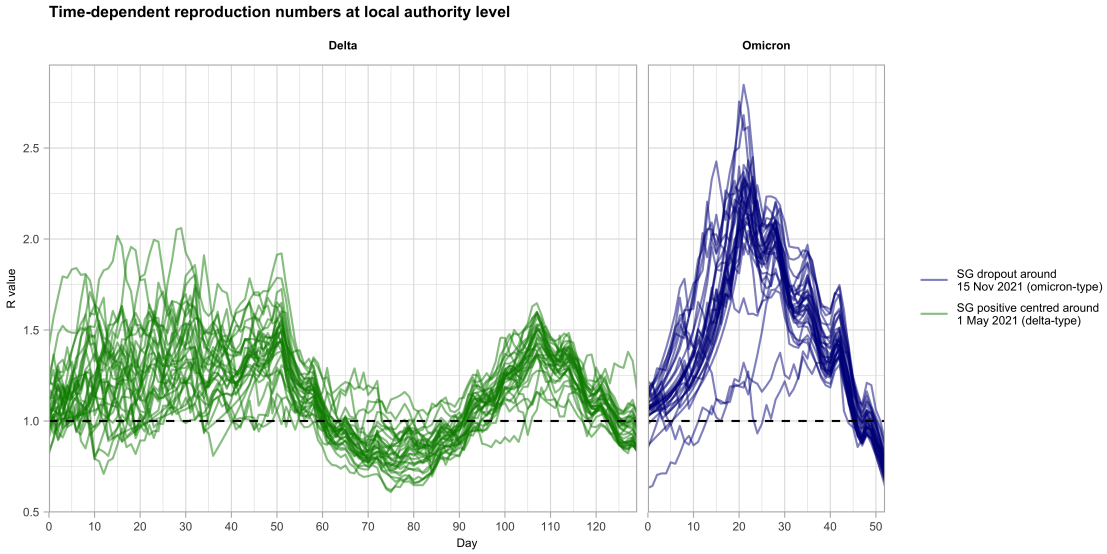

Figure A.2: Time-dependent reproduction numbers for the Delta (left) and Omicron waves (right), over each of the 32 separate local authorities.

of infectiousness obtained from the *SCoVMod* compartmental model (for more detail see Reference [12]).

As we estimate the infection reproduction number using the cases data, we implicitly assume that case ascertainment does not change over time, and does not account for the delay between infection, and registering a case.

B.2 Explanatory variables used in random forest regression model

The models described in Section 2.2 are informed with the following data, first at cohort resolution:

- Age range (five-year windows, up to 75+);
- Sex;
- Prior case status: the time of the last reported case, broken into three categories: never tested positive before, last tested positive in the 6 months prior to the first day of the outbreak, last tested positive over 6 months prior;
- Cohort population (derived using historical testing data for those testing positive before, and estimated populations as of mid-2020 collated by the National Records of Scotland [35], for the remainder that have not tested positive before).

At age/sex/DZ resolution, we then include:

- COVID-19 vaccination uptake (eDRIS) (see also Supplementary material B.4);
- Ethnicity (% population belonging to a minority ethnicity), as per the most recent Scottish census data (2011);
- The per-population, time-aggregated number of *negative* LFD tests reported.

Finally included are the following at DZ resolution or broader:

- Measures of DZ-level deprivation (obtained from Scottish census data, and the 2020 *Scottish Index of Multiple Deprivation* [34]);

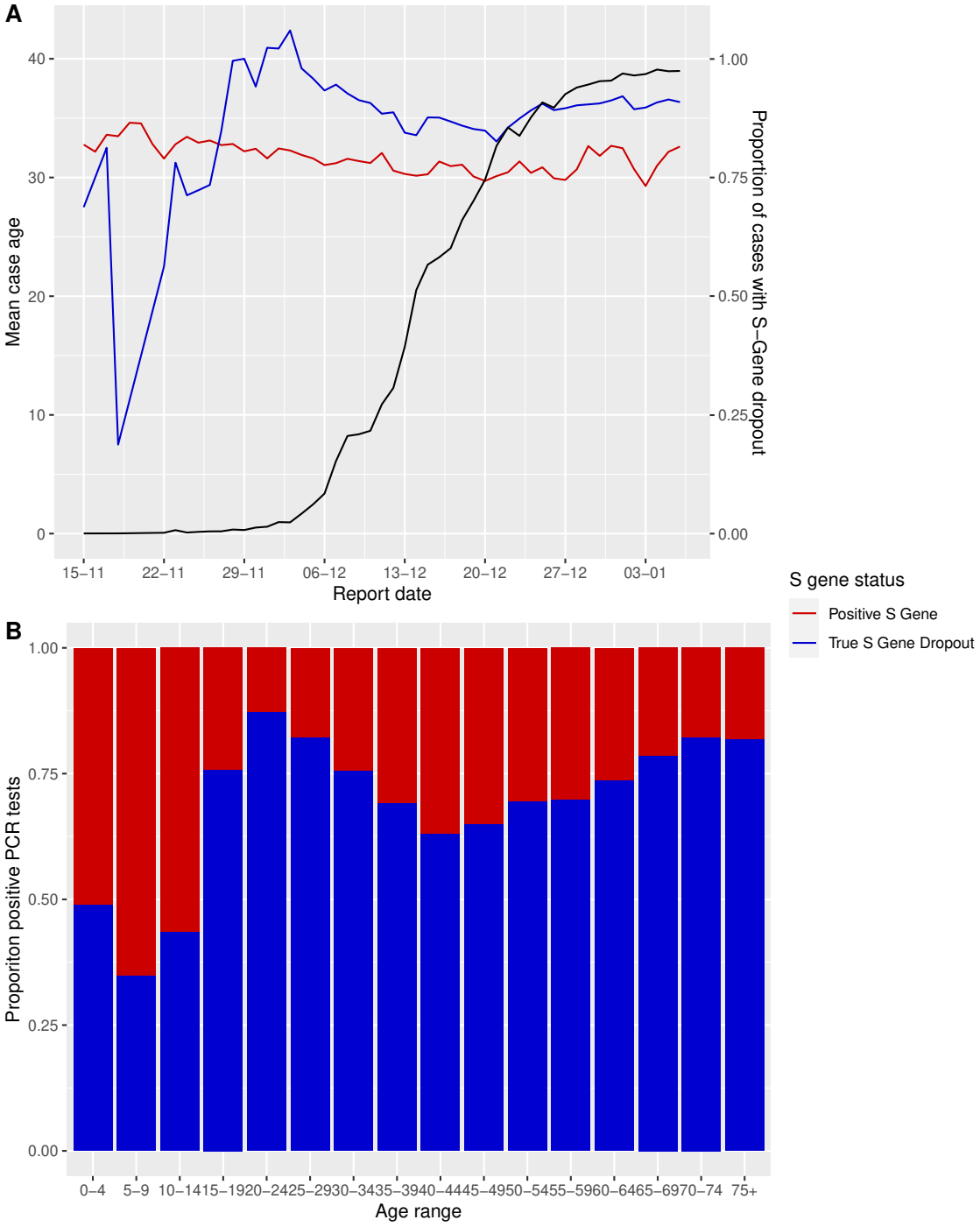

Figure A.3: PCR positive cases over the period 15th November 2021 to 6th January 2022 that were S-gene dropout or true S-gene positive. (A) Daily mean case age for the two definite PCR S-gene outcomes (blue and red lines) against the proportion of the daily cases that were true S-gene dropout (presumed Omicron-type). (B) the proportion of the cases over the period by 5-year age bracket.

- The time at which the first variant case was detected in that cohort's corresponding *intermediate zone* (*IZ*). An IZ typically contains 4–6 DZs, and 3,000–5,000 individuals, with

this granularity chosen to give a reasonable estimation of how long there had been local circulation;

- S-gene coverage (the proportion of cases with an accompanying S-gene result, required to associate a likely variant) at IZ level;
- Population density, at IZ-level.

We do not use data on *PCR* negative tests. In the Omicron wave PCR positivity peaked at 30% (per eDRIS data), with testing capacity being reached (resulting in a policy change on 5th January 2022 removing the need for a confirmatory PCR after an LFD positive [17]). Thus with this "ceiling" capacity being reached, we exclude negative PCR tests as a poorer proxy for propensity to test as compared to LFD negatives, and being too closely related to overall cases (requiring an S-gene sequenced positive PCR test).

B.3 Model hyperparameters

From 6,976 DZs, 2 sexes, 16 age ranges, and 3 prior case states, there were a total of 669,696 cohorts (of which a fraction will have population zero and are excluded). Cohorts from 90% of DZs were randomly sampled for the fit, with 10% reserved to test model performance against data it explicitly did not fit. The fit was made to $\sqrt{\text{cases} + 1}$. The model comprised 500 trees, with data rows sampled for each tree weighted by cohort population. 5 variables were tested at each split, and each tree had a maximum of 30,000 terminal nodes, with a minimum node size of 300. This choice of hyperparameters maximises accuracy at a whole-DZ level across areas with both high and low case numbers, both on the data the model were informed by, and the "unseen" data.

The model is fit in *R* (version 4.1.0), using the *randomForest* package [18] (version 4.6-14), and ALEs analysed using the *ALEPlot* package [20] (version 1.1).

B.4 Vaccination uptake as an explanatory variable

Scotland's COVID-19 vaccination programme began on December 8th 2020, with initial priority given to healthcare workers, the elderly and those otherwise especially vulnerable to COVID-19, then generally by decreasing age [37]. All first doses had been offered and administered to willing adults by 18th July 2021 [38], with rates of first dose administration declining thereafter. By 15th November 2021, then, the first dose date may have differed between two individuals by up to 11 months. This likely led to substantial variation in protection offered by the first dose at the time of the Omicron wave, given both evidence of efficacy waning over timescales of six months, and high rates of breakthrough for Omicron against vaccines originally designed against earlier "wild-type" SARS-CoV-2 lineages, particularly for non-mRNA vaccines [39, 40, 41]. This, combined with high uncertainty in the cohort-level population denominator used to determine uptake, leads us to exclude first and second dose uptake (being highly correlated with first dose uptake) as a predictor for Omicron cases. We do, however, include third/booster dose uptake, as the proportion receiving a first dose to have *returned* for a third/booster dose by 15th November 2021 (and zero if nobody in the cohort had yet received a first dose). This definition eliminates uncertainty in the underlying population. Prior to the detection of Omicron, those aged 50+ or otherwise vulnerable to COVID-19 were due to be offered a third or booster dose, twelve weeks after their second [42]. The booster programme began on September 20th 2021, and a snapshot on 15th November 2021 shows substantial variation between different cohorts, particularly by age. With these doses being delivered more recently, as well as evidence of this dose proving more protective against Omicron [39, 43], we include this definition of third/booster dose uptake as a reasonable proxy for vaccine-induced protection against Omicron at the time.

The initial Delta wave occurred while the bulk of first and second doses were still being administered, thus we include second dose uptake on 1st May 2021 as a predictor, as the proportion of individuals that had returned for a second dose, having received a first (and zero, if nobody in the cohort had yet received their first dose).

C Random forest variable importance

Figures C.5 and C.6 show feature importance measures extracted from the *RandomForest* function. Age, population and prior case status have much higher node purity than the other variables, indicating that splits in individual trees using values of these variables in particular are characteristically more "effective" at separating cohorts with fewer cases from those with more cases. These are consistent with explicit variation in Fig. 1 with respect to age and prior case status (and population trivially). Fig. C.6 then shows random permutation of each of the variables results in appreciable increase in fit error, confirming that this larger collection of variables are important to explain finer patterns in the data.

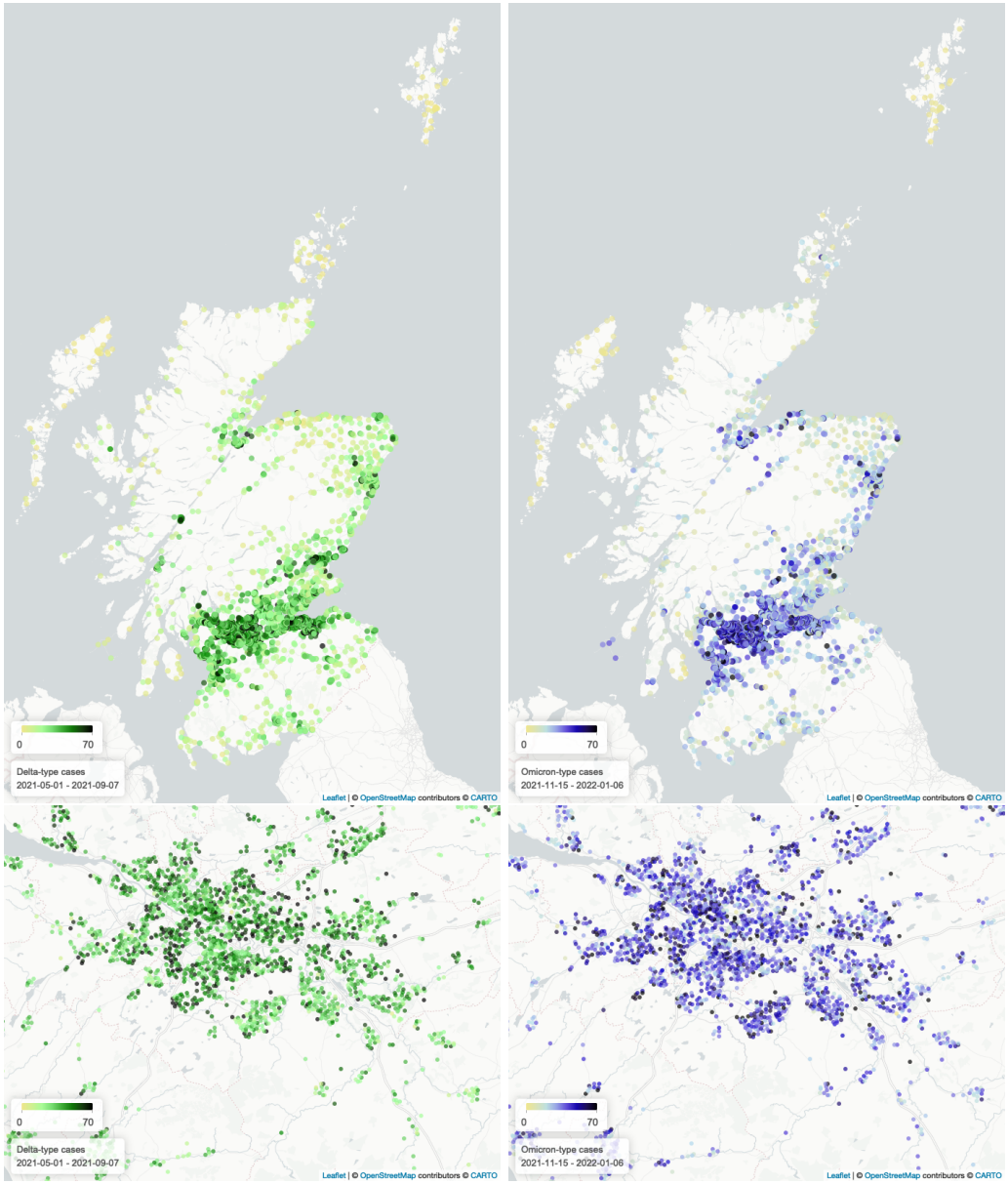

Figure A.4: Omicron-type COVID-19 cases in Scotland (top) between 15th November 2021 and 6th January 2022 (right), compared to Delta-type cases between 1st May and 7th September 2021 (left), as well as focus on the Greater Glasgow region (bottom). Each point indicates the population *centroid* of a DZ, with the colour representing the number of cases reported. The dense clustering of DZs in the south is the *central belt*.

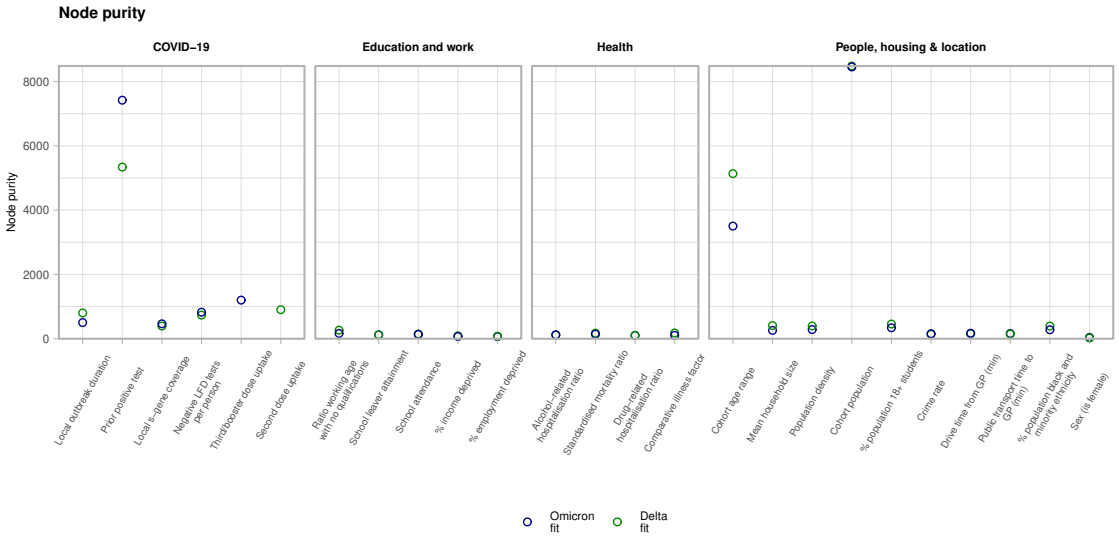

Figure C.5: Explanatory variable node purity.

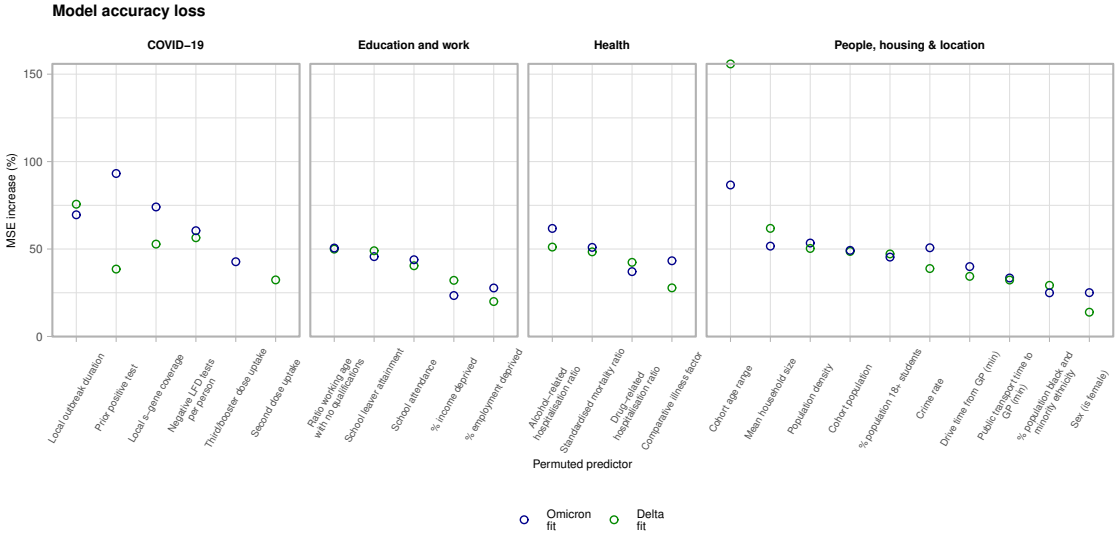

Figure C.6: Explanatory variable mean squared error (MSE) increase on random permutation: for variable i, the increase in MSE on data not trained in each tree, if the entries of i were instead randomly permuted.