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

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

August 3, 2022

Abstract

For a disease such as COVID-19, it is important to identify individuals in a population at heightened risk of infection, as well as broader patterns of infection spread. This is both to estimate burden on healthcare systems (given substantial variation in disease severity from person to person), and to better control the spread of infection. In Scotland, the circulation of SARS-CoV-2 continues to place sustained pressure on healthcare systems, even after a comprehensive vaccination programme and earlier strict non-pharmaceutical interventions.

To better understand individuals at heightened risk, we analyse the spatio-temporal distribution of over 450,000 cases of COVID-19 registered in Scotland in the waves of the B.1.1.529 *Omicron* lineage from November 2021, and an earlier wave of the B.1.617.2 *Delta* lineage from May 2021. These cases are taken from a uniquely fine scale national data set specifying individual tests. We use random forest regression on local case numbers, informing the model with measures relating to local geography, demographics, deprivation, COVID-19 testing and vaccination coverage. We can then identify broader risk factors indicative of higher case numbers. Despite the Delta and Omicron waves occurring around six months apart, with different control measures and immunity from vaccination and prior infection, the overall risk factors remained broadly similar for both.

We find that finer details and clusters in the case distribution are only adequately explained when incorporating a combination of all these factors, implying that variation in COVID-19 cases results from a complex interplay of individual-level behaviour, existing immunity, and willingness to test for COVID-19 at all. On comparing testing patterns to subsequent COVID-19 hospitalisations, we conjecture that the distribution of cases may not be representative of the wider pattern of *infection*, 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 immediately designated as a variant of concern (VOC). 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 [1, 2]. The Omicron variant became the dominant variant globally in a matter of weeks after its first reporting [3].

While there is now strong evidence that Omicron infection typically leads to less severe illness compared to earlier SARS-CoV-2 variants [4, 5, 6], the rate of its initial spread resulted in the imposition of many measures for its control, and the volume of infection led to pressure on healthcare

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

systems comparable to waves of earlier VOCs. Its potential for both rapid spread and immune evasion remains matters of considerable international concern, providing impetus to understand characteristics of its of spatio-temporal spread.

The first COVID-19 cases with an S-gene result consistent with Omicron (BA.1 sub-lineage) were detected in Scotland 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 were imposed in order to curb the spread, however these mostly took the form of guidance [7, 8], rather than stricter legal restrictions and widespread closures of business seen 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 and island settlements as well as dense urban areas.

In this work, we study the growth rate and spatial patterns of the first 227,286 cases in Scotland identified as Omicron-type, in the 52-day period 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 129-day period between 1st May and 7th September 2021.

Throughout the epidemic, cases have not been evenly distributed amongst the population [9, 10, 11]. The main contribution of this work is to explain variation at the resolution of neighbourhoods. 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 demographics. With this, 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 time-dependent case reproduction number for Omicron far exceeded that of Delta, the risk factors for cases remained broadly consistent.

2 Methodology

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 specimen date, S-gene test result, if known (PCR only), and a pseudonymised ID specifying age range, sex, and residing data zone (*DZ*, a geographic region typically comprising 500–1,000 individuals, with 6,796 DZs in total). These IDs link repeat tests by the same individuals over time. 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. To the best of the authors' knowledge, this is the first study probing the *nationwide* distribution of cases at such scales. Data on more severe outcomes relating to COVID-19 (hospitalisation, ICU admission and mortality) are provided at the same resolution.

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 S-gene dropout cases are assumed Omicron. Remaining S-gene positive cases are presumed to be Delta, consistent with broader nationwide sequence data [3]. Finally, 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 less than 1% of fully sequenced cases in the UK [12], 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 [13], 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 distribution of cases using random forest regression in *R*, using the *randomForest* package [14]. With this, we can fit the distribution without specifying any explicit relation between cases and other variables. These methods are suitable here due to their ability to capture interactions between different (but often highly correlated) variables which may be highly non-linear in nature, as well as random forests' known robustness to overfitting (with random sampling with replacement of the data used to fit each tree), given appropriate hyperparameter choice [15].

We term the group of individuals with a given age range, sex, DZ and when (if at all) they last tested positive as a *cohort*, usually comprising 0–50 individuals. For each of these cohorts, we aggregate the total number of cases of interest from the eDRIS data. For the Omicron outbreak, 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 (truncating here 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 Section B.2.

To assess the explanatory variables, we extract two metrics for importance from the *RandomForest* function output: the node purity, and the loss of model accuracy on random permutation of a predictor's values in the data set. We also calculate the *accumulated local effects* (ALEs) of each predictor. ALEs describe how the RF model fit value changes, in response to changing one predictor variable in isolation, averaged over many different entries in the data set [16]. In this context, they provide an indicator as to whether a variable value is associated with fewer or more cases.

Hyperparameters are detailed in Section 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, and 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).

We have chosen day "zero" for each of these periods as the first date from which there are consistent rises in cases with S-gene results associated with that variant. Per regression on $\log_{10}$(cases) (Figure A.1), Omicron-type cases had a case 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, in terms of new DZs having reported at least one case over time (also Figure A.1). At the scale of Scotland's 32 local authorities (LAs), the time-dependent reproductive number R_t (methodology detailed in Section 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 period in which both variants were circulating, the *age* distributions by variant differed. Taking the mid-points of the five-year age brackets, we see that the mean ages of the Delta-type cases is 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-type accounted for at least 5% of cases (Figure A.2A). However, the median ages are equal (both 32.5 years), because 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-type, but then a peak in the 20–29 age group (Figure A.2B).

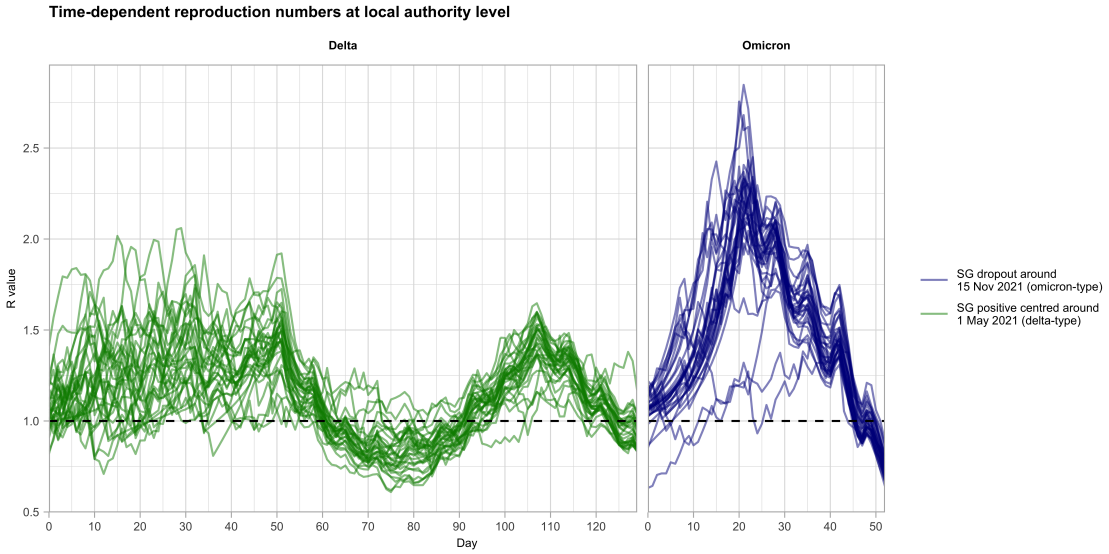

Figure 1: Time-dependent reproduction numbers for the Delta (left) and Omicron waves (right), over each of the 32 separate local authorities (LAs).

3.2 Detailed case distribution analysis

Case data are presented at DZ level in Figure 2, with focus on the dense *Greater Glasgow* area in Figure 3. The *central belt* (containing the two largest cities, Glasgow and Edinburgh) has a clustering of DZs with high case rates, relative to the less dense, more rural regions outside the belt.

Fitting this data using our model, Figure 7 shows model performance at a DZ level, comparing observed cases to fit cases. Points deviating from the diagonal show DZs where the model was less accurate. Our full model explains 70% (fit: 71%, test: 62%) of the DZ-level variation in the data case distribution. As a comparison, we simultaneously fit three "univariate" models, under equivalent setups 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%). Figure 7 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 statistic [17] on the residuals. We compare nearby DZ-aggregated residuals over physical distances (from 1km to 100km), using the population-weighted centroids as the point position of each DZ), as well as network distance (from 1 to 20 nearest neighbours away, defining nearest neighbours as DZs that share a boundary). Figure 6 shows for the full model, mean autocorrelation of residuals within 1km of each other (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 scale of a local authority). For the single variable models, age and deprivation-only models exhibit much higher residual clustering, with the density-only model performing better, but persisting over larger distances. These are all 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).

Figure 4 shows the case distributions over age, sex, prior cases, deprivation and health board. Omicron cases were concentrated in younger adults, with a peak of 90 cases/1,000 in age range 20–24. There was no marked difference in total cases between men and women. Case rates were much lower amongst those that had tested positive for COVID-19 previously.

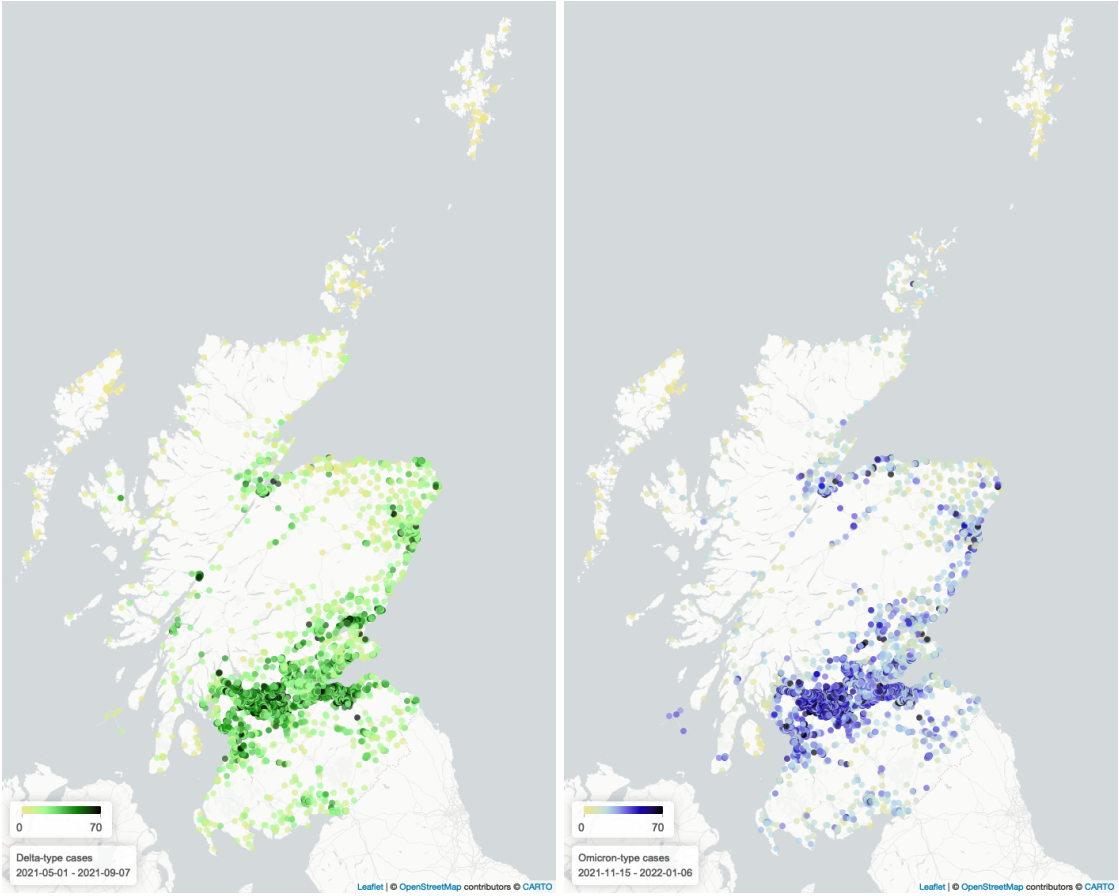

Figure 2: Omicron-type COVID-19 cases in Scotland between 15th November 2021 and 6th January 2022 (right), compared to Delta-type cases between 1st May and 7th September 2021 (left). Each point indicates the population *centroid* of a DZ, with the colour representing the number of cases reported. The dense clustering of DZs in the south is the *central belt*.

There is also variation in case rates over Scotland's 14 health boards, most notably in Orkney, Shetland and the Western Isles (all island communities without a road link to the mainland), as well as a decrease with increasing rurality (in line with Figure 2). Aggregating DZs with respect to deprivation rank gives a bimodal distribution, with higher rates in both the most deprived, and least deprived communities.

The distributions in Figure 4 are reasonably 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).

The accumulated local effects (ALEs) help us to identify *risk factors* (variable values that the model views as associated with more cases), and which emerge as *protective* (with fewer). For the range of explanatory variables in Figure 5, 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*) straightforwardly reflect the distributions in Figure 4.

Low *local S-gene coverage* (see Section B.2 for precise description) is protective. At the time S-gene coverage was broadly consistent for cases in mainland Scotland (90% overall per eDRIS data), but significantly lower in the LAs of Orkney Islands, Shetland Islands and Na h-Eileanan

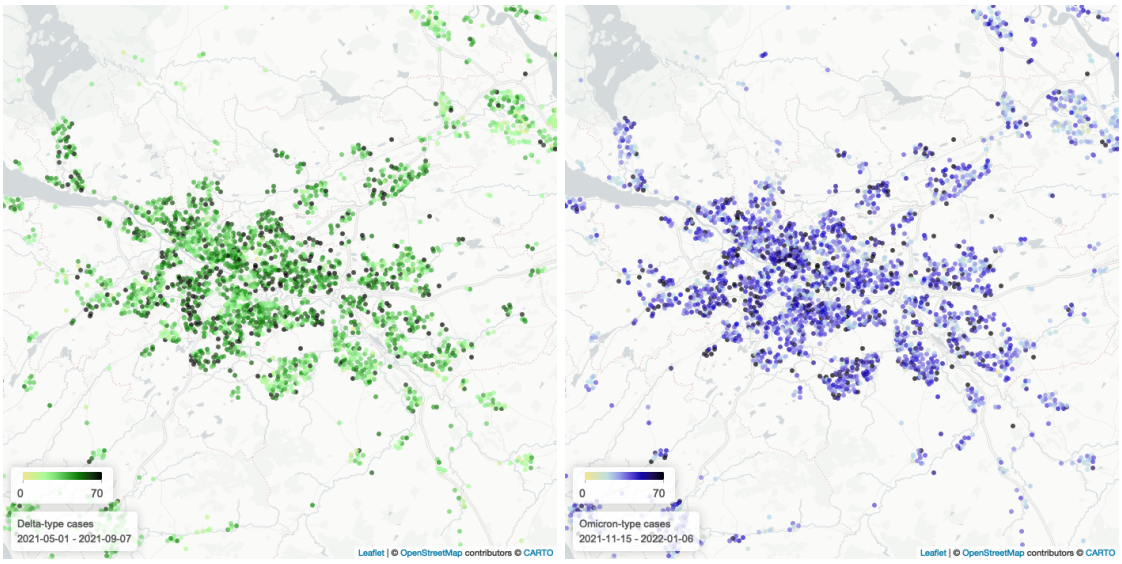

Figure 3: Case distributions from Figure 2, focusing on Greater Glasgow.

Siar (74%, 20% and 23% respectively). Increasing *local outbreak duration* (the time difference between January 6th and the first day an Omicron-type 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 [18, 19], and increased risk of inter-household transmission relative to other close contacts outside of the home [20].

The fit shows that third/booster vaccination uptake has a generally protective ALE for increasing uptake, but *zero* uptake also protective. This should be interpreted with caution, as some younger age groups had zero uptake as they were not yet eligible for a booster, but generally reported fewer cases than the average. See Section 4 for further discussion.

There is an association between cases, and the number of *negative* LFD test results reported per person. This is an important relation which is discussed further in Section 4 in the context of case *ascertainment*; the proportion of infections that eventually manifest as a reported positive test.

Individual indicators associated with deprivation appear less influential. However, deprivation-level variation in case rates in Figure 4, while subtle, is well captured, and the deprivation-associated variables did emerge as important in improving overall model accuracy (Figure C.4). There is strong correlation between different measures of deprivation (as a community suffering with one form of deprivation is more likely to be suffering with others), but also these measures 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 then plausible the fit does not take a single deprivation measure in isolation as influential, but rather a combination of several correlated predictors.

Comparing this to Delta-type cases from 1st May to 7th September 2021, the spatial distribution (Figures 2, 3) shows high case rates again within DZs in the central belt, with evidence of a small number of localised outbreaks in more remote communities. Figure 4 shows cases skewed slightly younger, with the highest rates within ages 15–19. With respect to deprivation, 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 the single exception that third/booster dose uptake was replaced with second dose uptake (third/booster doses were only administered later). Model performance was similar, explaining 72% (fit: 73%,

test: 61%) of DZ-level variation. Figure 5 shows the risk factors for Omicron cases (such as high rates of negative LFD test reporting, or high mean household sizes) broadly emerged as risk factors for Delta as well. With the exception of zero uptake (which is likely associated with the very young), higher second dose uptake emerged protective. A notable difference from Omicron is that cohorts with high 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 distribution of hospital admissions in the same period

Included alongside the case distribution in Figure 4 is the distribution of COVID-19 related hospital admissions, spanning the same periods, lagged by one week. 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, rather are presented to show only the *distribution* of admissions.

For both periods, admissions were highly skewed towards older age groups. This is consistent with age-dependent rates of severe COVID-19 outcomes observed across previous stages of the epidemic [21, 22]. 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 hospital admissions did not have an associated second case, in the past.

In both periods there was a strong deprivation skew. For Omicron, the bottom 50% of DZs by deprivation reported 63% of admissions, and for Delta, the proportion was 64%. Given this pattern is not seen in cases (rather, for the subset of cases presented the distribution is bimodal), the skew likely manifests from variation in both infection-to-hospitalisation rates, and case ascertainment, with respect to deprivation. Understanding the relative contributions of these two factors is central to understanding the overall pattern of infection, and we reserve further discussion for Section 4.

4 Discussion

The detection of the Omicron variant in Scotland in November 2021 preceded an explosive wave of cases, consistent with its enhanced ability 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 level of neighbourhoods.

Our spatial distribution model shows that variation in case counts is associated with broad factors such as age structure, but also with factors relating to testing, vaccination, geography and demography, all of which vary on much finer spatial scales. We have shown that the full spatial distribution of COVID-19 cases is only reasonably explained after considering a combination of all of these variables.

At the individual level, the fundamental 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 being infected, if exposed. Finally, the propensity of individuals to test, and report the results of those tests, affects whether an infection actually manifests as a registered case. Evidence from random community testing suggests that only a minority of infections do so [23, 24]. Our distribution model uses explicit data on COVID-19 vaccination and tests, as well as data on age, sex, geography and deprivation 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 spatio-temporal analyses provide some important insights.

Living in a deprived community has been hypothesised as a risk factor for COVID-19 infection (e.g., [25, 26, 27]), however in our data this does not manifest in case numbers. A clear skew is only seen in admissions, where approximately two thirds are reported from the bottom half of neighbourhoods when ranked by deprivation. From the data used to inform the model, we suggest this skew arises from a combination of an inherently higher risk of hospitalisation

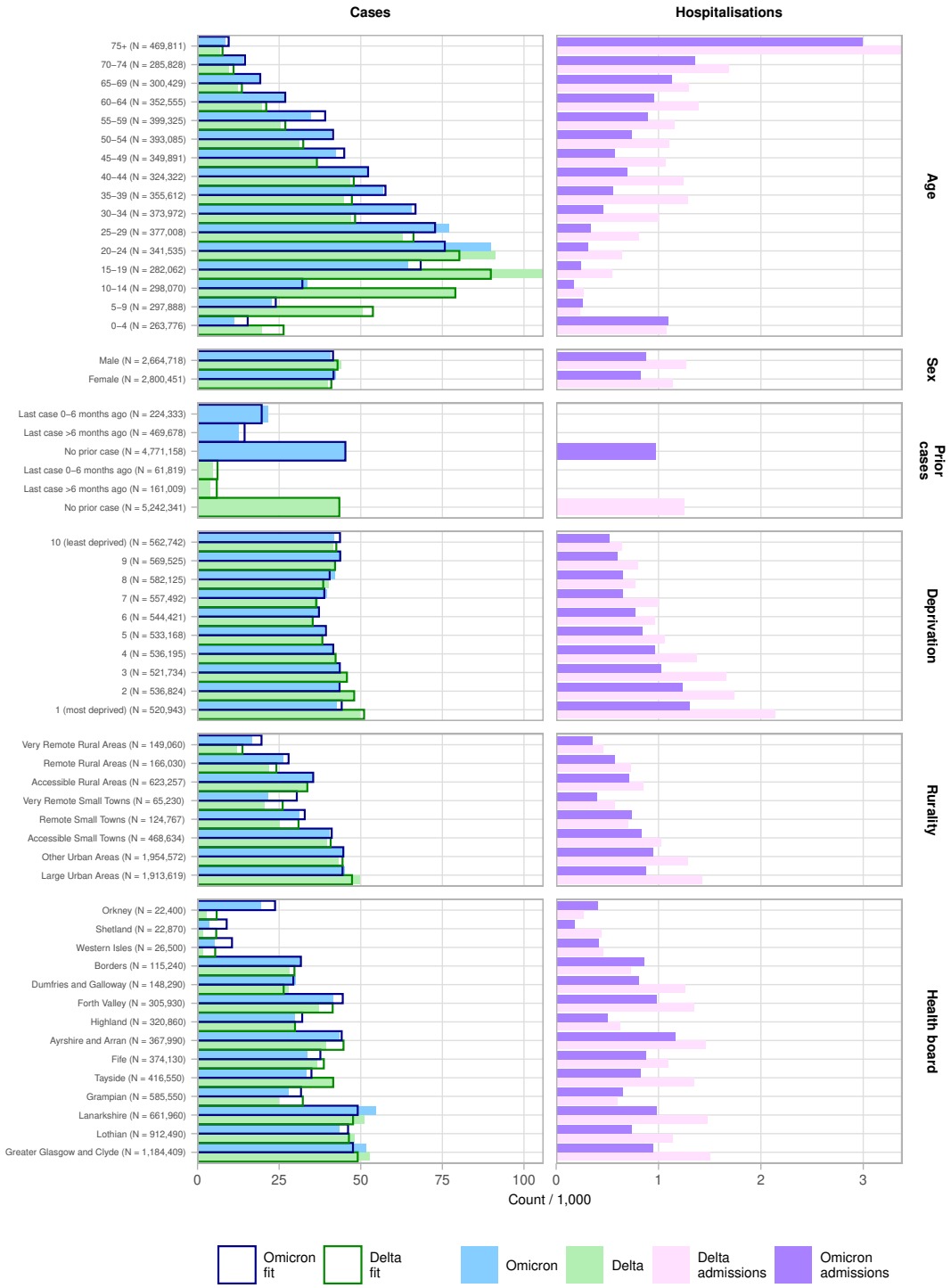

Figure 4: Cases (left) and hospitalisation admissions (right) by age, prior cases, deprivation, rurality and location. Note that hospital admissions count all COVID-19 admissions in that period, and may be related to cases excluded from these analyses due to the lack of an S-gene result, or may not have an associated case at all.

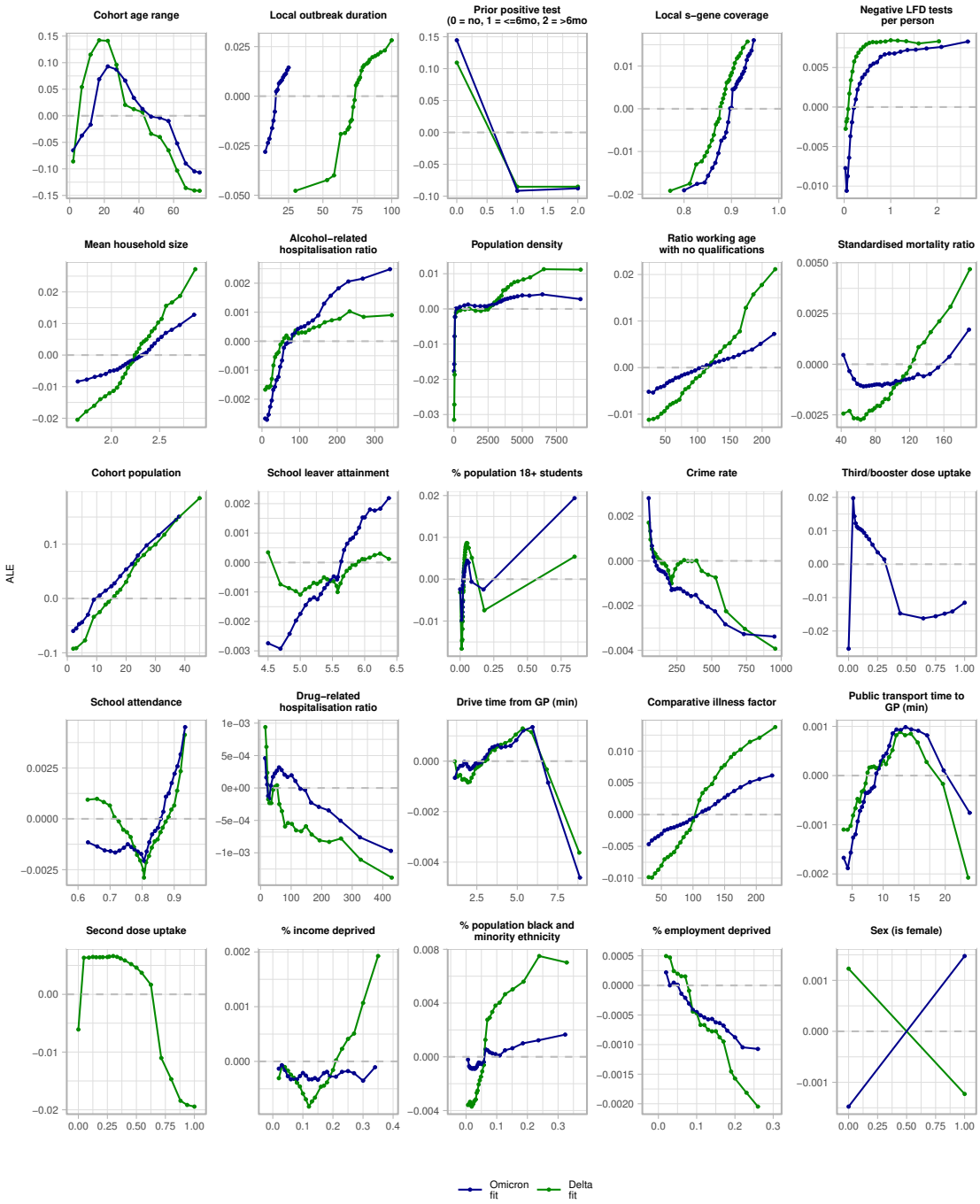

Figure 5: Accumulated local effects across all explanatory variables. For each variable, the x-axis represents the range of values of that variable in the data, and the y-axis is the univariate ALE for that variable value.

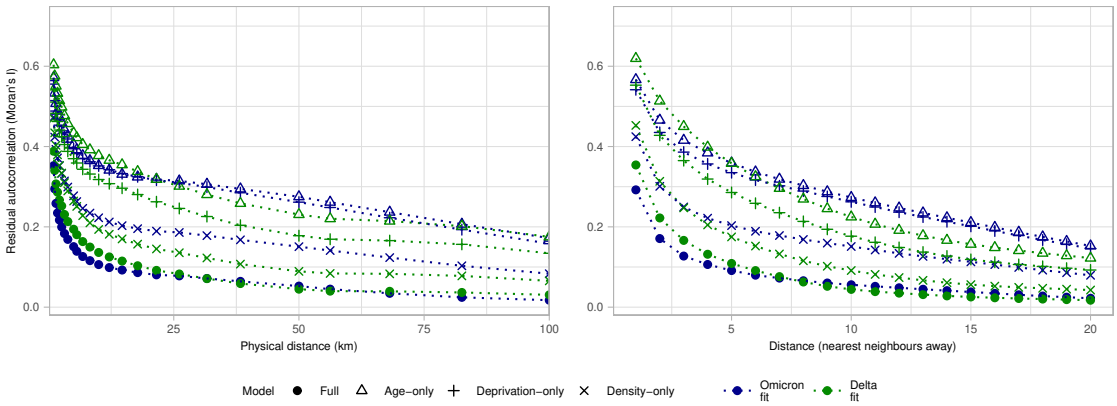

Figure 6: Residual clustering as measured by the Moran's I statistic, at different physical (left) and network-based distances (right). Higher values represent higher autocorrelation between residuals within a given locus.

given *infection* with higher deprivation, but also a potential deprivation-level variation in case *ascertainment*. From Spring 2021 to Spring 2022, LFD tests were abundantly available in Scotland, free-of-charge. Individuals were encouraged to test and report results regularly regardless of symptoms or outcome. Thus, with PCR testing mainly reserved for symptomatic testing, the reporting of LFD negatives may relate to propensity for testing overall, and higher reporting of asymptomatic or mild infections that would not result in hospital admission. Per the eDRIS data (e.g. [28]), rates of LFD reporting have consistently been lower in more deprived areas, and LFD positivity higher. This is consistent with an earlier mass LFD testing pilot in Liverpool, England, where participation fell from 58.0% in the least deprived quintile of neighbourhoods, to 38.4% in the most [29]. This suggests the skew in case-hospitalisation rates may not entirely be from variation in the *infection*-hospitalisation rate.

The ALE (Figure 5) for negative LFD test reporting exhibits a plateau as the number of tests per person increases. Aggregating PCR positive test results (any S-gene result) by LFD tests reported in each period, individuals that had reported no LFD test 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%). This peaking at 1–2 LFD tests implies a complex interaction between risk of infection and test-seeking behaviour, at the individual level, and that LFD reporting rates may not have a direct linear relationship with case ascertainment.

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 as indicative of fewer cases at local level, but the inverse when a cohort has *exactly* zero uptake. We suspect this is an artefact arising from differences in eligibility; in both periods studied, eligibility was not even across the population (Section B.4). Cohorts with zero uptake will then likely be not (yet) eligible for vaccination, specifically those with age ranges 0–14; ages that reported fewer cases generally, but also may also be protected for other reasons associated with young age. In the data vaccinations and cases are not individually linked, meaning we do not have the exact vaccination status of those reporting cases. Had they been linked, we may have been able to capture a more direct association with vaccination status.

While the Omicron wave was much more rapid than earlier waves, Figure 5 shows that the risk factors for testing positive were broadly similar to those for testing positive for Delta, after analysing a similar number of cases from the initial outbreak (albeit over a longer time period). This is despite material differences in time of year, imposed NPIs, and overall outbreak trajectories

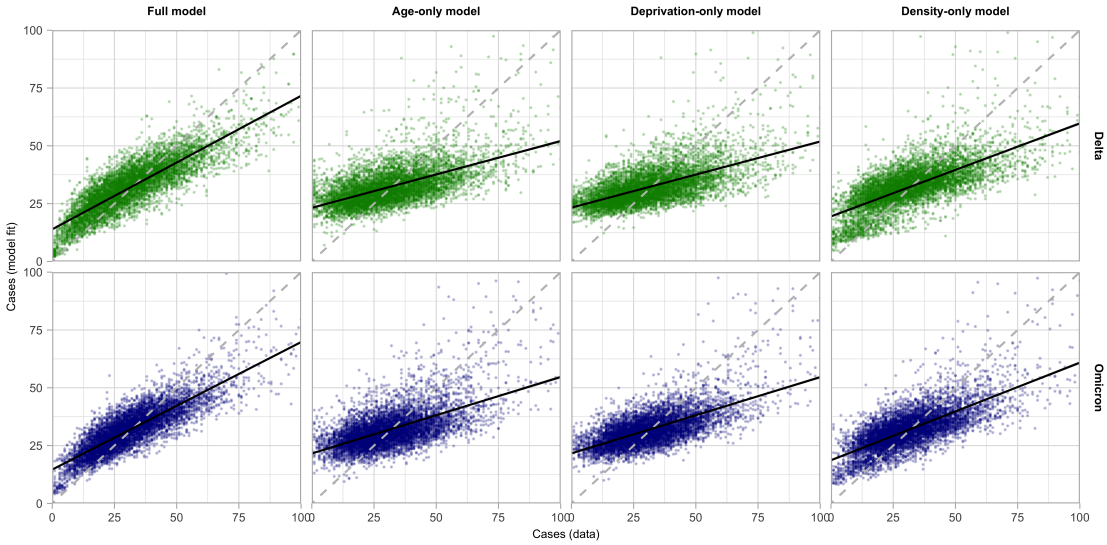

Figure 7: Performance of different models in terms of modelled cases (y-axis) in comparison with the observed cases (x-axis), for both Omicron and Delta. Each point represents a DZ. Also included are comparisons to models informed with only population, and one of either age, overall deprivation rank, or population density.

(as per the time-dependent reproduction numbers in Figure 1, with the Omicron outbreak being more rapid and more consistent across different local authorities, as compared to Delta).

The model is by design not informed with position data, nor by network characteristics (e.g., data on nearest neighbours). It is likely that the number of cases recorded by an area is influenced not just by the characteristics of that area, but also those 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 study cases amongst a population of 5.5 million, at the level of 0–50 individuals. However, regardless of resolution, cases only partially represent the full underlying pattern of infection, and variation in testing behaviour suggests the distribution of infections may be characteristically 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 are distributed amongst the population.

5 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] Dandan Tian, Yanhong Sun, Huihong Xu, and Qing Ye. The emergence and epidemic characteristics of the highly mutated sars-cov-2 omicron variant. *Journal of Medical Virology*, 94(6):2376–2383, 2022.

[2] Wanwisa Dejnirattisai, Jiandong Huo, Daming Zhou, Jiří Zahradník, Piyada Supasa, Chang Liu, Helen ME Duyvesteyn, Helen M Ginn, Alexander J Mentzer, Aekkachai Tuekprakhon, et al. Sars-cov-2 omicron-b. 1.1. 529 leads to widespread escape from neutralizing antibody responses. *Cell*, 2022.

[3] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[4] Nicole Wolter, Waasila Jassat, Sibongile Walaza, Richard Welch, Harry Moultrie, Michelle Groome, Daniel Gyamfi Amoako, Josie Everatt, Jinal N Bhiman, Cathrine Scheepers, 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.

[5] Lamprini Veneti, Håkon Bøås, Anja Bråthen Kristoffersen, Jeanette Stålcrantz, Karoline Bragstad, Olav Hungnes, Margrethe Larsdatter Storm, Nina Aasand, Gunnar Rø, Jostein Starrfelt, 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*, 27(4):2200077, 2022.

[6] Tommy Nyberg, Neil M Ferguson, Sophie G Nash, Harriet H Webster, Seth Flaxman, Nick Andrews, Wes Hinsley, Jamie Lopez Bernal, Meaghan Kall, Samir Bhatt, 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*, 399(10332):1303–1312, 2022.

[7] Scottish Government. Nightclubs to close. https://www.gov.scot/news/nightclubs-to-close/.

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

[9] Duncan Lee, Chris Robertson, and Diogo Marques. Quantifying the small-area spatio-temporal dynamics of the covid-19 pandemic in scotland during a period with limited testing capacity. *Spatial Statistics*, page 100508, 2021.

[10] Christopher J Banks, Ewan Colman, Thomas Doherty, Oliver Tearne, Mark Arnold, Katherine E Atkins, Daniel Balaz, Gaël Beaunée, Paul R Bessell, Jessica Enright, et al. Scovmod–a spatially explicit mobility and deprivation adjusted model of first wave covid-19 transmission dynamics. *Wellcome Open Research*, 7(161):161, 2022.

[11] Nazir I Lone, Joanne McPeake, Neil I Stewart, Michael C Blayney, Robert Chan Seem, Lorraine Donaldson, Elaine Glass, Catriona Haddow, Ros Hall, Caroline Martin, 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*, 1:100005, 2021.

[12] UK Health Security Agency. Sars-cov-2 variants of concern and variants under investigation in england: technical briefing 35. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1050999/Technical-Briefing-35-28January2022.pdf.

[13] Scottish Government. Self-isolation and testing changes. https://www.gov.scot/news/self-isolation-and-testing-changes/.

[14] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Dan Apley and Maintainer Dan Apley. Package 'aleplot'. 2018.

[17] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[18] Neda Jalali, Hilde Kjelgaard Brustad, Arnoldo Frigessi, Emily Ann MacDonald, Hinta Meijerink, Siri Laura Feruglio, Karin Maria Nygård, Gunnar Øivind Isaksson Rø, Elisabeth Henie Madslien, and Birgitte Freiesleben De Blasio. 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.

[19] Jannik Fonager, Marc Bennedbæk, Peter Bager, Jan Wohlfahrt, Kirsten Maren Ellegaard, Anna Cäcilia Ingham, Sofie Marie Edslev, Marc Stegger, Raphael Niklaus Sieber, Ria Lassauniere, 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*, 27(10):2200181, 2022.

[20] Julien Dupraz, Audrey Butty, Olivier Duperrex, Sandrine Estoppey, Vincent Faivre, Julien Thabard, Claire Zuppinger, Gilbert Greub, Giuseppe Pantaleo, Jérôme Pasquier, 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*, volume 8, page ofab149. Oxford University Press US, 2021.

[21] Public Health Scotland. Covid-19 statistical report, as at 15 november 2021. https://www.publichealthscotland.scot/publications/covid-19-statistical-report/covid-19-statistical-report-17-november-2021/.

[22] Office for National Statistics. Coronavirus (covid-19) weekly insights: latest health indicators in england, 12 february 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19weeklyinsights/latesthealthindicatorsinengland12february2021.

[23] Ewan Colman, Jessica Enright, Gavrila A. Puspitarani, and Rowland R. Kao. Estimating the proportion of sars-cov-2 infections reported through diagnostic testing. *medRxiv*, 2021.

[24] Emily S Nightingale, Sam Abbott, and Timothy W Russell. 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*, 22(1):1–14, 2022.

[25] Richard Blundell, Monica Costa Dias, Robert Joyce, and Xiaowei Xu. Covid-19 and inequalities. *Fiscal studies*, 41(2):291–319, 2020.

[26] Clare Bambra, Ryan Riordan, John Ford, and Fiona Matthews. The covid-19 pandemic and health inequalities. *J Epidemiol Community Health*, 74(11):964–968, 2020.

[27] Jose Miguel Baena-Díez, María Barroso, Sara Isabel Cordeiro-Coelho, Jorge L Díaz, and María Grau. Impact of covid-19 outbreak by income: hitting hardest the most deprived. *Journal of Public Health*, 42(4):698–703, 2020.

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

[29] Mark A Green, Marta García-Fiñana, Ben Barr, Girvan Burnside, Christopher P Cheyne, David Hughes, Matthew Ashton, Sally Sheard, and Iain E Buchan. 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*, 6:100107, 2021.

[30] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516, 2004.

[31] National Records of Scotland. Mid-2020 small area population estimates for 2011 data zones. 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/.

[32] Scottish Government. Scottish index of multiple deprivation 2020. https://www.gov.scot/publications/scottish-index-of-multiple-deprivation-2020v2-indicator-data/.

[33] Scottish Government. Coronavirus (covid-19): vaccine deployment plan 2021. https://www.gov.scot/publications/coronavirus-covid-19-vaccine-deployment-plan-2021/.

[34] Scottish Government. Major milestone in vaccination programme. https://www.gov.scot/news/major-milestone-in-vaccination-programme/.

[35] Nick Andrews, Julia Stowe, Freja Kirsebom, Samuel Toffa, Tim Rickeard, Eileen Gallagher, Charlotte Gower, Meaghan Kall, Natalie Groves, Anne-Marie O'Connell, et al. Covid-19 vaccine effectiveness against the omicron (b. 1.1. 529) variant. *New England Journal of Medicine*, 2022.

[36] Sandile Cele, Laurelle Jackson, David S Khoury, Khadija Khan, Thandeka Moyo-Gwete, Houriiyah Tegally, James Emmanuel San, Deborah Cromer, Cathrine Scheepers, Daniel G Amoako, et al. Omicron extensively but incompletely escapes pfizer bnt162b2 neutralization. *Nature*, 602(7898):654–656, 2022.

[37] Eleftheria Vasileiou, Colin R Simpson, Ting Shi, Steven Kerr, Utkarsh Agrawal, Ashley Akbari, Stuart Bedston, Jillian Beggs, Declan Bradley, Antony Chuter, 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*, 397(10285):1646–1657, 2021.

[38] Cabinet Secretary for Health and Social Care. Scotland's autumn/ winter vaccination strategy 2021. 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.

[39] Aziz Sheikh, Steven Kerr, Mark Woolhouse, Jim McMenamin, and Chris Robertson. Severity of omicron variant of concern and vaccine effectiveness against symptomatic disease: national cohort with nested test negative design study in scotland. 2021.

A Timeseries evolution of Omicron and Delta waves

Supplementary plots of the time evolution of the Omicron and Delta waves are presented in Figures A.1, A.2.

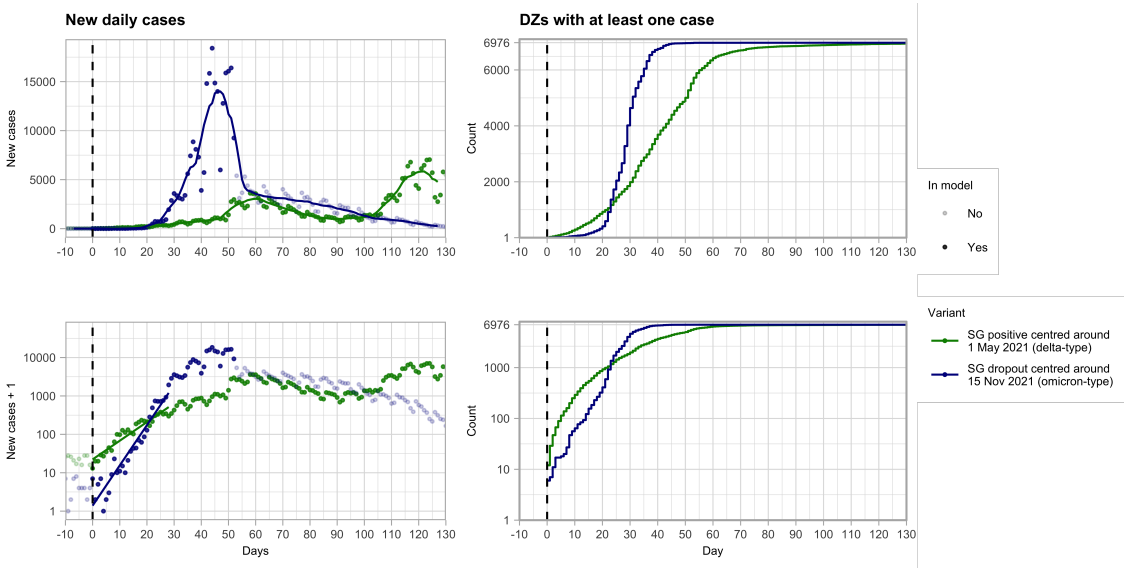

Figure A.1: Timeseries of the initial outbreaks of the Delta and Omicron variants.

B Methodology

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* as a result of 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 days [30]:

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

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 of infectiousness obtained from the *SCoVMod* compartmental model (for more detail see Reference [10]).

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.

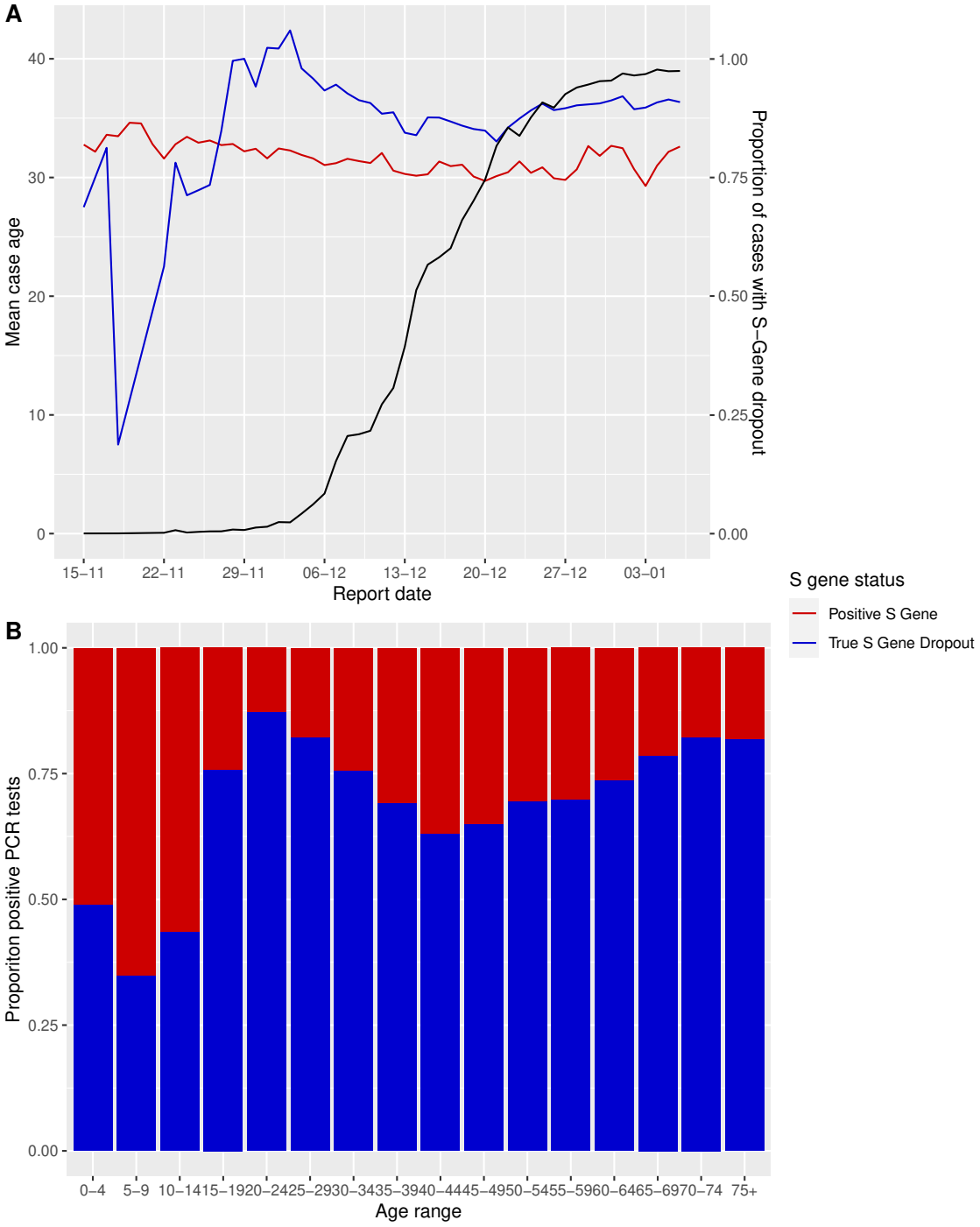

Figure A.2: PCR positive cases over the period 15th November 2021 to 6th January 2022 that were S-gene dropout or true S-gene positive. (A) Daily mean case age for the two definite PCR S-gene outcomes (blue and red lines) against the proportion of the daily cases that were true S-gene dropout (presumed Omicron-type). (B) the proportion of the cases over the period by 5-year age bracket.

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 (5 year window, 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 the 2020 *Small Area Population Estimates* collated by the National Records of Scotland [31] for the remainder that have not tested positive before).

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

- COVID-19 vaccination uptake (eDRIS) (see Section B.4 for details);
- 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* [32]);
- 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;
- IZ-level population density.

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 [13]). 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. 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 RF 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 [14], and ALEs analysed using the *ALEPlot* package [16].

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 [33]. All first doses had been offered and administered to willing adults by 18th July 2021 [34], 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 [35, 36, 37]. 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 discovery of Omicron, those aged 50 and above or otherwise vulnerable to COVID-19 were due to be offered a third or booster dose, at least twelve weeks after their second [38]. 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 [35, 39], we instead include this definition of third/booster dose uptake as a reasonable representation of 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.3 and C.4 show feature importance measures extracted from the *RandomForest* function the model is fit using. 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 Figure 4 with respect to age and prior case status (and population trivially). Figure C.4 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.

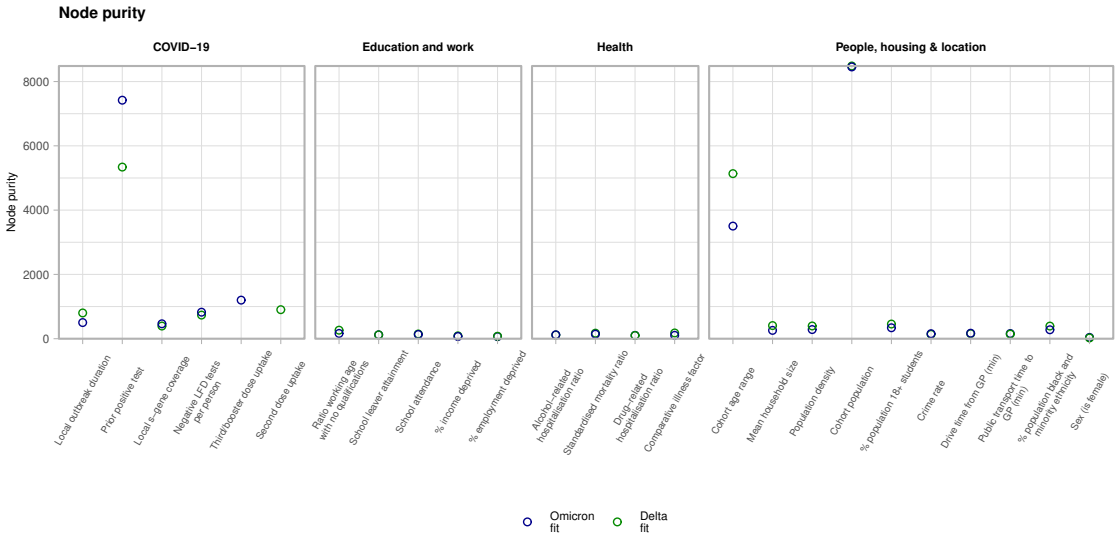

Figure C.3: Explanatory variable node purity.

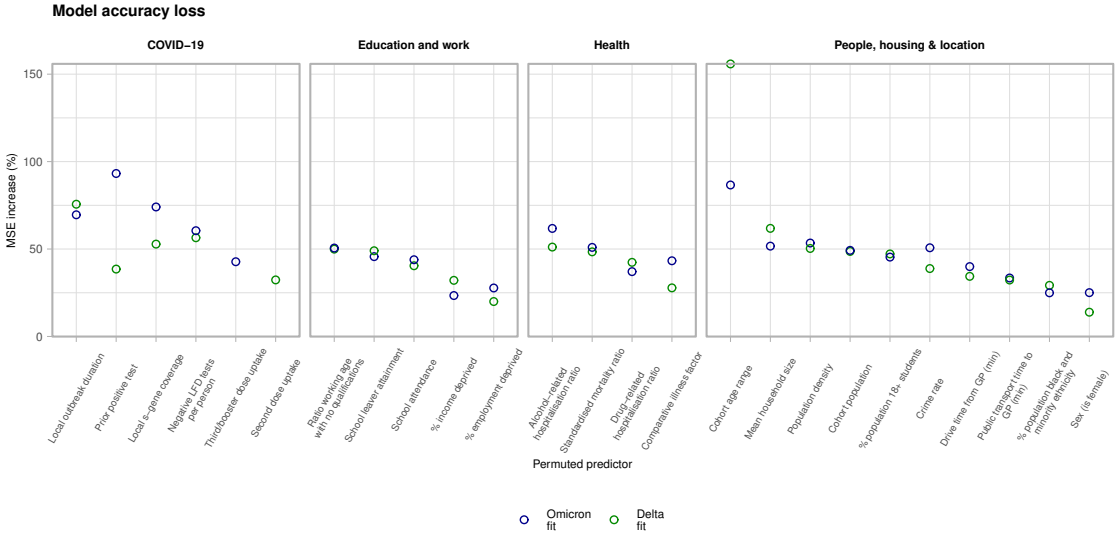

Figure C.4: Explanatory variable mean squared error (MSE) increase on random permutation: for variable i, the increase in MSE on data not trained in each tree, if the entries of i were instead randomly permuted.