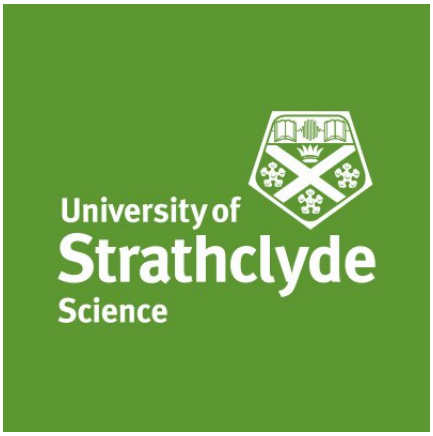
Ecological studies of *Clostridioides difficile* and COVID-19 infection with the application of space-time risk models



University of Strathclyde
Department of Mathematics and Statistics

Florence Tydeman

Doctor of Philosophy
September 2021

Declaration

I, Florence Tydeman, declare that the contents within this thesis, titled 'Ecological studies of *Clostridiodies difficile* and COVID-19 infection with the application of space-time risk models', is my own work. I confirm that I have clearly referenced any work published by others.

This thesis is the product of a 4-year EPRSC fully funded doctoral training programme which provided opportunities to experience additional training and innovation placements. I took part in two placements, which have resulted in 4 papers accepted for publication or published so far. The first paper is from the work presented in Chapter 6: a project on *Clostridiodies difficile* with the University of Bristol (for which I was granted Honorary Status at Bristol Medical School) in collaboration with Public Health Wales (PHW), which is published in the Journal of Antimicrobial Chemotherapy [1].

The second placement was with the Academic Department of Women and Children's Health, Guy's and St Thomas' Hospital and King's College London, on a project of childhood adiposity, published in Pediatric Obesity [2]. Finally, I was involved in a study on Hyperemesis Gravidarum with King's College London using data from a BBC survey, which is published in the American Journal of Obstetrics and Gynecology [3], with another paper from these data recently accepted for publication in Obstetric Medicine: 'Termination of wanted pregnancy and suicidal ideation in Hyperemesis Gravidarum: a mixed methods study.' The work relating to previous 3 papers are not presented as chapters in this thesis as they are not in line with the objectives of the main body of work.

1. Florence Tydeman, Noel Craine, et al, Incidence of Clostridioides difficile infection (CDI) related to antibiotic prescribing by GP surgeries in Wales, Journal of Antimicrobial Chemotherapy, Volume 76, Issue 9, September 2021, 2437–2445

2. Dalrymple KV, Tydeman FAS et al and the UPBEAT consortium. Adiposity and cardiovascular outcomes in three-year-old children of participants in UPBEAT, an RCT of a complex intervention in pregnant women with obesity. Pediatr Obes. 2021 Mar;16(3) e12725

3. Nana M, Tydeman F, Bevan G, Boulding H, Kavanagh K, Dean C, Williamson C. Hyperemesis gravidarum is associated with increased rates of termination of pregnancy and suicidal ideation: results from a survey completed by $>$ 5000 participants. Am J Obstet Gynecol. 2021 Jun;224(6):629-631.

I was invited to present the work of Chapter 3 at the GEOMED conference, University of Glasgow, where I was awarded 'Best Student Talk' in 2019. This work was also presented at the Research Students' Conference in Probability and Statistics (RSC) and Young Statisticians Meeting (YSS) conferences in 2018. Finally, the work from Chapter 7 was presented at the Royal Statistical Society (RSS) conference, September 2021.

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Chris Robertson and Dr Kim Kavanagh, for your constant guidance and advice. You have both helped me develop as a researcher and I sincerely appreciate all of your support and encouragement throughout this experience.

I would also like to extend my appreciation to the Department of Women and Children's Health, at Guy's and St Thomas' hospital. The work of this department inspires me and I would like to thank them for all of the opportunities they given me. Similarly, to the Department of Population Health, University of Bristol, for providing me with the opportunity to broaden my knowledge and academic network.

I would also like to thank my friends in the department at the University of Strathclyde for keeping me smiling throughout, and particularly to Kate for being my statistical fairy godmother. I feel lucky to have had you as a friend throughout this experience. You were always there with a cup of tea when I needed it and for that, I am grateful.

Thank you to my friends and my family, who have given me endless support. To Alice and Jodie who have been there for all the highs and the lows. To Mum, Daisy and Doris for your constant reassurance - I promise I will stop phoning you all so much and to Eric for always keeping me grounded.

Finally, I want to thank my dad. For your constant guidance, relentless optimism and unwavering belief in me. I hope you know how grateful I am for your never-ending support. You really are marvellous.

Abstract

Infectious diseases continue to pose major global health threats. With the recent devastation from the COVID-19 pandemic and growing concerns of healthcare-associated infections (HAIs), there is a worldwide requirement for stringent techniques to monitor and understand the key drivers for infections. Infectious diseases have an inherent spatial dimension due to the contagious nature of viruses and bacteria. This thesis aims to explore the use of spatial and spatio-temporal techniques applied to infections, specifically *Clostridiodies difficile* infection (CDI) and COVID-19, to identify risk factors at an ecological population-based level. A mixture of open-sourced and routinely collected data, at different spatial scales, were used to understand the surveillance capacities of observational public health data.

Antimicrobial prescribing and stewardship have been a global focus in the last decade as concerns have grown with emergent novel antibiotic-resistant infections. CDI has been shown to have a well-defined association with certain broad-spectrum antibiotic classes and other environmental factors, however, there is a gap in the literature aiming to understand these relationships ecologically and spatially. The main focus of this thesis was to use spatio-temporal models to investigate spatial risk factors of CDI incidence, such as GP antimicrobial prescribing, in Scotland and Wales. Similar spatial techniques were then applied to investigate the spatial distribution of COVID-19 testing during the first wave of the 2020 epidemic in Scotland. The relevant spatial and spatio-temporal models applied throughout this thesis were initially discussed in Chapter 2.

The spatial distribution of Scottish GP antibiotic prescribing rates, from 2016 to 2018, was investigated in Chapter 3 using spatial point-location correlation methods. Risk factors of increased GP antibiotic prescribing were explored, showing GP practice de-

mographic information as key drivers of increased antibiotic prescribing. These analyses were followed by an exploration of Scottish CDI incidence data, from 2014 to 2018, at a small areal level (intermediate zones (IZ)), to understand spatial auto-correlation and temporal trends of CDI incidence in Chapter 4. Population demographic risk factors, as highlighted in the literature, were obtained at the same spatial scale and assessed as ecological risk factors of CDI incidence using conditional autoregressive (CAR) models.

The next phase of this thesis then combined the previous two analyses, introducing a multi-level spatial problem, which aimed to explore central risk factors of CDI that were not available at the same spatial scale in Chapter 5. Spatial interpolation methods were applied to manipulate GP antibiotic prescribing point-location data and areal-unit cattle density data to match the CDI incidence at an IZ spatial scale. These data could then be explored as ecological risk factors of CDI incidence, carrying forward the previously defined CAR model from Chapter 4 and adjusting for demographic confounders.

Welsh CDI incidence and primary care antibiotic prescribing data offered the opportunity to compare between two countries in the UK. The retrospective ecological study in Chapter 6 used aggregated disease surveillance data to understand the impact of total and high-risk Welsh GP antibiotic prescribing on total and stratified inpatient/non-inpatient CDI incidence. Location and health board information were anonymised preventing a formal spatial analysis, however, the results were comparable to previous chapter findings and supported the hypothesis of an increased risk of CDI incidence reflected in GP antibiotic prescribing rates, particularly high-risk antibiotics, and population demographics.

Finally, at the beginning of the COVID-19 pandemic, it became evident that the methodologies applied in this thesis could support the investigation of the spread of COVID-19 infections. The work presented in Chapter 7 aimed to explore how best

to capture spatial patterns of community COVID-19 infection by conducting a spatio-temporal analysis on three data streams – positive test rates, relevant NHS24 calls and COVID Symptom Study (CSS) predicted cases, to assess which was best for early disease surveillance. Results showed both sources to identify similar trends of COVID-19 and gold-standard testing data, particularly when used in parallel.

This thesis has provided new insights into the associated risks between CDI incidence and GP antibiotic prescribing in Scotland and Wales, demonstrating the capabilities of open-source and routinely collected public health data when applied in a spatial framework. These results support the requirement of stringent measures to reduce antibiotic prescribing in the community. It also highlights the beneficial use and suitability of analysing infectious disease data with spatial techniques to address gaps in the literature to understand population-based risk factors of disease. There is a strong argument for future research into methods of analysing multi-level spatial data, particularly in the application of observational public health data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An infection is the invasion or multiplication of microorganisms, where an infectious disease is the illness caused by an infection. The infectious process can be explained by a series of steps: beginning with a pathogenic causal organism such as bacteria, fungi, viruses or parasites. The organism must be in sufficient number and of sufficient virulence (severity) to damage normal tissues in the host. Tissues then provide a medium for the organism to reproduce. If the pathogen leaves the host, through routes such as a respiratory or intestinal tract, it can then be transported and spread to a new host or other reservoirs including contaminated food and waste from human or animals. For an individual to become infected, the organism might enter the host through some portal of entry such as eating contaminated food or inhaling airborne particles. Skin contamination though hand shaking or touching common surfaces is another means of transmission as the new host can then acquire the organism through an oral route. The pathogen can then attach to the next susceptible host (e.g. someone who does not have immunity or adequate resistance). The new host will then go through a series of physical changes and reactions in response to the body fighting the causative organism by producing antibodies [8].

Medical treatment of infectious diseases is largely dependent on the causative organism. Bacterial infections are commonly treated with antibiotics to kill the bacteria, or to prevent them multiplying (bactericidal or bacteriostatic), whereas treatment of viral infections mainly relieve the associated symptoms while a host's immune system fights the virus [9]. Infections vary in transmissibility. SARS-CoV-2 is the causative virus associated with coronavirus disease (COVID-19). COVID-19 is a respiratory tract infectious disease that results in coughing and sneezing which provide vehicles of transportation for the virus to new hosts, through small droplets of infection, however other modes of transmission include fomite (contaminated surfaces) and other bodily fluids do occur [10]. Some viruses, such as human immunodeficiency virus (HIV), are a blood-borne and cannot survive in air, water or saliva for any significant amount of time, making these viruses less transmittable, however can be contagious through other forms of human contact such as sexual transmission or needle sharing [11]. Bacterial agents such as Escherichia coli (E.coli) are primarily transmitted through consumption of contaminated food and are commonly spread from person to person from hand to mouth contact however do not survive in the air, on surfaces and as such are not spread by coughing [12].

Methods to control and prevent infectious diseases depends on a thorough understanding of the how disease spreads in communities, severity of illness, drivers for infections and effectiveness of clinical interventions [13, 14]. Epidemiology is the study of the distribution and determinants of disease in different groups of people. The knowledge gained by epidemiology informs the management of disease outbreaks for populations at risk [15]. Infectious disease epidemiology specifically relates to the complex relationship between hosts and infectious agents which is primarily concerned with minimising the impacts of pathogens on public health [16]. Infectious disease surveillance is an epidemiological tool to help monitor disease burden. The structure of infectious disease surveillance is largely driven by the epidemiology and clinical presentation of diseases,

however a central response is the analysis and reporting of surveillance data to aid fast and efficient public health actions [17]. Infectious disease surveillance can take different forms such as aggregated or case-based disease surveillance, with both forms providing useful information. Aggregated disease surveillance is a strong tool for monitoring the number of cases and overall burden of disease, however it may lack the detail of individual-level case based disease surveillance. For example, measles was classed as an endemic for many countries until 1990 and as such surveillance was historically of aggregated form as this was the most feasible for the primary goal of reducing mortality. However, by 2016 most WHO countries changed their goal to be disease elimination and therefore a case-based surveillance became more suitable [17]. This is a primary example of goal driven surveillance, which has been highlighted across the world during the COVID-19 pandemic. Countries such as New Zealand have largely focused on case-based surveillance throughout as the number of cases have remained low. This means that contract tracing and localised lockdown have been quickly implemented [18]. Although the primary goal in the UK has also been case-based surveillance, during the peaks of infection this has been much less feasible. Efforts were made through the use of the track and trace systems however estimate show that tracers in England fail to contact approximately 1 in 8 people who test positive, with 18% of those reached providing no details for close contacts [19].

Infectious diseases continue to pose major global health threats. With the recent devastation from the COVID-19 pandemic and growing concerns that novel resistant bacterial infections could emerge as healthcare associated infections (HAI), there is a worldwide requirement for stringent techniques to monitor and understand the key drivers of the spread, incidence and prevalence of these diseases. The European Centre of Disease Prevention and Control (ECDC) have estimated that approximately 4 million patients a year suffer from a HAI in Europe, with roughly 37,000 deaths that are a direct consequence of these infections and many more deaths thought to be related [20]. At the

time of writing, the worldwide total number of COVID-19 cases is estimated to be 219 million and an associated 4.54 million COVID-19 related deaths [21].

The main aim of the work discussed throughout this thesis is to highlight the capabilities of routinely collected and open sourced data in the understanding of population-based risk factors for community spread and healthcare associated infectious diseases. Identifying ecological risk factors aids public health surveillance and control of infections. This thesis focuses on the use of spatial and spatio-temporal methods for analysing infectious disease data.

In this introduction, background information on community spread and healthcare associated infectious diseases, specifically *clostridioides-difficile* infection (CDI); the important role of antibiotic prescribing and COVID-19 infection are discussed in section 1.1. Section 1.2 summarises the use of spatial methods in the application of infectious diseases and section 1.3 highlights benefits associated with analysing routinely collected healthcare data with section 1.4 concluding this chapter with an outline of this thesis.

1.1 Community Spread and Healthcare Associated Infectious Diseases

HAI's are infections that people acquire from an interaction with a healthcare setting, for an unrelated reason [22], with bacterial infections being the most common type of HAI's [23]. These types of infections can occur from direct intervention such as surgical or medical treatment, or as a result of transmission in a healthcare setting. HAI's are a risk for everyone involved in healthcare settings including healthcare professionals, patients and visitors and are therefore is the responsibility of everyone involved [23]. Interventions within healthcare settings, such as improving hand hygiene, are known to reduce the presence of HAI's although compliance is a common problem [24]. HAI's are

estimated to cost the NHS 1 billion pounds each year, with the additional implications of increased use of NHS resources, patient morbidity and a reduction in patient safety. Many HAIs are deemed to be preventable and there are national and global initiatives to reduce these avoidable illnesses [25, 26].

The most common types of HAI's are surgical site infections (SSIs), urinary tract infections (UTIs) and blood infections, caused by a variety of bacteria. Some of the most well known include *Methicillin-resistant Staphylococcus aureus* (MRSA), *Clostridioides difficile (C-difficile)* and *Escherichia coli (E.Coli)* [27]. MRSA is a bacterium resistant to many antibiotics including methicillin, amoxicillin and penicillin [28]. MRSA is commonly referred to as one of the 'superbugs', which are a class of bacteria that have mutated to protect themselves from most antibiotic classes, making treatment difficult [29]. However, transmission of COVID-19 in hospitals has been a burden throughout the pandemic, despite efforts to contain infections and minimise spread, estimates show the hospital-acquired infection rate for SARS-CoV-2 to be approximately 12–15% of all positive COVID-19 cases in hospital [30]. In March 2020, a freedom of information request provided by 81 acute hospital trusts stated that a total of 32,307 patients, who were admitted to hospital with other conditions, contracted COVID-19 in hospital and 8747 (27%) of these patients died within 28 days. This is notwithstanding that this was during the peak of the first wave of the pandemic, when personal protective equipment for staff was in short supply and there were rising infection rates in the community, however does highlight the anxiety surrounding the spread of COVID-19 in hospital settings [31].

Healthcare associated infection's (or healthcare-acquired infections) is the umbrella term used for infections that occur as a result of contact with the healthcare system, however these infections can be defined separately by hospital acquired (HA) and community acquired (CA) infections. These are broadly defined: hospital (nosocomially)

acquired infections are infections that were not present or incubating in an individual at the time of admission to a hospital [32], whereas a community acquired (CA) infection are infections that were contracted outside of a hospital or diagnosed within 48 hours of admission without any previous health care encounter [33].

This thesis primarily focuses on *Clostridioides-difficile* Infection (CDI), which is globally considered to be major burden in community and hospital settings [34, 35, 36]. Health Protection Scotland (HPS) report the numbers and trends of CDI using two categories: Healthcare-associated CDI (HA-CDI) is a CDI patient with onset of symptoms at least 48 hours following admission to a hospital or up to twelve weeks after discharge from a hospital and a Community-associated CDI (CA-CDI) is a CDI patient with onset of symptoms while outside a hospital and without discharge from a hospital within the previous 12 weeks – or with onset of symptoms within 48 hours following admission to a hospital without stay in a hospital within the previous 12 weeks [37]. Defining infections separately can provide more detailed information of the ecology of the infection.

1.1.1 *Clostridioides-difficile* Infection

Clostridioides difficile (C. difficile) is a bacterium that colonises the bowel in approximately 5% of adults [38]. In those that develop symptomatic *C. difficile* infection (CDI), diarrhoea, fever, and abdominal pain are common. The majority of reported CDI cases are related to a hospital or care-home stay, however recent studies indicate that the incidence of community-associated *C. difficile* infection (CA-CDI) is increasing and may account for up to 30% of all CDI cases [39, 40]. CA-CDI population-based studies report similar risk-factors to hospital associated *C. difficile* (HA-CDI) however, there have been key differences reported with CA-CDI cases linked to younger patients and less severe illness [40].

Mandatory surveillance of CDI was introduced in Scotland in 2006 due to the rapid increase in cases at the beginning of the century in the United Kingdom (UK), Europe and North America. Initially, surveillance was only enforced for patients aged 65 years and older, however was extended to include all patients aged 15 years old or older in April 2009. There was high incidence and several serve hospital outbreaks around this time, primarily in older patients [41], with the Vale of Leven hospital (VOLH) outbreak in Strathclyde being one of the most notable in Scotland. Between 1st January 2007 and 1st June 2008 there were 131 patients who tested positive for CDI in this hospital, and of these patients 68 had tested positive between 1st December 2007 and 1st June 2008 (6-months). During this 6 month period, 28 of those 63 patients died with CDI as a causal factor of death. Over a 2 year period (1st January 2007 to 31st December 2008) there was a total of 34 CDI related deaths at the VOLH, which is even more impactful given that the VOLH is a small hospital with only 136 beds. The inquiry into this outbreak was a key driver for the Scottish Government to shift focus onto *C-difficile* infection, with actions implemented such as the stewardship of associated high-risk antibiotic groups and strict surveillance of infection rates [42]. Since 2008 there has been a dramatic and sustained decrease in the incidence of CDI, although the disease remains a reportable illness, as vulnerable patients continue to be at risk. [41].

Risk factors of CDI include: antimicrobial exposure, age, proton pump inhibitor (PPI) use, a previous stay in hospital or nursing home and comorbidities such as hypertension, diabetes and Chronic obstructive pulmonary disease (COPD) [39, 43, 44]. The epidemiology of *C. difficile* is likely to be shaped by a range of factors [45], however evidence from a broad range of studies, including those using whole genome sequencing, supports the central role of antibiotic use as the most important cause of symptomatic CDI [46]. Whole genome sequencing has shown considerable genetic diversity in CDI cases, with many hospital infections having no evidence of transmission from another symptomatic CDI patient which implies that the infection may have arisen as a result of

activation of otherwise quiescent organisms for a person who was previously colonised asymptomatically. This finding has important implications for antibiotic stewardship and for infection prevention both outside and within the hospital setting [47]. Research carried out in North Wales in 2015 found three quarters of cases sampled from both hospital and community settings were unrelated to any other identified cases [48].

CDI bacteria often lives harmlessly in the gut, as other bacteria keep it under control in the bowel, however symptomatic CDI can often cause a person to experience diarrhoea which therefore allows the bacteria to spread, as the vehicle of transmission for *c-difficile* infections is faecal matter. Once the bacteria have left the body (or host), it turns into resistant cells called spores which can survive on hands, surfaces (such as toilets) and clothing for long periods of time unless these are thoroughly cleaned. Hand to month contact can then lead to the spread of infection to others. Studies of molecular typing and contact tracing have estimated between 10% - 38% of hospital-onset CDI (occur $\geq$ 48 hours after admission) can be attributed to transmission from known symptomatic sources within the hospital. However, these estimates suggest that a large proportion of CDI cases originate from other sources, such as community exposure or transmission from asymptomatic patients. A transmission model of CDI found that hospitalised patient with CDI transmit the infections at a rate 15 (95% CI 7.2 - 3.2) times that of an asymptomatic patient, although persons in the community transmit infection at a rate of 0.1% (95% CI 0.002 - 0.2 %) that of hospitalised patients [49].

Evidence surrounding the acquisition and transmission of CDI is largely focused in hospital settings, with risk factors including patient to patient transmission; room assignment; tube-fed patients and room square-footage [50, 51, 52]. Spatial modelling of community-acquired CDI (CA-CDI) has shown an association with environmental

exposures such as proximity to livestock farming, farming raw materials and nursing homes [53, 54].

Environmental factors may be a common cause of *C. difficile* exposure and asymptomatic colonisation with transition to symptomatic CDI being mediated by antibiotic use. A recent environmental survey from the USA identified a high prevalence of toxigenic *C-difficile* from community environments that were similar to clinical isolates, including 24.6% of swabs from recreational parks showing toxigenic *C-difficile* [55]. However, another survey of *C. difficile* amongst healthy volunteers in the community from UK reported a low prevalence of 0.5% [56]. Notwithstanding the possibility of geographical difference in *C. difficile* ecology between the UK and USA, these data suggest that the environmental exposure of individuals may be common, although the colonisation of *C-difficile* from environmental exposure may be minimised by protective factors such as a healthy gut flora or hygiene practices in the community. Individual-level risk factors of CDI are well understood [57], and these studies highlight potential environmental transmission links in the community, however there is a paucity of information on risk factors of CDI at an ecological level.

1.1.2 The Role of Antimicrobial Prescribing

The associations between the use of broad spectrum antibiotic and CDI has been consistently reported [58, 59, 60]. Broad spectrum antibiotics are likely to disrupt the existing microbial ecology of the gut leading to an overgrowth of pre-existing, previously asymptomatic *C. difficile* or newly acquired organisms. Antibiotic stewardship is a core component of national responses to CDI. In primary care, this may be interpreted as encouraging low rates of GP antibiotic prescribing by raising the clinical threshold for initiation of antibiotics or by delaying prescription. Recent research provides strong evidence that the declines in CDI observed in England were associated with changes in antibiotic use, particularly fluoroquinolones [46]. A meta-analysis compris-

ing of 32 studies found antibiotic stewardship programs to have reduced the incidence of *c-difficile* by 32% ($p = 0.0029$) alongside a 37% ($p = 0.0065$) reduction in the incidence of MRSA [61]. However, a study in a hospital in Scotland between 2006 and 2010 found a significant reduction in the use of high-risk antimicrobials (HRA) although of the six comparable studies, only 2 showed a decrease in CDI rates, concluding that despite the large reduction in prescribing it is challenging to demonstrate the real-world impact [62]. Nevertheless, ongoing stewardship of particular broad spectrum antimicrobials that are associated with a high risk of CDI is recommended to reduce the number of patients predisposed to CDI and lower transmission rates [63]. The four broad-spectrum antibiotics targeted by stewardship are collectively called the '4C antimicrobials' which includes Cephalosporins, Clindamycin, Co-amoxiclav and Ciprofloxacin [64]. Ciprofloxacin belongs to a group of antibiotics called Fluoroquinolones and stewardship targets are commonly directed towards the entire group of antibiotics.

There is an overall goal to raise awareness of unnecessary and inappropriate use of all antibiotics in Scotland, which aims to prolong the efficacy of currently available antibiotics, minimise the rise in antimicrobial resistant (AMR) organisms and to control infections such as CDI [65]. Monitoring the amount of primary and secondary care antibiotic prescribing is imperative in the control of HAI's and AMR. In Scotland in 2019, Primary Care antibiotic prescribing accounted for 83% of all antibiotics prescribed (daily defined dose, DDD), whereas Secondary Care (acute and non-acute combined) accounted for 17% of total antibiotic prescribing. GP prescribing accounted for 71.5% of all primary care prescribing, with 7.9% prescribed by Nurses and 2.9% prescribed by Dentists [66]. Consequently, antimicrobial use is a community issue rather than just a health care setting issue.

Over-prescribing of antibiotics is a particular problem in primary care, with respiratory tract infections (RTI's) (commonly a self-limiting illness) being the leading cause for prescribing [67] and accounting for 60% of all primary care antibiotic prescriptions. RTI's are most commonly viral infections and therefore prescribing antibiotics is often futile which provides an opportunity to reduce antibiotic prescribing [68]. During the first-wave of the COVID-19 pandemic, Scotland saw a 44% increase in primary care prescribing of antibiotics commonly prescribed for RTI's, when compared to the same week in 2019. However, after this initial surge, prescriptions were less than previous years rates [69]. NHS England reported a 15.5% decrease in overall GP antibiotic prescribing, however when these figures were compared to the absolute number of appointments, the number of prescriptions were actually 6.7% higher than expected ($p < 0.0001$) [70].

Antibiotic prescribing in the community is clearly sensitive to changes, however the reduction of primary antibiotics is a multifaceted issue. A survey of 1000 general practitioners (GPs) in England found that 55% felt pressured to prescribe antibiotics, mainly from patients, regardless of whether they knew if the prescription was necessary and 44% admitted to prescribing an antibiotic just to get a patient to leave the surgery. A third of these GP's admitted to prescribing antibiotic several times a week, primarily due to not knowing whether the infection was viral or bacterial and others claiming a lack of diagnostic tools [71]. The Scottish Antimicrobial Prescribing Group (SAPG) have been working with NHS Scotland health boards to improve antibiotic use across Scotland since 2008, working closely with NHS Education for Scotland (NES) to create material to engage with healthcare staff, patients and the general public to raise antibiotic awareness. Campaigns in Scotland include adverts in pharmacies and social media platforms as well as broadcasts on local radio station [72].

1.1.3 COVID-19

On the 30th of January 2020 the Director-General of the World Health Organisation (WHO) declared the novel coronavirus outbreak a public health emergency of international concern (PHEIC), WHO's highest level of alarm [73]. A novel coronavirus was identified in samples on the 12th of January 2020 and was referred to as SARS-CoV-2, with COVID-19 as the associated disease [74]. COVID-19 was initially classed as a high consequence infectious disease (HCID) in January 2020 with examples of HCID's including Ebola Virus and Middle East respiratory syndrome (MERS), however as early as March 2020 the virus was no longer considered a HCID in the UK. As the WHO continued to class COVID-19 as a PHEIC, stringent measures remained in place to control the spread of infection.

The first positive COVID-19 case was reported in Scotland on 1st March 2020 [75]. In the early stages of the pandemic the key symptoms of COVID-19 infection were fever, a new and continuous cough and shortness of breath [74]. As time passed, other common symptoms highlighted included anosmia (loss of smell), ageusia (loss of taste), loss of appetite and fatigue. Other non-specific symptoms include myalgia (muscle pain), sore throat, headache, nasal congestion, diarrhoea, vomiting and nausea. Symptoms of COVID-19 may appear 2-14 days after exposure to the virus and the risks of COVID-19 are increased for those aged 60+, male sex and those who have health conditions such as lung or respiratory illness, diabetes or conditions that effect the immune system [76]. Sociodemographic factors including deprivation, ethnicity, population density and obesity were also recognised risk factors of COVID-19 [77, 78]. Although these factors define those who are at a higher risk of COVID-19 and critical illness, everyone is susceptible to the infection. There is currently no clear understanding of the drivers for symptomatic and asymptomatic people in general public. A meta-analysis of 13 studies from seven countries found that the percentage of asymptomatic cases was 17%

(95% CI 16% - 20%), with the relative risk of transmission 42% lower for asymptomatic people compared to symptomatic people [79].

As previously mentioned, COVID-19 is a respiratory tract infectious disease therefore the disease can be spread from coughing, sneezing and singing, through small droplets or aerosols containing the virus [10], however can also be passed purely by standing too close to someone and talking. The implementation of strict social distancing rules and mask wearing have been enforced as source control throughout the pandemic. Multi-layer cloth masks can prevent the exhalation of respiratory particles alongside the microorganisms they carry, while also protecting against 50-70% of the inhalation of small participles [80]. During the first wave of the pandemic, these rules were strictly enforced as an attempt to slow the spread of infection as quickly as possible and protect the NHS from becoming overrun with cases.

Surveillance of COVID-19 during the first wave of the pandemic was assessed through use of positive testing rates, mortality and telehealth data (such as NHS 24 and 111). The gold standard test for COVID-19 is the RT-qPCR which is an antigen test that provides fast results. However, during the first wave of the pandemic there was very little access to testing which was narrowly focused and very targeted, prioritising key workers such as NHS staff in the very early stages, however, anyone who presented with one of the COVID-19 key symptoms could book a test on the Scottish Government website. Those who did not have symptoms but sought a test were advised to ask employers or their place of study to provide a test, or order home tests. Testing capacities were low at the beginning of the pandemic, with Scotland averaging 1900 tests per day, however, this grew quickly from 1700 per day noted on the 1st of April 2020 to 4400 per day by the 30th of April 2020 [81]. Surveillance therefore had to rely

on a combination of sources to understand the burden of COVID-19 during these early stages.

As time passed, serology testing (testing for antibodies) became more widely implemented as a surveillance tool for COVID-19, with an updated Public Health Scotland 2021 surveillance report suggesting the seroprevalence (antibody prevalence) between 66.6% and 69.7% among those attending community healthcare settings. Positive testing rates were closely monitored across the country, with telehealth services remaining a core component for route to care for those experiencing symptoms of COVID-19 [82].

1.2 Spatial Analysis of Infectious Diseases

Infectious diseases are inherently spatial due to their contagious nature. Infectious disease data are commonly described in terms of person (or population) incidence, with an associated place and time therefore lend themselves to spatial and spatio-temporal analysis methods.

This thesis aimed to explore ecological risk factors of *C-difficile* infection, particularly focusing on the relationship between CDI and GP prescribing of high-risk antibiotics using non-identifiable aggregate data over time. An additional aim was to investigate the relationship between high areas of COVID-19 infection and two COVID-19 symptom reporting platforms during the first wave of the pandemic to assess surveillance capacities. All of the data obtained for this thesis were therefore of a spatio-temporal structure and hence motivated the use of spatial and spatio-temporal methods.

The control of infectious diseases is heavily reliant on the understanding of key drivers of infection, and their association with demographic, behavioural, socioeconomic and environmental factors. Tobler's Law, The First Law of Geography, states:

> "Everything is related to everything else, but near things are more related than distant things" [83].

This is the fundamental concept behind all spatial analysis. The goal of spatial analysis is to try to understand and account for, any underlying spatial structure so that real-world relationships can be seen. Modelling data in a spatial framework allows the possibility to further understand the characteristics of places and the relationships between them. Spatial mapping of disease is a historically applied technique, with one of the most famous epidemiological studies conducted by Dr John Snow during the cholera outbreak in London in 1854. Dr John Snow collected and mapped data by street addresses where there were high incidences of cholera deaths, which eventually led to the identification of the epidemic source - a contaminated water pump [84].

Disease mapping is an important and useful methods for the management of disease as the simple visual nature helps to raise awareness of issues quickly. For example, visualising COVID-19 positive testing rates by local authorities across the UK, shows that Scotland and Northern Ireland are currently experiencing higher rates of COVID-19 in comparison to other areas in the England and Wales. Areas such as Glasgow and Lanarkshire are experiencing the highest rates in the UK, while some of the Scottish islands appear to be much lower, with Orkney showing the lowest rates of COVID-19 across the whole of the UK (figure 1.1).

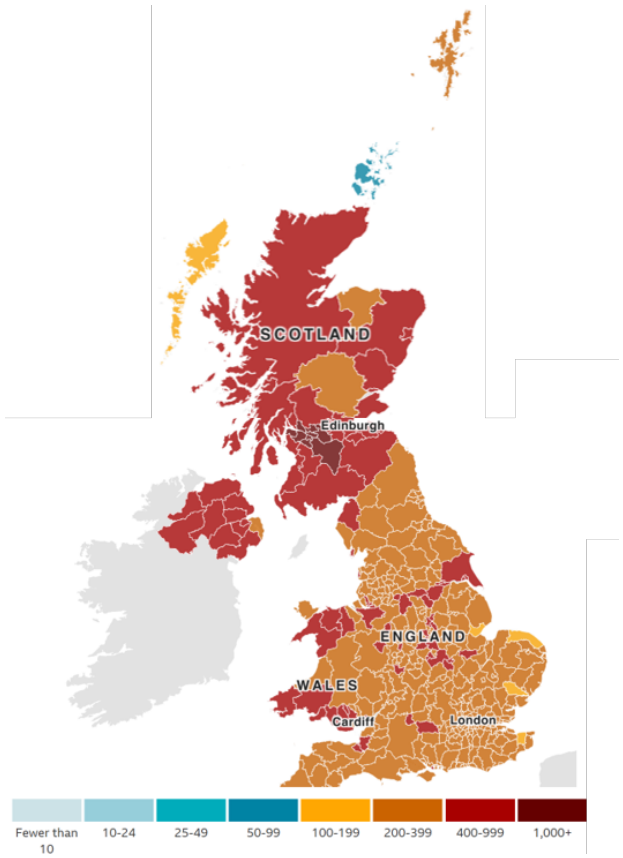


Figure 1.1: Rate of COVID-19 (cases per 100,000 per week) by United Kingdom Local Authorities for September 2021 [4].

Mapping disease prevalence and incidence has been used in public health for a long time, however the statistical challenge of obtaining reliable estimates of risk remains. The goal of disease mapping is to obtain small area estimation at fine scale geographic resolution [85].

The spatial granularity at which spatial data are explored can range from small area units, such as electoral wards, to global trends in the spread of infection. Spatial analysis techniques also have strong predictability capacities and provide the opportunity for real-time infectious disease surveillance. The application of spatial techniques to contemporary data sources such as social media, in combination with environmental and epidemiological data can assist in the real-time updating of spatial maps [86]. Rapid

decision making about containment, management and prevention is the primary goal with all infectious disease.

The analyses of infectious disease with spatio-temporal techniques has dramatically increased in popularity in the last two decades and with modern day computational power combined with spatio-temporal modelling advances alongside a diverse ranges of epidemiological data, it has been said that society is on the cusp of the development of a fully integrated approach for epidemic and intervention early detection [87].

1.3 Use of Open-Sourced and Routinely Collected Data

Electronic Health Record (EHR) refers to a digital version of patient medical records that allow instant and secure access to a patients medical history from remote locations [88]. The use of EHRs has allowed the development of novel approaches for large-scale epidemiological studies and public health interventions, benefiting a wide range of clinical areas. In Scotland, all health-related activities, from prescribing to test data to surgical procedures and dental appointments, are recorded electronically. The richness of linking these data provide opportunities to understand patterns in health and spread of disease, while ensuring data remain anonymised and protected as a core focus [89]. In Scotland, personal health data is protected and anonymised by the Community Health Index (CHI), which is a population register that is used to uniquely identify a person [90]. The envisioned future of data-linkage and electronic health data for Scotland is to set a international standard for safe and secure use of EHR and other routinely collect population-based data for research purposes [89].

Routinely collected data (RCD) are health data that are collected for purposes other than research or without a defined research question, prior to collection. RCD are increasingly used in medical research as these data are easy to access, collected under real-world circumstances and are inexpensive. There are vast amounts available with extensive data-linkage opportunities such as linking with genetic data. However, RCD presents a number of challenges despite efforts focused on improving the quality of data collection, storage and linkage, as well as other technical challenges including information security, confidentiality and methods of working with big data [91, 92].

Open sourced healthcare data are freely accessible data that are available for modification and sharing. In Scotland, the information service division (ISD) provide many open-sourced data sets, including Scottish healthcare and social care data, that are free to download, supported by Public Health Scotland and the NHS [93, 94]. Other forms of open source data include census data, which are data that provide a picture in time of the national population to aid the understanding of population demographics. These records support other open source data-bases, such as the Scottish Index of Multiple Deprivation (SIMD). The SIMD is the Scottish Government's statistical tool to help identity areas of multiple deprivation by measuring health, employment, income, skill and training, housing and crime all of which are represented by data zones (DZ = 6976).

The data-linkage these data provide endless opportunity to conduct large scale health studies, with the goal to aid public health and implement beneficial interventions. Randomised control trials (RCTs) are likely to remain the gold-standard for establishing the strength of associations for health related research, however there is untapped potential in these aggregated data structures, leading to cost-effective ways of defining population-based risk factors of infection [92].

1.4 Thesis Outline

This thesis utilises a number of different statistical techniques to analyse routinely-collected, and open-sourced, data with an aim to understand population-based risk factors of clostridioes-difficle infection and the surveillance potential of symptom reporting of COVID-19 infection during the first wave of the pandemic. The methodological focus of this thesis is applying spatial and spatio-temporal techniques in the analysis of infectious disease data, while additionally handling a mutli-level spatial data problem.

An overview of the existing methodology, that will be applied throughout this thesis, is presented in Chapter 2. In particular reference to correlation and modelling techniques for spatial and spatio-temporal data structures, and reviewing the relevant literature.

Chapter 3 presents an analysis of routinely collected EHR data, which aimed to understand the spatial distribution of general practitioner (GP) antibiotic prescribing by individual practices in Scotland. GP practice population and demographic data were obtained and linked together for three consecutive years: 2016 to 2018. These data were then combined with antibiotic prescribing records and aggregated by GP practice. An analysis was performed to investigate risk factors of increased total antibiotic prescribing rates and to understand health-board difference between prescribing of specific antibiotic classes. A mixture of point-location and areal-unit spatial data were obtained for these analyses.

Following this, routinely-collected aggregated quarterly counts of CDI were obtained at intermediate zone (IZ) level, between 2014 and 2018, to explore spatial and temporal patterns of CDI in Chapter 4. IZ population and sociodemographic data were collected to assess risk factors of CDI. To adjust for age and sex differences within IZs expected CDI counts were calculated using indirect standardisation. Total CDI cases were ob-

tained by IZ, but were also analysed separately for community-acquired CDI (CA-CDI) and healthcare-acquired CDI (HA-CDI). These data were then modelled using spatial and spatio-temporal techniques to assess risk factors of CDI. A final spatio-temporal model was applied to explore clustered trends of CDI incidence in Scotland, to provide insight into whether CDI incidence was increasing, decreasing or remaining constant overtime and whether it was consistent for all areas in Scotland.

Chapter 5 then utilised the data and results discussed in Chapters 3 and 4. This study assessed ecological risk factors of CDI that are on the causal pathway but were not available at the same spatial scale (IZ). This presented a multi-level spatial problem which motivated this chapter to explore methods of spatial interpolation as a tool for transforming spatial point-level data to areal-unit data. Spatio-temporal modelling of these data then assessed these variables as risk factors of CDI, preforming a sensitivity analyses to compare model results.

Chapter 6 presents a retrospective ecological study that used routinely-collected aggregated disease surveillance data to understand the impact of total and high-risk GP antibiotic prescribing on total and stratified inpatient/non-inpatient CDI incidence in Wales. GP practice population demographic data were collected and adjusted for as confounders. Initially, a trend analyses was preformed before exploring GP antibiotic prescribing rates as an ecological risk factor for CDI incidence. Regression modelling of these data then assessed antibitoic prescribing as primary risk factors of total CDI and then stratified for CA-CDI and HA-CDI. This study provided between country comparisons for ecological risk factors of CDI in Scotland and Wales.

Chapter 7 then conducts a spatial and spatio-temporal analyses of COVID-19 positive testing data at postcode districts (PCD) level in Scotland. These analyses aimed to assess the correlation between test positivity against COVID-19 related NHS 24 calls

and predicted COVID-19 cases from the COVID Symptom Study (CSS) app users. The aim was to determine the strength of these data as surveillance tools in the initial months of a pandemic. The methods that had been previously applied throughout this thesis lent themselves to the exploration COVID-19, therefore, when the pandemic began it was thought appropriate to preform an analyses on COVID-19 data for this thesis.

Finally, Chapter 8 provides a summary of the key findings of this thesis and discusses possible future work.

Chapter 2

Methods of Assessing Correlation and Modelling Spatial Data

This chapter outlines the central statistical methodologies applied throughout this thesis, with particular emphasis on spatial data types and their associated methods, as referenced in the literature. Section 2.1 introduces the two spatial data structures used in this thesis, spatial point-location and areal-unit data. Section 2.2 then introduces commonly applied methods of assessing correlation and conducting spatial prediction for point-location spatial data. Section 2.3 then similarly describes methods of analysing spatial areal-unit data, and references the literature surrounding spatial and spatio-temporal modelling of count data in a Bayesian framework. The analyses conducted in this thesis uses R programming language [95], and will make reference to application in R throughout.

2.1 Types of Geospatial Data

Geospatial data is a phrase used to describe data that contains, or relates to, information given at a specified location. There are a few forms of geospatial data, including Vector, Raster and Attributes: this thesis focuses on the use of spatial, or spatio-temporal, data

in vector form. Spatial vector data provides a snapshot of information on the earth's surfaces which can then be graphically represented by 2-dimensional maps and used to conduct inference on the known values at specified locations or areal units. A temporal aspect can be introduced when these data are repeated at specified time-points for the same locations. There are three geospatial vector data types: point-location, line and areal. Real world examples of these data types include: house postcodes and field site locations (point); rivers and roads (line); villages and cites (areal) [96]. This thesis will only work with point-location and areal-unit data as all the data acquired have either been collected at postcode level or at predefined regions such as postcode district or intermediate zone.

2.1.1 Point-Location Data

Point-location, or geostatistical, data are data represented in vector, or list, form for information at a specific location with associated coordinates: Longitude and Latitude (or Easting and Northing). Longitude represents the X (Easting) coordinate and Latitude represents the Y (Northing) coordinate, where the units of Longitude/Latitude is degrees, minutes and seconds and Northing/Easting is metres.

To assess point-location data in R, a `SpatialPointsDataFrame` object must be created. This can be computed directly from a `data.frame` by specifying which columns contain coordinate information, using the function `coordinates`:

```
coordinates(DataFrame) = c("Longitude", "Latitude")
```

These R funtions are all contained within the R package **sp** [97]. The point-location data used in this thesis are primarily represented by postcode. To obtain coordinate information for each postcode, an online converter is used to geocode postcodes to longitude and latitude coordinates [98].

2.1.2 Areal-Unit Data

Areal-unit data represents information for a defined area or region. A simple example of this is Scottish health-boards: Scotland is divided by 14 regions to partition population health responsibility [99] (figure 2.1).

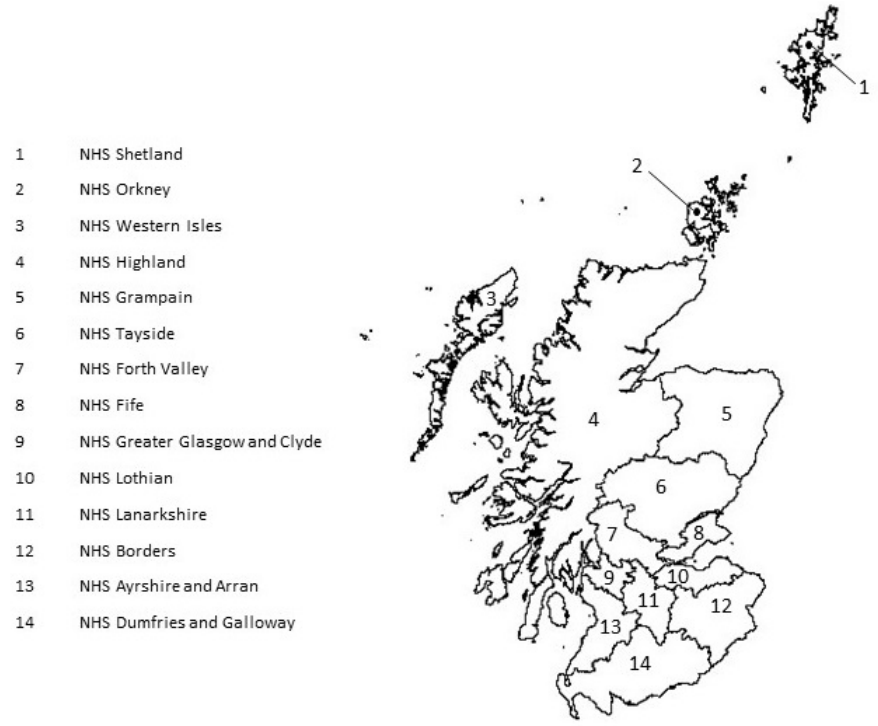


Figure 2.1: Polygons of spatial object: Scottish health boards. [5]

To analyse a spatial areal unit data frame in R, the data must be linked to areal boundary information and transformed to be a `SpatialPolygonsDataFrame` object. To create this object, the spatial data are linked to a shapefile. This thesis uses shapefiles from the Scottish government website [100], which include boundary information of the

required regions. The `combine.data.shapefile` function is then applied to combine spatial data and boundary information: `shp` represents the boundary information for each of the K areas that relate to the `data` and `dbf` is the lookup table that links the areas to the `data`.

```
SpatialDataFrame <- combine.data.shapefile(data=data, shp=shp, dbf=dbf)
```

This function is contained within the R package **CARBayes** [101]. This thesis will mention multiple areal representations of Scotland throughout, including: intermediate zones (IZ, n = 1279), data zones (DZ, n = 6976), postcode districts (PCD, n = 429), health-boards (HB, n = 14) and Agricultural Parishes (AP = 891). Data Zones divide Scotland into 6976 zones with population sizes ranging between 500 - 1000 household residents. Intermediate zones are an aggregated version of data zones, with 1279 zones for population sizes ranging between 2500 and 6000 household residents.

DZs are nested within IZ, both are nested within Scotland Local Authorities (LA, n = 32) and all nested within NHS Scotland health boards (HB, n = 14) [102]. There are 429 Postcode Districts (PCDs) that are nested within the 16 Scottish postcode areas which are represented by the first 4 digits of a postcode, e.g. "AB10" but PCDs are not nested within IZs as they span multiple IZs. This leads to difficulties when using multiple spatial data scales. These areal units vary greatly in spatial granularity, however, the choice of granularity is often determined by the requirements of the analyses or by accessibility to data and are not always compatible. The areal-unit data collected for this thesis are a mixture of population demographic, environmental and disease data, to represent people who reside within predefined regions. Data are available at a number of spatial scales, where some are easily linked using nested spatial scales such as DZ to IZ. Whereas others, such as environmental cattle density data, are only available at an Agricultural Parish (AP) level which is not always compatible with other spatial scales,

therefore, requiring more attention. The data obtained for this thesis are mostly count data which are aggregated to show a measure of disease incidence or rate over time, or by percentage of area-population, for each areal-unit.

2.2 Point-Location Methods

The typical goal of working with point-location data is to understand the spatial pattern between information at known locations and conduct spatial predictions of unknown information at known locations. This section presents methods for illustrating and testing spatial autocorrelation for spatial point-location data. Furthermore, this section introduces methods of spatial prediction. For the purposes of this thesis, there is a particular focus on utilising these spatial prediction methods to transform point-location data to spatial areal-unit data and handle incompatible spatial scales.

2.2.1 Variograms

In geostatistics the spatial correlation is modelled using the variogram. The variogram, $\gamma(h)$, measures the spatial dependence as a function of a separating distance h such that the correlation is measured between multiple pairs of points $(z(s_i), z(s_j))$ at a separating distance $h = s_i - s_j$, where, $s_1, ..., s_n$ denotes the locations of observed data, with $Z(s)$ denoting the value of the spatial variable at location s. The variogram is defined by,

$$\gamma(h) = \frac{1}{2}E[(Z(s_i) - Z(s_j))^2] \tag{2.1}$$

This translates as half the expected squared difference of the spatial variable at two locations specified by distance h. γ is known as the **semivariance** when evaluated at h [103].

An *intrinsic stationary* assumption is made which assumes a spatial variable $Z(s)$, at location s, is composed of a mean and residual with constant mean $E(Z(s)) = \mu$:

$$Z(s) = \mu + e(s) \tag{2.2}$$

In practice, the **sample variogram** will be calculated. $\gamma(h)$ is estimated from n observations of Z at locations $s_1, s_2.....s_n$ and denoted $\hat{\gamma}$.

$$\hat{\gamma}(h \pm \delta) = \frac{1}{2|N(h \pm \delta)|} \sum_{i,j \epsilon N(h \pm \delta))} [(Z(s_i) - Z(s_j))^2] \tag{2.3}$$

The average semivariance is calculated over an interval, $h \pm \delta = [h - \delta, h + \delta]$, given that very little (or no) points are separated by exact distance h. This process is called **binning**, similar to constructing histograms. Given,

- $N(h_{\pm\delta})$ represents the sets of pairs of coordinates s_i and s_j separated by h.
- $|N(h_{\pm\delta})|$ is the number of pairs of coordinates within each bin.
- The summation adds up the squared differences.
- $\hat{\gamma}(h)$ is the average squared difference between collected data separated by h.

When visualising the variogram, it tends to show higher variability at larger separating distances due to there being fewer pairs of points contributing to the semivariance in that bin [103].

Variogram Models

Variogram models formalise the sample variogram by allowing the spatial dependence to be estimated for any separating distance. The variogram models also help to understand specific features of the spatial dependence structure. There are four main variogram parameters (see figure 2.2):

- Range - This is the distance at which there is no spatial dependency (where the variogram "levels-off").
- Nugget - The short range variability (discontinuity at the origin caused by measurement error or small scale spatial variation).
- Sill - The variance between points that are not spatially dependent (value at which the data "level-off").
- Partial Sill - The variance of the data with the Nugget effect removed (value at which the data "level-off" subtract Nugget).

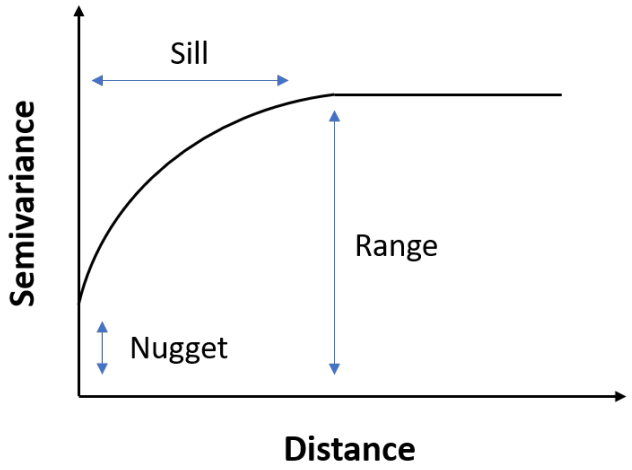


Figure 2.2: Variogram model parameters diagram [6].

There are multiple types of variogram models that can be fitted with some of the most common including: Nugget, Exponential, Spherical and Gaussian. Each of these models represent varying levels of spatial dependency e.g. Nugget model assumes no spatial

dependency, whereas, Exponential and Spherical assume high levels of spatial correlation between points at close distances which then decays with increasing distances: a similar structure to figure 2.2 where the semivariance increases with increasing binned distances of points [103].

Monte Carlo Envelopes: Test for Spatial Association

The variogram provides a visual representation of the spatial dependence within the data. A formal measure of spatial dependence can be achieved using Monte Carlo Envelopes, by permutation of the data on the spatial locations. Monte Carlo is a computational mathematical technique that relies on repeated random sampling to obtain numerical results that would otherwise be difficult to solve. The applications of Monte Carlo revolve mostly around three problems: optimization, numerical integration and generating samples from a probability distribution [104]. This instance adopts the latter, Monte Carlo is applied in conjunction with the flexible function `envelope` [103], to compute upper and lower limits for an estimated variogram to assess spatial correlation between pairs of points.

For each simulation, the data are randomly assigned to new spatial locations and a variogram is calculated applying the same spatial lags as the variogram that was originally calculated for the data. An $\alpha\%$ significance is then assumed and the $\alpha/2$ and $1-\alpha/2$ percentiles are computed for each bin of the variograms in the randomly simulated data. These lower and upper limits can then be superimposed to the original variogram and are known as the 'envelopes'. This is the Monte Carlo test for spatial association, in which the envelopes hold the assumptions of no spatial correlation under spatial randomness, hence if the variogram points fall outside of the limits this indicates evidence of spatial autocorrelation.

2.2.2 Interpolation

Spatial interpolation is the process of using known information from a set of points to predict a value at another location, or locations. There are many methods of spatial interpolation, with varying complexity, however, this chapter will focus on two methods: Inverse-Distance Weighted (IDW) interpolation and Kriging interpolation. Interpolation methods are divided by deterministic and statistical approaches, but share the same underlying concept: that new information is predicted dependent on the points close by. This relates to Toblers Law [83], mentioned in section 1.2.

Kriging interpolation relies on modelling the spatial structure of the data using a variogram function, however, in the absence of strong spatial association the variogram cannot always be adequately modelled. Inverse-distance weighted interpolation is a deterministic interpolation method that does not rely on a variogram model, and in some cases has been shown to out preform Kriging [105]. Most interpolation methods require some form of subjective process and it is important to quantitatively compare spatial predictions [106]. This thesis applies both a deterministic and statistical approaches of spatial interpolations. There is currently no rule of thumb for choosing an interpolation method as it varies greatly between specific situations. A recommended approach is to not assume one method for analyses but to try different methods where possible [107].

Inverse-Distance Weighted Interpolation

Inverse-distance weighted interpolation assumes that things close to each other are similar. This method of interpolation makes predictions at new locations based on a weighted average of the points surrounding it. The weighting of points is determined by the euclidean distance between observed locations $(s_1, ..., s_n)$ and new prediction location (s_0), raised by a power value, p. This power value determines the strength of

association such that a low p implies heavily smoothed data and high value of p implies more granularity. Predictions of a spatial variable $Z(s)$ at a new location s_0 is denoted by $\hat{Z}(s_0)$. The inverse-distance weighted (IDW) formula is described below:

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{n} w(s_0, s_i)^{-p} Z(s_i)}{\sum_{i=1}^{n} w(s_0, s_i)^{-p}} \tag{2.4}$$

Where $w(s_0, s_i)$ is the weighting factor determined by the euclidean distance,

$$w(s_0, s_i) = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \tag{2.5}$$

and $\hat{Z}(s_0)$ is the prediction of the response variable at new location s_0, dependent on the observed values of the variable $Z(s_i)$ at locations $s_1, ..., s_n$ [103].

The power value, $p \geq 0$, is a subjective decision and is the most influential part of this method. The chosen value of p will determine whether further away points are influential on the prediction or not (figure 2.3).

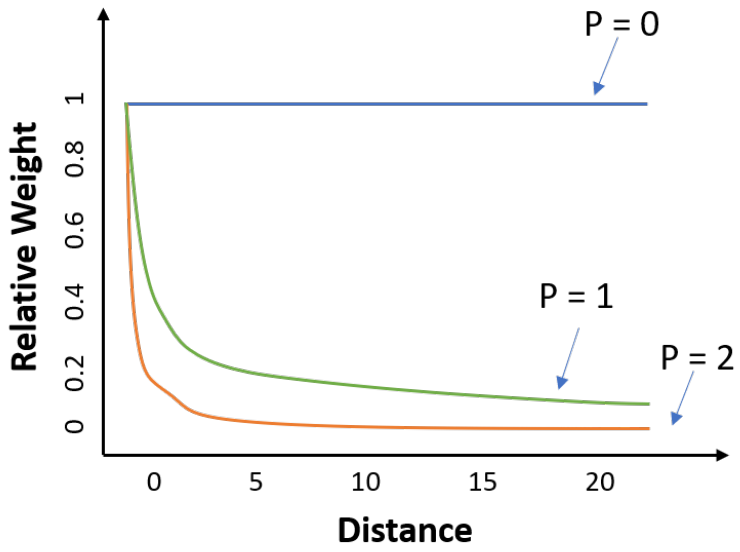


Figure 2.3: Inverse-distance weighted interpolation power diagram [7]

IDW is a simple and intuitive method that is quick to compute, however, it is a purely deterministic approach. It does not provide variance estimates or measures of uncertainty. It is also sensitive to outliers and the choice of parameters are intuitive [108].

Kriging Interpolation

Kriging is another method of spatial interpolation which is more statistically sophisticated than IDW. Kriging is an adaptation of IDW which incorporates a more formal measure of spatial dependence, allowing for uncertainty estimates to be produced.

IDW interpolation is determined by the distance between the prediction location and samples at observed location, whereas Kriging interpolation weights are determined by the sample variogram (equation 2.1), and are dependent on the spatial structure of the data [109]. The prediction equation for Kriging is similar to IDW in that the predicted value $\hat{Z}(s_0)$ at a new location s_0 is given by a linear combination of the observed values of the variable, $Z(s)$, at the observed locations, s_i:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{2.6}$$

The weights are represented by λ_i, and are under the constraint, $\sum_{i=1}^{n} \lambda_i = 1$, ensuring unbiased predictions. The Kriging weights are chosen so that the variance of the prediction error is minimised:

$$var[\hat{Z}(s_0) - Z(s_0)] = E[(\hat{Z}(s_0) - Z(s_0))^2] \tag{2.7}$$

By substituting equation 2.1, and rearranging, the prediction variance can be rewritten in terms of the sample variogram:

$$E[(\hat{Z}(s_0) - Z(s_0))^2] = 2\sum_{i=1} \lambda_i \gamma(h_{0i}) - \sum_{i=1}^{n}\sum_{i=j}^{n} \lambda_i \lambda_j \gamma(h_{ij}), \tag{2.8}$$

such that, $\gamma(h_{0i})$, is the variogram evaluated at the distance, h_{0i}, between new location, s_0, and observed location, s_i and $\gamma(h_{ij})$ is the variogram evaluated at the distance, h_{ij}, between observed points i and j.

The prediction error is then minimised by a set of weights:

$$\boldsymbol{\lambda} = \boldsymbol{A}^{-1}\boldsymbol{b}, \tag{2.9}$$

where $\boldsymbol{A}$ is the $n \times n$ covariance matrix for the observed data: $\boldsymbol{A}$ has diagonal entries representing the variance of $Z(s_i)$ for observed locations $s_i = s_1, ...s_n$ and all other entries represent the covariances between all points: $Cov(Z(s_i), Z(s_j))$ for $i = 1 : n$ and $j = 1 : n$. Vector $\boldsymbol{b}$ then represents the covariances between values at the new prediction location ($Z(s_0)$) and values of $Z(s_i)$ for observed locations $s_i = s_1, ...s_n$. The observed data covariances can then be manipulated so that they are estimated from the spatial data by defining the relationship between the variogram and covariances:

$$\gamma(h) = \sigma^2 - Cov(Z(s_i), Z(s_j)) \tag{2.10}$$

Rearranging this then allow the covariances to be represented in terms of the variogram:

$$Cov(Z(s_i), Z(s_j)) = \sigma^2 - \gamma(h), \tag{2.11}$$

where the diagonal of the covariance matrix $\boldsymbol{A}$ is now represented by constant variance (σ^2) across all location (estimated by the sill in the variogram), and all other entries are equal to $\sigma^2 - \gamma(h_{ij})$. Vector $\boldsymbol{b}$ then contains the covariances between observations at the prediction location and the sampled data ($\sigma^2 - \gamma(h_{0j})$).

The weights are now estimated in a form that will model the spatial dependency in the data. Using these weights in a prediction procedure is called Kriging, and produces Best Linear Unbiased Predictions (BLUP): it minimises the prediction error; every prediction is a linear combination of observations and are unbiased because the expected difference between $\hat{Z}(s_0)$ and $Z(s_0)$ is zero.

There are three forms of Kriging: Simple, Ordinary and Universal, although, Ordinary Kriging is the most commonly applied method [109]. Below shows the steps to preforming Universal kriging as Simple and Ordinary Kriging are simplified versions of the Universal process. Simple Kriging makes the assumptions that the spatial process has mean zero and is has a simple form that is quite restrictive. Ordinary Kriging assumes that the spatial surface has an unknown mean that needs to be estimated and Universal Kriging is similar to Ordinary Kriging, however, it can also account for covariate information, which is similar to a linear regression analyses at the observed locations. The steps of performing Universal Kriging is shown below. These steps are dependent on the variogram *intrinsic stationarity* assumption in equation 2.2:

Steps of Universal Kriging

$$Z(s) = \mu + \sum_{j=1}^{p} X_j \beta_j + \epsilon(s) \tag{2.12}$$

1. Estimate the parameters $\mu, \beta_1, ..., \beta_j$. For example, from linear regression model with spatial predictors.

2. Obtain the binned sample variogram for the spatial values $Z(s) - \mu - \sum_{j=1}^{p} X_j\beta_j$ (e.g. residuals) and fit an appropriate variogram model to obtain $\gamma(h)$

3. Use the fitted variogram, $\gamma(h)$, to construct the covariance $\boldsymbol{A}$ and $\boldsymbol{b}$

4. Calculate the kriging weights $\boldsymbol{\lambda} = \boldsymbol{A}^{-1}\boldsymbol{b}$ as described in equation 2.11

5. Calculate the predictions using $\hat{Z}(s_0) = \mu + \sum_{j=1}^{p} X_j\beta_j + [Z(s) - \mu - \sum_{j=1}^{p} X_j\beta_j]^T\boldsymbol{\lambda}$

Ordinary Kriging is then a special case of Universal Kriging when $\beta_j = 0$ in equation 2.12 and additionally setting $\mu = 0$ gives Simple Kriging.

Kriging interpolation is a sophisticated method of interpolation, that can quantify errors and there are multiple versions available to help configure data. It is more computationally intensive than other interpolation methods and relies on the sensitive input of the spatial correlation structure (fitting the variogram) [106].

2.2.3 Application of Point-location Methods in R

Chapter 3 explores GP antibiotic prescribing rates in Scotland, obtained by postcode. These data are transformed to a `SpatialPointsDataFame`, as described in section 2.2, by linking longitude and latitude coordinates. The spatial correlation amongst these data is then assessed by fitting a sample variogram using the function `variogram` from the **Gstat** R-package and simulating Monte Carlo envelopes using the `variog.mc.env` function from R-package **geoR**.

Chapter 5 then assesses the GP antibiotic data again, with an aim to convert these data from point-location postcode data to an intermediate zone (IZ) level. This is achieved by using IDW interpolation as a method of spatial prediction to IZ centroids, for multiple values of power p and comparing the RMSE using cross-validation. Centroids are extracted from each IZ to form a prediction grid which allow these data to be obtained

as IZ areal-level data. The antibiotic data did not display any spatial correlation and, therefore, Kriging interpolation is not an appropriate, nor possible, method for interpolating these data. IDW provides a method of computing spatial predictions at an areal level using interpolation, however, it is noted that IDW also assumes a level of spatial dependence. This is discussed further in section 5.2.3, where multiple values of power values p are compared.

Cattle density data are obtained for Chapter 5, by Agricultural Parish (AP, n = 891), however, these data are not compatible with IZs (n = 1279) as previously discussed in section 2.1.2. The cattle density data are converted from a `SpatialPolygonDataFrame` to a `SpatialPointsDataFrame` by extracting the cattle density data at AP centroids. Ordinary Kriging interpolation is initially applied to make spatial predictions of cattle density at IZ centroids. IDW interpolation is then also applied, for multiple power values p. The spatial predictions from both interpolation methods can then be compared by calculating the RMSE using cross-validation. These analyses are also implemented in R utilising the **Gstat** R-package, with functions `idw` and `krige`.

2.3 Areal-Unit Methods

Areal-unit data are a set of non-overlapping polygons that hold information regarding a specific area. Methods of areal-unit analysis also relate strongly to Toblers First Law of Geography, see section 2.4, however, instead of measuring the similarities between individual points in terms of distance, areal-unit analysis introduces the concept of Spatial Adjacency. Spatial Adjacency is a fundamental concept of areal-unit spatial analyses, which defines the spatial relationship between two or more polygons dependent on whether they share a boundary. Polygons that share a boundary are known as neighbours. This is expanded upon in section 2.3.1, before presenting a common method of measuring spatial autocorrelation for areal-unit data. Section 2.3.2 then

discusses methods of spatial modelling, commenting on the role of Bayesian inference and appropriate models for this thesis. Finally, methods for modelling spatio-temporal data are presented in section 2.3.3.

2.3.1 Spatial Adjacency

In spatial statistics, two areal units are considered to be adjacent (or neighbours) if they share a common edge/border (this is known as rook adjacency) or if they share a common vertex/corner (this is known as queen adjacency). This thesis adopts the former: areal units will be defined as neighbours if the spatial polygons share a common border.

An adjacency matrix or neighbourhood matrix ($\boldsymbol{W}$), is a summary of the relationships between pairs of polygons such that,

$$w_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and } i \text{ and } j \text{ share a common } \textbf{border} \\ 0 & \text{Otherwise} \end{cases} \tag{2.13}$$

The adjacency matrix is a $n \times n$ matrix, for n areal-units, that is symmetric and has zero diagonal, given that an areal-unit cannot be a neighbour of itself. The row and column sums of the matrix then represent the total number of neighbours for each spatial polygon. To conduct any formal analyses using the spatial adjacency matrix, all areal units to have as least one neighbour ($\sum_{i,j=1}^{n} \geq 1$).

This causes problems when working with Scottish data due to the number of islands. Those that are isolated, do not share a common border, will return zero neighbours. For this thesis, the closest areal unit was manually linked using euclidean distance e.g. working with Scottish health board data, NHS Shetland is assigned to be a neighbour with NHS Orkney, then NHS Orkney is linked to NHS Highlands. This ensures that

the Scottish islands are included in the spatial correlation structure. This becomes particularly important later for assessing spatial correlation and defining spatial random effects for modelling spatial data. The same approach has been adopted for other studies using Scottish areal-unit data [110].

The adjacency matrix, $\boldsymbol{W}$, has direct implications in spatial modelling given that if $\omega_{ij} = 1$ areal units (i, j) are modelled to be as spatially correlated. Similarly if $\omega_{ij} = 0$ then areal units (i, j) will be modelled as conditionally independent.

Spatial Adjacency Example

There are 14 NHS Scotland health boards, as seen in figure 2.1. NHS Lankarshire is a central Scottish health board, as shown in figure 2.4.

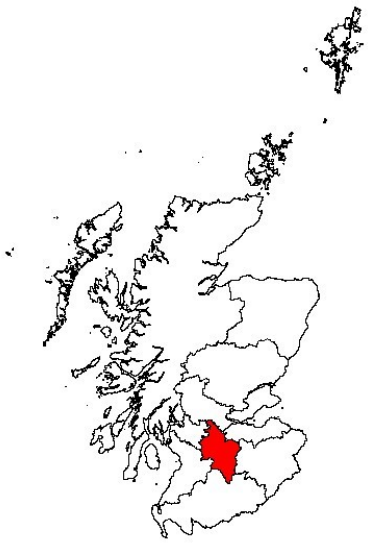


Figure 2.4: Spatial adjacency diagram 1: NHS Scotland health boards with NHS Lanarkshire highlighted in red.

Using the rook adjacency definition, areal units that share a common border, the neighbouring health boards can then be defined as shown in figure 2.5.

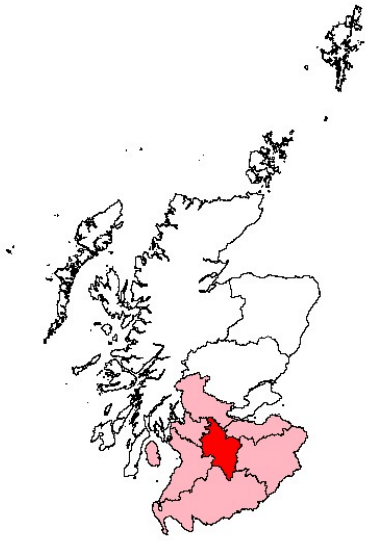


Figure 2.5: Spatial adjacency diagram 2: NHS Lanarkshire highlighted in red and neighbouring NHS Scotland health boards shaded in pink, as defined by rook adjacency.

The adjacency matrix, $\boldsymbol{W}$, is defined using the equation 2.13 such that $\omega_{ij} = 1$ if $i \neq j$ and share a common border:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
S08000015	0	0	1	0	0	0	1	0	1	0	0	0	0	0
S08000016	0	0	1	1	0	0	0	0	1	1	0	0	0	0
S08000017	1	1	0	0	0	0	0	0	1	0	0	0	0	0
S08000029	0	1	0	0	1	1	0	0	0	1	0	0	1	0
S08000019	0	0	0	1	0	0	1	1	1	0	0	0	1	0
S08000020	0	0	0	1	0	0	0	1	0	0	1	1	1	1
S08000021	1	0	0	0	1	0	0	0	1	0	0	0	0	0
S08000022	0	0	0	0	1	1	0	0	0	0	1	1	1	1
S08000023	1	1	1	0	1	0	1	0	0	1	0	0	0	0
S08000024	0	1	0	1	0	0	0	0	1	0	0	0	1	0
S08000025	0	0	0	0	0	1	0	1	0	0	0	1	0	1
S08000026	0	0	0	0	0	1	0	1	0	0	1	0	0	0
S08000030	0	0	0	1	1	1	0	1	0	1	0	0	0	0
S08000028	0	0	0	0	0	1	0	1	0	0	1	0	0	0

Figure 2.6: Adjacency matrix for NHS Scotland health boards.

The geocode for NHS Lanarkshire health board is "S08000023" and is also represented in column 9, see figure 2.6. Summing across the rows and columns of this adjacency matrix's shows NHS Lanarkshire to have 6 neighbours which is consistent with figure 2.5. The Isle of Arran is linked to NHS Aryshire and Arran as one health board.

Moran's I Test for Spatial Autocorrelation

The Moran's I statistic is a correlation coefficient which measures spatial autocorrelation of areal-unit data.

The Moran's I statistics is defined by:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.14}$$

where,

- n represents the number of areal units
- w_{ij} is the ij^{th} element of the adjacency matrix $\boldsymbol{W}$
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of the data $y_1, ..., y_n$

Moran's I values range between $-1 < I < 1$, where $I = -1$ represents perfect negative correlation, $I = 0$ represents spatial randomness and $I = 1$ represents perfect positive spatial correlation. Negative correlation values are unexpected as we expect things close to one another to be similar, although it is possible and is sometimes strongly informative.

Moran's I test for Spatial Association is then determined using a similar process to Monte Carlo Envelopes (section 2.1) such that Moran's I is calculated for K separate permutations of the data. A series of Moran's I values, simulated under spatial randomness, are then compared to the observed Moran's I in a permutation test. This is carried out under the Null Hypothesis of no spatial correlation and determined by a chosen significance level, $\alpha\%$ [111].

2.3.2 Spatial Modelling

Spatial modelling of areal-unit data aims to account for underlying spatial correlation structures and allow for more of the variability to be accounted for, in turn providing more stable estimates in comparison to linear modelling which assumes independence of spatial units.

A common practise when modelling spatial data is to initially use a generalised linear model (GLM) with independent errors, under the assumption that spatial covariates are sufficient to account for the spatial correlation. Residuals from these models can then be assessed using a spatial correlation test such as Moran's I. If the residuals show evidence of spatial correlation then the independence assumption is not valid and a spatial model is then required to account for this residual spatial correlation. This can be achieved by extending the GLM to include a set of spatially correlated random effects which produces a spatial generalised linear mixed model (GLMM) [112].

This spatial modelling section initially describes the generalised linear model (GLM), and the generalised additive model (GAM), as this thesis utilises them throughout as exploratory modelling methods, commonly as a prerequisite to performing spatial modelling. Spatial modelling requires more sophisticated estimation techniques and a Bayesian framework is commonly used, therefore, this section follows with an introduction of Bayesian statistics in the context of spatial modelling. The spatial generalised linear mixed model is then described with a discussion of the most popular conditional autoregressive (CAR) models. Defining the most suitable for the analyses in this thesis.

Generalised Linear Models

The GLM is a generalisation of ordinary linear regression that allows the response to have an error distribution other than normal. The GLM assumes the linear predictor is linked to mean μ through a *link function* $g(.)$ such that $g(\mu) = \eta$,

$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{2.15}$$

An *offset* term is commonly added to the linear predictor when modelling count data, if the number of events need to be adjusted for by some factor such as the population size of an area. For example, the expected annual number of *E.coli* cases μ_i in an area i is dependent on the given population P_i of the area, since areas with large populations would be expected to have more cases each year. The number of *E.coil* per unit of population, assuming a logarithmic link function, can be modelled: $\log(\mu/P) = \eta$, for the linear predictor η (equation 2.15) and then rearranged for μ such that $\log(\mu) = \log(P) + \eta$ [113].

Generalised Additive Models

Generalised additive models (GAMs) are a non-linear regression modelling technique that allows for complex relationships to be described. GAM's are an extension of GLMs such that the linear predictors are replaced with multiple smooth non-linear functions called splines. GAMs are commonly applied for modelling non-linear relationships and this thesis utilises their flexibility for exploratory purposes. GAMs are useful in the initial stages of exploration to visualise trends between independent and dependent variables. By allowing full flexibility of these models, GAMs help to expose the relationships between variables and assess highly influential data points. The function of the mean is presented as a linear combination of smooth functions of the explanatory variable(s) [114]:

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) \tag{2.16}$$

where equation 2.16 is seen to have a similar form to equation 2.15: g(.) represents the link function for a probability distribution from the exponential family E_Y. However, for GAM's the $\beta_j x_j$ has been exchanged for a flexible function $s_j(x_j)$. These flexible functions are known as splines. A spline is a piece-wise polynomial curve e.g. joins multiple polynomial curves together, where the location of the joins are known as 'knots'. In practice, the smoothness of a spline may be determined by directly controlling the number if knots (k) or by estimating the number based on predictive accuracy. The spline, $s(x)$, is defined below:

$$s(x) = \sum_{k=1}^{k} \beta_k b_k(x)$$

Here, β represents a weight and $b(x)$ represents a *basis expansion*. A *basis expansion*, or basis function, determines the flexibility of the curve being fit to the data [115]. This thesis adopts the use of the **mgcv** R-package for implementing GAM's, using the function `gam()` [116]. This function allows for models to be estimated using generalized cross validation (GCV), which aims to balance an optimal fit while maintaining interoperability. The estimated degree's of freedom (e.d.f) is summary statistic of the fitting GAM which provides feedback on the degree of non-linearity of the curve fit to the data, and tests whether this deviates from zero (i.e. no relationship). This thesis adopts default processes to visually assess the relationship between dependent and independent variables, however, also explored varying degrees of curvature by forcing high k, as this provided insight into the influence of outliers within the data.

Bayesian Statistics

Spatial modelling is commonly conducted using a Bayesian framework, where inference is based on Markov chain Monte Carlo (MCMC) simulations. This section briefly introduces Bayesian inference, the key components of setting up, and interpreting results from a Bayesian model.

Bayesian statistics presents a way of performing statistical analyses which allows prior knowledge or information to inform the analyses. It is derived from conditional probability with the fundamental concept of Bayesian statistics reliant on Bayes' Theorem [117]. $\boldsymbol{Y}$ represents the observed data and $\boldsymbol{\theta}$ represents any parameters that we wish to estimate (e.g. mean, variance, regression model coefficients, residual variance, etc.).(See equation 2.17).

$$f_1(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{f_2(\boldsymbol{Y}|\boldsymbol{\theta})f_3(\boldsymbol{\theta})}{f_4(\boldsymbol{Y})}, \tag{2.17}$$

$f_2(\boldsymbol{Y}|\boldsymbol{\theta})$ represents the likelihood of the data conditional on the distributional assumptions of θ. In frequentist statistics, inference is primarily made based on the likelihood. The likelihood depends on the type of data, e.g. if we wish to model count data then a Poisson distribution may be considered, whereas a linear regression model would apply a Normal distribution. Likelihood is, therefore, driven by the type of data acquired. $f_3(\boldsymbol{\theta})$ denotes the prior distribution which represents previous knowledge or beliefs about the parameters before seeing the data. $f_4(\boldsymbol{Y})$ is a normalising constant which is used to ensure the resulting posterior distribution integrates to 1. Finally, $f_1(\boldsymbol{\theta}|\boldsymbol{Y})$ gives the posterior distribution which combines information from both the likelihood and the prior distribution. The posterior distribution is used to make inferences regarding the parameters of interest.

The key difference between Bayesian and frequentist frameworks is defined by the way these methods treat model parameters. A frequentist approach would treat an unknown quantity as fixed, although unknown, whereas a Bayesian approach would treat an unknown quantity as a random variable whose randomness accounts for the uncertainty [118]. In practice, simulation techniques are applied to sample the posterior distribution for complex models to avoid mathematical challenges. For the purposes of this thesis, Markov Chain Monte Carlo (MCMC) simulations are used and all parameters are drawn from a *Gibbs* or *Metropolis-Hastings Sampler* or a mixture of both [101].

Monte Carlo is a collection of computational methods that help to solve difficult mathematical problems through the use of simulation. Markov chain Monte Carlo (MCMC) is a method that generates a series of sequential *Markovian samples* from a population where the probability of each new sample is dependent on the sample drawn in the previous process. MCMC provides a way of sampling from any probability distribution: in this instance the goal is to sample from the posterior distribution. In practice, direct sampling is often complex or not possible due to a lack of conjugate priors, therefore, algorithms are implemented too approximate from a given multivariate probability distribution.

Gibbs sampling is a commonly applied MCMC algorithm, where *Metropolis-Hastings* is a more generalised version of *Gibbs*. The *Gibbs* sampler relies on conditional distributions to construct a Markov chain: the probability of the next sample in the chain is calculated as the conditional probability given the previous sample. Therefore, *Gibbs* sampling is only appropriate for full conditional distributions which limits its applicability. For example, discrete models are simple to approximate using a *Gibbs* sampler whereas many continuous models have conditional distributions that do not have the parametric form suitable for this method of sampling. The *Metropolis-Hastings* sampling algorithm provides a solution to this problem. Unlike *Gibbs*, which assumes

a target distribution (conditional distribution) to select the next samples from, the *Metropolis-Hastings* sampling algorithm incorporates 'surrogate' or 'proposed' probability distributions alongside an acceptance criteria to decided whether a sample is accepted or rejected, allowing a more flexible method of sampling. In the scenario where the 'proposed distribution' is the conditional distribution, the *Metropolis-Hastings* algorithm is equivalent to *Gibbs* [101, 119]. Bayesian inference is commonly conducted by implementation of these computational methods. This style of Bayesian framework is well prepared to handle large complex models, which are common in spatial data and, therefore, is particularly suited for spatial analyses [118].

One of the biggest challenges in Bayesian modelling is knowing when an MCMC chain has reached convergence. One of the measures taken to achieve this is done by removing the 'burnin' period, whereby the first N observations in the MCMC chain are removed. The size of this burnin period is dependent on the time it takes for the chain to reach convergence towards an equilibrium, however removing the first 10% of samples is commonly recommended to start [118]. Another measure taken to improve convergence is to *thin* the number of samples to minimise between sample autocorrelation. This can be done by only saving every k^{th} sample and discarding the rest, however, the cost of thinning is paid in computational time. For example, to achieve 1000 samples, a thinning of 10 (save 1 in 10 samples) would require the MCMC to generate 10,000 values opposed to an unthinned chain of 1000 samples.

In practice, the number of samples required is dependent on the complexity of the underlying model e.g. a spatio-temporal model requires a higher number of samples and increased burnin to reach convergence compared to a spatial model due to the increased size and complexity. One method of assessing convergence is to examine trace plots of the samples from each MCMC run for individual parameters. These trace plots should show no evidence of a trend, showing that the sampler has moved between separate re-

gions of high probability, known as mixing [119]. Another tool for assessing convergence is the geweke diagnostics. This is a test that compares the mean at the beginning and the end of the Markov chain (commonly the first 10% and last 50%). The z-score is then calculated under the independence assumption and, therefore, the geweke diagnostic is a Z-test of the equality between two means [119]. The geweke diagnostic is adopted in this thesis a method of assessing convergence alongside checking trace plots.

Inference from a Bayesian model can be made by extracting key values from the simulated posterior distribution: the median is used as a measure of central location and 95% credible interval (evaluated by extracting the 2.5^{th} and 97.5^{th} percentiles of the distribution) are calculated for each parameter.

It is noted that a Bayesian framework with MCMC simulation using a *Gibbs* or *Metropolis-Hastings* sampler is not the only method, or the most efficient, for assessing spatial and spatio-temporal models. *Stan* is a programming language that allows Bayesian models to be written using statistical notation which can be used in R, Python and Matlab. *Stan* is an alternative sampler and optimiser platform which allows the implantation of Bayesian models using *Hamiltonian Monte Carlo*, which uses a multi-step process to eliminate the effects of autocorrelation between samples [120]. *INLA* is another Bayesian platform that stands for Integrated Nested Laplace Approximation which performs Bayesian inference for a wide class of hierarchical models. *INLA* avoids the use of MCMC simulation by taking advantage of numerical approximation methods which largely improves the computation time of running models and avoids convergence problems associated with MCMC [118]. Nevertheless, this thesis uses MCMC with a *Gibbs/ Metropolis Hasting* sampler throughout as the analyses was primarily conducted using the well-established R-packages **CARBayes** and **CARBayesST** for implementing spatial and spatio-temporal models, which adopts this Bayesian framework [112, 121].

Application of Bayesian Spatial Models in R

The **CARBayes** and **CARBayesST** R-packages are specifically designed to implement spatial and spatio-temporal generalised linear mixed models for areal unit data in R using a Bayesian framework with MCMC simulations [112, 121]. These packages are provide a straightforward interface for implementing these models that is easy to access and interpret, with strong supporting material [112, 121] and adaptable for multiple response distributions.

These packages model the spatial and spatio-temporal autocorrelation using random effects that are assigned a conditional autoregressive (CAR) style prior distributions, with a number of different random effect structures available. The output of these models are presented in a neat structure which is comparable to a `glm` summary output in R. The `print()` function of a model object will present a `summary.results` table which includes: the posterior median (`Median`) with `2.5%` and `97.5%` credible intervals, for each of the model parameters. The output also provides a measure of effective samples (`n.effective`) from the **coda** R-package and the `Geweke.diag`, as a measure of convergence for each parameter: a parameter is consider to have converged for values in the range $[-1.96, 1.96]$ [112].

Spatial Generalised Linear Mixed Models

A spatial GLMM model is defined by extending the GLM (equation 2.15), by including a set of spatially varying random effects for n areal-units, represented by $\phi_i = (\phi_1, ..., \phi_n)$:

$$g(\mu_i) = \boldsymbol{x_i}^T\boldsymbol{\beta} + \boldsymbol{O_i} + \phi_i \tag{2.18}$$

where $g(\mu_i) = \boldsymbol{x_i}^T\boldsymbol{\beta}$ represents the general form of a GLM with link function $g()$. This is the form of a spatial generalised linear mixed model. $\boldsymbol{O_i}$ represents a vector of known *offsets* for areal unit i, if required.

Conditional Auto-regressive (CAR) Models

In practice, a spatial GLMM can be implemented using a Bayesian hierarchical model with random effects represented by a conditional autoregressive (CAR) prior. This CAR prior incorporates spatial autocorrelation through the use of the adjacency matrix ($\boldsymbol{W}$) [112].

The simplest CAR prior is the Intrinsic Model, which was proposed by *Besag et al.* in 1991 [122]. The CAR prior is usually defined as a set of i univarite full conditional distributions of the form, $f(\phi_i|\boldsymbol{\phi_{-i}})$ for $i = 1, ..., n$, where $\boldsymbol{\phi_{-i}} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ...\phi_n)$, representing the random effects in equation 2.18. The Intrinsic CAR model is defined in equation 2.19:

$$\phi_i|\boldsymbol{\phi_{-i}}, \boldsymbol{W}, \tau^2 \sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}}\right) \tag{2.19}$$

If two areas (i, j) are neighbours, $\omega_{ij} = 1$ and the conditional expectation of area i is then represented by the mean of the data points of the neighbouring areas. This ensures that the mean of area i is similar to the surrounding areas. The conditional variance τ^2 is also dependent on the number of neighbours for i such that the variance decreases with increasing number of areas. This makes sense as the more neighbouring areas

there are for a region i, the more information is provided thus reducing uncertainty in the estimation of the mean for region i.

This model is a relatively restrictive model as it is a single parameter model that does not estimate the strength of the spatial correlation between the spatial random effects. This model is not suitable for data with a weak spatial structure and is best suited to strongly spatially correlated data [123].

There have been a number of extensions to the *Intrinsic* CAR model proposed, with the three most common suggested to be: Convolution CAR, Cressie CAR and Leroux CAR [123]. The Convolution CAR (*Besag et al. 1991* [122]) is an extension of the Intrinsic CAR which presents a linear combination of random effects with a CAR prior and a set of independent random effects: $\phi_i = \phi_i^{(1)} + \phi_i^{(2)}$, where $\phi_i^{(1)}$ is equivalent to equation 2.19 and $\phi_i^{(2)} \sim N(0, \tau_2^2)$. This improves the over smoothing problem with Intrinsic CAR but each data point is represented by two random effects and only the sum is identifiable. The Cressie CAR, or Proper CAR, (*Cressie et al. 2000* [124]) uses a single set of random effects and introduces a ρ parameter to allow the model to control the level of spatial correlation, again improving on the Intrinsic CAR's strong spatial correlation requirement. The Cressie CAR is presented in equation 2.20:

$$\phi_i | \boldsymbol{\phi_{-i}}, \boldsymbol{W}, \tau^2 \sim N\left(\frac{\rho \sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{ij}} \right) \tag{2.20}$$

where $\rho = 1$ represents strong spatial correlation, similar to Intrinsic CAR and $\rho = 0$ represents independence. The conditional variance parameter is equivalent to the Intrinsic CAR model and does not depend on ρ. This can cause problems as ρ trends closer to zero, as the conditional variance then becomes disproportionately dependent

on the number of neighbouring areas, whereas this would not be expected if areas were spatially independent.

Leroux CAR (*Leroux et al. 2000* [125]) is an alternative CAR prior which also allows the modelling of spatial autocorrelation to vary in strength, only using one set of random effects, while also ensuring that if $\rho = 0$ then τ^2 simplifies to a constant: if areas are spatially independent, the conditional variance is not modelled to be dependent on the number of neighbouring areas. The Leroux CAR model is defined in equation 2.21:

$$\phi_i|\boldsymbol{\phi_{-i}}, \boldsymbol{W}, \tau^2, \rho \sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\rho\sum_{j=1}^{n} w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + 1 - \rho}\right) \tag{2.21}$$

A study comparing these four models on performance, simulated disease data to assess each model in three separate scenarios: spatial independence, moderate spatial dependence and strong spatial dependence. The Intrinsic model performed the worst in the spatial independence scenario and the Convolution model preformed the worst in the third scenario, suggesting it is not appropriate for modelling data with strong spatial dependence. The concluding results showed the Leroux CAR model to consistently preform well in each of the scenarios, suggesting it can handle independence and strongly correlated data [123].

Leroux CAR is described in the literature as a flexible model that is suitable for modelling varying strength of spatial autocorrelation, which is appropriate for the data modelled in this thesis. This model is applied in Chapters 4, 5 to account for spatial autocorrelation between intermediate zone CDI incidence and in Chapter 7 which assesses the proportion of individuals who test positive for COVID-19 by postcode district.

Application of Spatial Models R

All spatial modelling in this thesis were implemented using the **CARBayes** R-package using the `S.CARleroux` function to conduct a Leroux CAR spatial model. This model can be fit for multiple distributions, which for this thesis includes Poisson and Binomial. The link functions for these are defined below in reference to the spatial GLMM in equation 2.18:

- **Poisson** - $Y_i \sim Poisson(\mu_i)$ with $g(.)$ link function $ln(\mu_i)$
- **Binomial** - $Y_i \sim Binomial(N_i, p_i)$ with $g(.)$ link function $ln(p_i/(1-p_i))$

The distribution family is specified in the *formula* sections of the `S.CARleroux` function. Other specifications include: the `data.frame` that contains the data of interest; the distributional family; the adjacency matrix ($\boldsymbol{W}$); the burnin period (if required); the number of samples and any thinning to be applied. Trials are specified specifically for modelling proportional *Binomial* data. For example:

```
SpatialGLM <- S.CARleroux(formula = formula, data = data, family =
"binomial", trials = trials, W=W, burnin = burnin, n.sample = n.sample,
                                    thin=thin)
```

All priors used for this thesis are uninformative priors, however, it is possible to specify specific priors using the functions such as `prior.mean.beta`, `prior.var.beta`, `prior.tau2`. The uninformative prior for this thesis are given below:

- $\boldsymbol{\beta} \sim N(0, 100000)$ for all regression parameters.
- $\tau^2 \sim Inverse - Gamma(1, 0.01)$
- $\rho \sim Uniform(0, 1)$

All models were initially run for `burnin = 20,000`, `n.sample = 120,000` and `thin = 10` to produce 10,000 samples. This would then be adapted if necessary.

2.3.3 Spatio-Temporal Modelling

Data that vary by area and time pose the problem of modelling both spatial and temporal auto-correlation. As previously described in section 2.3, neighbouring areas are expected to be related to one another, however, with spatio-temporal data it is expected that neighbouring areas and time periods to have more similar values than non-neighbouring areas and further apart time periods. More simply, things that are close to each other, at the same time, are more similar than those that are far away at different times. A Bayesian framework is again adopted for implementing spatio-temporal models with MCMC simulations (section 2.17), applying Bayesian hierarchical models with a conditional-autoreggresive (CAR) priors.

Spatio-temporal modelling can be thought of an extension of the spatial model outlined in equation 2.18. The spatio-temporal auto-correlation is accounted for by a latent component for k areal units and t time periods for one or more sets of spatio-temporally autocorrelated random effects ψ_{kt}: denoted by $\boldsymbol{\psi} = (\boldsymbol{\psi_1}, ..., \boldsymbol{\psi_N})$ for $t = 1, ..., N$, where $\boldsymbol{\psi_t} = (\psi_{1t}, ..., \psi_{kt})$ for $k = 1, ..., K$. The spatio-temporal generalised linear mixed model is then presented in equation 2.22:

$$g(\mu_{kt}) = \boldsymbol{x}_{kt}^T \boldsymbol{\beta} + O_{kt} + \psi_{kt} \tag{2.22}$$

where $g(\mu_{kt}) = \boldsymbol{x_{kt}}^T \boldsymbol{\beta}$ represents the general form of a GLM with link function $g(.)$ with $x_{kt} = x_{kt1}, ..., x_{ktp}$ for p known covariates for areal unit k at time point t and $\boldsymbol{O_{kt}}$ represents a vector of known *offsets* for areal unit k and time period t, if required. This model can be implemented for *Binomial*, *Gaussian* and *Poisson* data models with similar specification as described in section 2.3.2, with the addition of temporally varying time periods t.

There are a number of different spatio-temporal structures for ψ_{kt} where the spatial autocorrelation is still defined by the spatial adjacency matrix ($\boldsymbol{W}$), similar to a purely spatial analyses. This thesis applies three spatio-temporal models which were chosen upon the assumptions in the specific analyses. The main goal is to model the spatio-temporal variability within the data, while allowing for varying levels of spatial and temporal autocorrelation. The *Rushworth et al. (2014)* spatio-temporal AR(1) is suitable as it models the spatio-temporal pattern in the mean and is flexible for varying levels of both spatial and temporal autocorrelation. This model only requires one set of random effects to be estimated:

$$\psi_{kt} = \phi_{kt}$$

where ϕ_{kt} substitutes the random effect in the spatio-temporal generalised linear mixed effects model in equation 2.22.

This model proposes a spatio-temporal structure with a first order autoregressive process with a spatially correlated precision matrix, although, a second order process is also available if required. $\boldsymbol{\phi_t} = (\phi_{1t}, ..., \phi_{kt})$ represents the vector of random effects for time period t, which evolves over time by a multivariate first order autoregressive process with an autoregressive parameter ρ_T. $\boldsymbol{Q}(\boldsymbol{W}, \rho S)$ represents the precision matrix, relating to the *Leroux* CAR model proposed in section 2.21:

$$\phi_t | \phi_{t-1} \sim N\left(\rho_T \phi_{t-1}, \tau^2 \boldsymbol{Q}(\boldsymbol{W}, \rho_S)^{-1}\right) \quad t = 2, ..., N, \tag{2.23}$$

the temporal autocorrelation is induced by the mean $\rho_T \phi_{t-1}$ and the spatial autocorrelation is induced by the variance $\tau^2 \boldsymbol{Q}(\boldsymbol{W}, \rho_S)^{-1}$, where τ^2 represents the overall spatio-temporal variance parameter. The set of random effects are specified for ϕ_1, time point 1, using a *Leroux CAR* prior as ϕ_0 does not exist: $\phi_1 \sim N\left(\boldsymbol{0}, \tau^2 \boldsymbol{Q}(\boldsymbol{W}, \rho S)^{-1}\right)$.

Uninformative priors were again used for $\tau^2 \sim Inverse-Gamma(1, 0.01)$ and $\rho_S, \rho_T \sim Uniform(0, 1)$.

The AR(1) model does not allow for the spatial and temporal variances to be modelled separately as it only provides an overall spatio-temporal variance τ^2. For the analyses in Chapter 7, which assesses the proportion of individuals testing positive for COVID-19 over time by PCDs, it is of interest to assess the spatial and temporal variance separately, while also including a space $\times$ time interaction, as the spatial autocorrelation of COVID-19 data fluctuate over time. A similar study of joint disease mapping of COVID-19 cases and death by local authorities in England adopted the *Knorr-Held* model including an interaction term, which found this model to be the best fitting model according to the the log-likelihood when compared to the AR(1) model, and carried forward as the final model for inference [126].

The ANOVA model, proposed by *Knorr-Held* in 2000 [127], is suitable for these purposes as this model decomposes the spatio-temporal variation into 3 components: an overall spatial effect common to all time periods (ϕ_k); an overall temporal effect common to all spatial units (δ_t) and a set of independent space-time interactions (γ_{kt}):

$$\psi_{kt} = \phi_k + \delta_t + \gamma_{kt}$$

For this model, a temporal neighbourhood matrix is defined similar to $\boldsymbol{W}$ such that $\boldsymbol{D} = (d_{tj})$, where $d_{tj} = 1$ if $|j - t| = 1$ and $d_{tj} = 0$ otherwise. The model specifications are then given by,

$$\phi_k|\phi_{-k}, \boldsymbol{W} \sim N\left(\frac{\rho_S \sum_{j-1}^{K} \omega_{kj}\phi_j}{\rho_S \sum_{j=1}^{K} \omega_{kj} + 1 - \rho_S}, \frac{\tau_S^2}{\rho_S \sum_{j=1}^{K} \omega_{kj} + 1 - \rho S}\right), \tag{2.24}$$

$$\delta_t|\delta_{-t}, \boldsymbol{D} \sim N\left(\frac{\rho_T \sum_{j-1}^{N} d_{tj}\delta_j}{\rho_T \sum_{j=1}^{N} d_{tj} + 1 - \rho_T}, \frac{\tau_T^2}{\rho_T \sum_{j=1}^{K} d_{tj} + 1 - \rho_T}\right)$$

$$\gamma_{kt} \sim N(0, \tau_I^2)$$

The spatio-temporal autocorrelation is modelled by a common set of spatial and temporal random effects: $\boldsymbol{\phi} = (\phi_1, ..., \phi_k)$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_T)$. Again, these are both modelled by a Leroux CAR prior. The set of independent space-time interactions, $\boldsymbol{\gamma} = (\gamma_{11}, ..., \gamma_{KT})$ are optional. (ρ_S, ρ_T) are assumed to have a uniform prior, with conjugate priors for $(\tau_S^2, \tau_T^2, \tau_I^2)$ priors. Uninformative priors were used: $\tau_S^2, \tau_T^2, \tau_I^2 \sim Inverse - Gamma(1, 0.01)$, and $\rho_S, \rho_T \sim Uniform(0, 1)$. Due to the number of parameters being estimated in this model, a large amount of data are required to model estimates effectively which is a common problem with this model.

Chapter 4 analyses *C-difficile* infection data over time by IZs. It is of interest to assess individual IZ time trends of CDI incidence to see if clusters of IZs followed similar time trends over the five year period. To model this, these analyses follow a study which proposed a Bayesian space-time model for analysing clustered areal units based on disease trends [128]. This model provides information on whether areas are presenting increased, decreased, or no change in risk of disease.

This spatio-temporal model estimates clustered trends amongst areas and is represented by two components: an overall spatial structure and multiple temporally varying trends including linearly decreasing, linearly increasing and constant. Covariates are not allowed in this model but *offsets* can be used. This model is to primarily identify clusters of areas with similar temporal trends and highlight any differences between areas.

The model proposed by *Napier et al. (2019)* which fits an overall spatial pattern $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$ that is common for all time periods, with *Leroux CAR* prior. The areas are clustered by specified temporal trends, defined by $S = (f_1(t|\gamma_1), ..., (f_S(t|\gamma_S))$ where the following relates to equation 2.22:

$$\psi_{kt} = \phi_K + \sum_{s=1}^{S} \omega_{ks} f_s(t|\gamma_s),$$

An area k is assigned to one of the S trends using the binary indicator $\omega_k = (\omega_{k1}, ..., \omega_{kS})$, where $\omega_{ks} = 1$ if k is assigned to trend s. The region-wide probabilities, e.g. the probability of being assigned to trend S, are associated with each proposed trend given $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_S)$:

$$\phi_k|\boldsymbol{\phi}_{-k}, \boldsymbol{W}, \rho, \tau^2 \sim N\left(\frac{\rho \sum_{j-1}^{K} \omega_{kj}\phi_j}{\rho \sum_{j=1}^{K} \omega_{kj} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^{K} \omega_{kj} + 1 - \rho}\right) \tag{2.25}$$

$$\boldsymbol{\omega}_k = (\omega_{k1}, ..., \omega_{kS}) \sim Multinominal(1; \boldsymbol{\lambda}),$$

$$\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_S) \sim Dirichlet(\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_S))$$

For the purposes of this thesis the chosen temporal trends were represented by a constant trend ($f(t) = 0$); linearly decreasing (LD) ($f_1(t|\gamma) = \gamma t$) and linearly increasing (LI) ($f_2 1(t|\gamma) = \gamma t$). Where, the LD and LI trends are distinguished by their priors, such that the LD has a normal positive prior and LI has a normal negative prior. Each k area is assigned to a trend S and an associated region-wide probability is provided for each trend for every areal unit. Uninformative conjugate priors are shown for ρ and τ^2.

Application of Spatio-Temporal Models in R

All spatio-temporal modelling is conducted using the **CARBayeST** R-package with the functions: `ST.CARar()`, `ST.CARanova()` or `ST.CARclustrends()` to assess spatio-temporal variability within the space-time data. These models can all be fit for multiple distributions, however, this thesis only used Poisson and Binomial and link to the spatio-temporal GLMM in equation 2.22. Model specification are the same as in the spatial only modelling with further adaptations including:

- Choose either `AR=1` or `AR=2` in the `ST.CARanova()` space-time model
- Specify `interaction = TRUE or FALSE` with default `TRUE` in the `ST.CARanova()` space-time model
- Select the desired trends to be assessed using `trends =` argument. Options include: constant ('Constant'); linear decreasing ('LD'); linear increasing ('LI'); Known change point, where the trend can increase towards the change point before subsequently decreasing ('CP'); or decrease towards the change point before subsequently increasing ('CT'); and monotonic cubic splines which are decreasing ('MD') or increasing ('MI'). At least two trends have to be selected, with the constant trend always included, for function `ST.CARclustrends()`.

All models are initially run for `burnin = 50,000`, `n.sample = 250,000` and `thin = 20` to produce 10,000 samples. This would then be adapted if necessary.

2.4 Conclusion

This chapter summarises the spatial analysis methods applied throughout this thesis, highlighting the justifications for the choice of these methods. It is acknowledged that this area of research is vast and, therefore, it is unfeasible to comment in detail on the breadth of all possible analysis techniques. However, this chapter describes the fundamental concepts to consider when analysing spatial and spatio-temporal data.

There are three main themes to methods in this thesis: (1) understanding spatial autocorrelation for point-location and areal-unit data; (2) handling multi-level spatial data or incompatible spatial data - data which is available for the same area but where the sub-areas overlap and are not completely nested within each other and (3) modelling spatial and spatio-temporal data accounting for the spatial and temporal correlations. For each of these themes, the choice of methods are motivated by their suitability to the data; recognition in the literature and efficiency of implementation.

This thesis presents novel applications of these established techniques intending to maximise the use of the routinely collected data in a spatial framework and infer on population health.

Chapter 3

Spatial Analysis of GP Practice Antibiotic Prescribing in Scotland

3.1 Introduction

In 2010, the Scottish Antibiotic Prescribing Group (SAPG) proposed a framework to support Antimicrobial Management Teams (AMTs) to provide local surveillance of antimicrobial use and resistance. A review of the Healthcare Associated Infection (HAI) surveillance program in 2015 agreed a mandatory local surveillance framework in Scotland which outlined policy requirements to be implemented across all NHS Scotland health boards. This included monitoring the rates of antibiotic prescribing in primary-care, specifically in the context of information on antimicrobial resistance and *c-difficile* infection. This framework highlighted the importance of analysing national prescribing indicators at GP practice level to identify outliers and for NHS health boards to provide feedback to prescribers. The top two primary care antibiotic prescribing indicators to be monitored on a annual basis included: total antibiotic prescribed items per 1000 patients per day and the use of agents associated with a higher risk of CDI (cephalosporins, fluroquinolones, co-amoxiclav and clindamycin) items per 1000 per day. These indicators were also encouraged to be expressed as % of total antibacte-

rial prescriptions. Other indicators included the use of recommend antibacterial agents such as amoxicillin items per 1000 per day and seasonal variation of the use of fluroquinolones [129].

The Scottish Reduction in Antimicrobial Prescribing programme (ScRAP) was launched in 2013 by SAPG, primarily aimed at GP practices to develop targets to reduce the volume of antimicrobial prescriptions in primary care and develop a more sustainable strategy for antibiotic use. The Scottish One Health Antimicrobial Use and Antimicrobial Resistance Annual Report in 2016 showed the total antibiotic prescribing rate in primary care, was 2.0 items per 1000 registered patients per day, with 29% of the Scottish population receiving an antibiotic prescription in 2016. It was also reported that primary care prescribing accounted for 79.8% of all antibiotic use in humans [130]. The 2018 report stated that antibiotic use in primary care had decreased to 1.84 items per 1000 registered patients per day with 27.3% of the population receiving an antibiotic course [131].

Primary care antibiotic prescribing is likely to be influenced by a range of factors, however GP practice population demographics have been shown to be important. A study exploring the variation of antibiotic prescribing by GP practices in the UK found that the majority of variation between practices could not be explained by the prevalence of practice population comorbidities. However, the consultation rates for respiratory tract infections and high prescribing rates for corticosteroids explained most of the variation, alongside other factors such as the age distribution of the practice population [132]. Prescribing differences have also been shown between dispensing and non-dispensing GP practices, with a study in England indicating that dispensing GP practices were more likely to prescribe higher cost drugs and highlighted a potential financial conflict of interest in treatment decisions [133], which motivated the analyses in this chapter to

compare between dispensing and non-dispensing practices with reference to the amount of antibiotics prescribed.

A Welsh study highlighted increased primary care antibiotic prescribing in areas of increased deprivation [134], with a study in Scotland drawing similar conclusions. In 2012, patients in the most deprived SIMD quintile showed a total antibiotic prescribing rate that was 36.5% higher than those in a least deprived quintile [135]. Furthermore, a study of prescribing in England attributed higher antibiotic prescribing in more deprived areas and highlighted the importance of accounting for areal differences when setting prescribing targets to ensure communities were not inappropriately penalised [136].

A study in the USA, from 2010 to 2017, showed that influenza vaccination uptake at a population level had an association with a reduction in antibiotic prescribing in the community, after controlling for confounders such as socioeconomic differences, access to healthcare and state-level differences [137]. Although influenza is a viral infection, the association with antibiotic prescribing may be due to a combination of factors such as appropriate usage in treating secondary bacterial infections due to influenza and inappropriate use due to mis-prescription of antibiotics [137]. A study in the UK used electronic health records (EHR) to explore the impact of influenza vaccination on amoxicillin prescribing in older adults, which showed a reduction in amoxicillin prescribing for the vaccinated group. This study also attributed antibiotic use for influenza due to secondary bacterial infection [138]. A study in Ontario showed similar results in the reduction of influenza-associated antibiotic prescriptions, comparing antibiotic rates pre and post introduction of the universal influenza immunisation program in 2000. [139].

This chapter examined Scottish GP antibiotic prescribing rates from 2016 to 2018 to investigate prescribing patterns at NHS health board and GP practice level, specifically:

1. How GP practice total antibiotic prescribing rates (items/1000/day) change over time.

2. The differences between GP practice prescribing rates within NHS Scotland health boards for total antibiotic and high-risk antibiotic prescribing groups.

3. The spatial association between antibiotic prescribing rates by neighbouring GP practices.

4. Whether GP practices prescribe similarly across all antibiotic groups.

5. The relationship between GP practice demographic information and total antibiotic prescribing rates including practice population age distributions; dispensing and non-dispensing GP practices and measures of most and least deprived populations.

6. The relationship between influenza vaccination uptake in registered patients aged over 65 and the rate of antibiotic prescribing by GP practices.

3.2 Methods

The aim of this analysis was to explore the variation, including spatial variation, in GP practice antibiotic prescribing in Scotland. This analysis was motivated by the SAPG prescribing indicators described in section 3.1. Spatial mapping and auto-correlation methods were applied to visualise the pattern of antibiotic prescribing, while also exploring the variability in prescribing with a principal component analysis (PCA) and generalised linear models (GLM). The goal was to understand the overall picture of antibiotic prescribing by GP practices.

3.2.1 Data

All GP practice prescriptions from 2016 to 2018 were obtained from NHS Open Data platform for Prescriptions in the Community [140]. This included the number of items prescribed by GP practices in Scotland, each defined by a unique GP practice code and associated NHS Scotland health board code. Items were described by British National Formulary (BNF) item descriptions. BNF codes represent the unique identifier for all prescribed items and provide information about each prescription. For the purpose of this analysis all BNF codes beginning with "0501" were included as this relates to Chapter 5, section 1: Infections and Antibacterial Drugs. The next two digits of the code represent the BNF paragraph describing specific antibiotic drug groups such as penicillins (figure 3.1) [141].

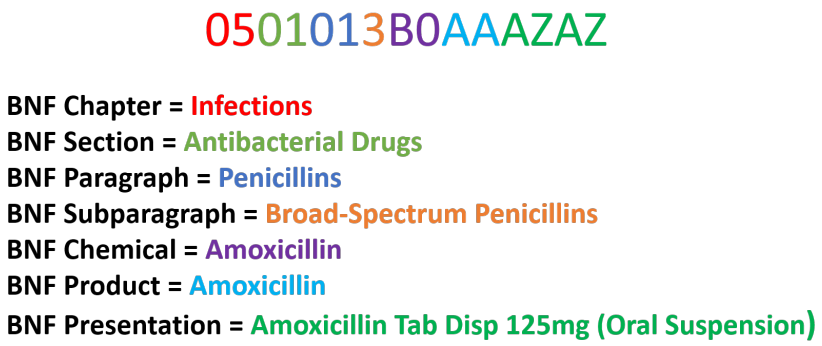


Figure 3.1: Example of the BNF code structure.

In this study the focus was on the 13 BNF antibiotic paragraphs (antibiotic groups) to aggregate prescribing, which include: penicillins; cephalosporins and other beta-lactams; tetracyclines; aminoglycosides; macrolides; clindamycin and lincomycin; some other antibacterials (unique or rarely prescribed); sulfonamides and trimethoprim; antituberculosis drugs; antileprotic drugs; metronidazole, tinidazole and ornidazole; quinolones and urinary-tract infections [142].

GP practice age distributions, dispensing, population and location data were obtained for 2016, 2017 and 2018 from NHS Open Data files for 'GP Practice Population Demographics' and 'GP Practices and List sizes' [143]. Location data included GP practice postcodes which were converted to latitude and longitude coordinates and linked to GP practices. SIMD deprivation scores by GP practice were obtained from ISD General Practice Data Tables. GP practice populations are summarised by the SIMD quintile associated with their place of residence, which provided the percentage of practice populations residing in the most and least deprived SIMD quintiles [144]. Finally, influenza vaccination data for GP practice populations aged over 65 were obtained from the respiratory team at Health Protection Scotland, for 2016 only.

Data Linkage for GP Antibiotic Prescriptions

All 2016, 2017 and 2018 prescriptions were downloaded and subset to include BNF codes beginning with *0501*, ensuring only antibiotics were included. These data were then merged to a BNF glossary file to introduce antibiotic group names (e.g. Penicillin). There were a number of BNF codes in the data that were not present in the glossary file, producing high numbers of missing values (NAs = 957) as a result of some BNF codes changing halfway through 2016. The drug descriptions matched, however, codes were slightly different. Comparing whether old and new BNF codes were present in 2015 and 2017, confirmed these differences which following recoding, reduced the missing data from NA = 957 to NA = 27. These 27 prescriptions could not be recovered and therefore removed from the analyses.

The number of prescribed items were aggregated by GP practice and antibiotic groups (13 BNF Paragraphs) then restructured for each year individually (2016 to 2018), from long data to wide data, to ensure the correct form for analyses: count of antibiotic items prescribed per GP practice, separated by 13 antibiotic groups and then summated across all antibiotic groups to calculate the total antibiotic items prescribed

also per GP practice. GP practice demographic information were then linked to prescribing records by GP unique practice code, however there were inconsistencies in the information available with many of the GP practices showing no practice population sizes. Contacting ISD helped locate some missing GP practice information however they could not all be recovered. There were also GP practices within the data that were inappropriate for these analyses due to the lack of consistency between practice population size reporting, e.g Access Practices for the Homeless, Behaviour Centres and University Health Services. These practices did not have realistic or consistent practice list sizes and therefore were removed.

GP practice demographic information were then merged to complete the data sets. Deprivation scores were obtained from Information Service Division (ISD) website, however the recorded GP practice populations within files were largely inconsistent to the previously merged GP practice populations [144]. Deprivation practice populations files were consistent between years and to other deprivation records. Similarly, the GP descriptive files had consistent populations and therefore it was decided that deprivation proportions would be calculated within the deprivation files independently then merged to the main GP practice data set.

This process was repeated for 2017 and 2018 with year specific GP practice demographic records, before combining all three years of data. Antibiotic rates (items/1000/day) were calculated by dividing GP practice antibiotic counts by GP practice population size, multiplied by 1000 and divided by the number of days in the year.

The Influenza data were merged by unique GP practice code and the percentage of Influenza vaccinations in registered patients aged over 65 was calculated. A subset of the antibiotic data was created for these analyses, including this variable, GP practice antibiotic prescriptions and demographic information for 2016.

The final data structure for these data is presented in figure 3.2, including the Influenza vaccination for 2016.

Practice.Code	PracticeListSize	Dispensing	HB2014	Aminoglycosides	Antileprotic Drugs	Antituberculosis Drugs	Cephalosporins and other Beta-Lactams	Clindamycin and Lincomycin	Macrolides
10002	7446	0	S08000030	8	9	1	31	5	295
10017	6803	0	S08000030	0	0	7	22	8	164
10036	4554	1	S08000030	0	11	0	34	1	112
10106	6054	0	S08000030	2	8	0	13	3	281

Macrolides	Metronidazole, Tinidazole & Ornidazole	Penicillins	Quinolones	Some Other Antibacterials	Sulfonamides And Trimethoprim	Tetracyclines	Urinary-Tract Infections	Total	Health.Board.Name
295	121	1749	77	7	466	1082	423	4274	NHS Tayside
164	122	1782	60	4	547	724	537	3977	NHS Tayside
112	85	1237	54	13	442	499	284	2772	NHS Tayside
281	95	2118	70	12	531	781	292	4206	NHS Tayside

Under4	Under15	Over85	Over74	QT1...Most.Deprived	Q5...Least.Deprived	lat	long	Level.6.Pop.65	Level.6.Vacc.65	Level.6.Rate
4.862324	16.601746	1.853593	8.099396	4	28	56.49558	-3.06606	1495	1066	0.7130435
3.391064	13.898996	4.002329	12.239849	0	15	56.37045	-3.84263	1671	1277	0.7642130
3.699650	13.988616	2.736427	10.858144	0	0	56.61851	-3.87052	1169	873	0.7467921
4.188841	15.193133	4.652361	14.695279	2	54	56.46868	-2.88297	1687	1270	0.7528156

Figure 3.2: Data structure for 2016 GP antibiotic prescriptions merged with influenza vaccination uptake in $\geq$ 65 years old.

3.2.2 Statistical Methods

Descriptive Statistics

Descriptive statistics were obtained by summarising GP practice population and antibiotic rates by year and NHS health board. GP practice demographic and total antibiotic prescribing were presented by Median (IQR) for each year. A line plot was created to show the change in total antibiotic prescriptions (items/1000/day) over three years,

separated by NHS health boards. GP practices were categorised as either *Low*, *Medium* or *High* prescribers based on tertiles of 2016 total antibiotic prescribing data, then assessed to see if GP practices changed category over time. A Sankey plot was also created to highlight the differences between GP practices' total antibiotic prescribing rates (items/1000/day) between each of the tertiles over three years [145].

Mapping and Exploratory Spatial Analysis at NHS Scotland Health Board Level

The antibiotic prescription data were aggregated by year and health board to show rates of GP practice antibiotic prescribing by health boards. The Scottish health board shape-files were obtained and merged to the aggregated data set to create a `spatialpolygondataframe` object where each polygon defines the outline of one health board (section Methods 2.1.2). The R package used for this process was **sp** [97]. These data could then be used to create maps of the health boards for antibiotic prescribing rates.

Firstly, total GP practice antibiotic prescribing rates were plotted by the 14 health boards for 2016, 2017 and 2018. Maps were then also created to visualise the difference in prescribing of broad spectrum high-risk antibiotic groups: Cephalosporins, Co-amoxiclav, Quinolones and Clindamycin. Moran's I test for spatial autocorrelation was applied to assess for spatial association between total and high-risk aggregated GP practices' antibiotic prescribing rates by health board (section 2.14).

Exploratory Spatial Analysis at GP Practice Level

Point maps for individual GP practices were created to compare neighbouring GP practice antibiotic prescribing rates for the whole of Scotland. Subset maps of the health boards containing Scotland's two largest cities (Glasgow and Edinburgh) were also cre-

ated to show a closer picture of GP practice prescribing: NHS Greater Glasgow and Clyde and NHS Lothian. GP practices were colour coded into categories of prescribing to compare GP practice antibiotic rates. All Scottish health boards were required to set antibiotic targets for 2016, 2017 and 2018 [146], however, NHS Lothian was the only health board to publish their targets ($\leq$ 1.9 prescribed items per 1000 registered patients per day for 2016, 2017 and 2018). Therefore, in this study, the Lothian targets were assumed for all NHS Scotland health boards. No additional factors were accounted for other than population size.

Sample variograms were plotted for the crude total antibiotic prescribing rates for 2016, 2017 and 2018 to visualise spatial dependence within the data and random permutation of the variograms were run to include Monte Carlo Envelopes in each of the variograms as a test for spatial autocorrelation (see chapter 2, section 2.1 and 2.2.1).

The R-package **ggmap**, and the function `get_googlemap`, were used to superimpose GP practice prescribing information onto Google maps of Scotland [147]. Coordinates of Scotland were used to determine the Google map position to which `ggmap` was applied, matching longitude and latitude coordinates for each GP practice.

Principal Component Analysis

A Principal Component Analysis (PCA) was conducted on the 13 antibiotic groups to explore whether GP practices followed similar prescribing trends for specific antibiotic drug groups.

PCA is a statistical tool that aids understanding of variability and patterns within a data set. It is an unsupervised learning method as it depends upon a set of variables $X_1, ..., X_p$ with no association to a response variable Y.

This technique reduces the dimensions of a correlation/covariance data matrix and returns linearly uncorrelated components which collectively summarise the variation in the data set.

PCA is a useful tool for visualisation. Often 2-dimensional scatterplots are used to visualise n observations for every combination of p variables however, as p grows large this becomes impractical and PCA provides an alternative approach. For example, if the prescribing rates of the p=13 antibiotic groups were plotted in 2-dimensions, with one for every combination of antibiotic groups, producing 78 scatterplots which would be ineffective in communicating the results. PCA recalculates a low dimensional version of a data set while maintaining important information (variability) from the original data.

The principal components (PCs) Z_1,...,Z_n for a set of variables $X_1, ..., X_p$, such that $n = p$, are normalised linear combinations of the variables such that the first component can be presented as:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_p 1X_p \tag{3.1}$$

$\phi_{11}, ..., \phi_{p1}$ are the *loadings* (or eigenvalues) of this first component (PC1) which define the direction in which the data vary most. The second principal component (PC2) is then the projection in the direction that explains the most variability that is orthogonal to PC1. Hence, the k$^{\text{th}}$ principal component (PCK) is in variance-maximising direction, orthogonal to the previous $k - 1$ components. Orthogonality is ensured by fact that eigenvectors must be orthogonal to one another and therefore returning uncorrelated principal components. The eigenvalue and eigenvectors are calculated using eigen-decomposition of the correlation matrix. For application in R, the command

`prcomp` uses a more generalised version of eigen-decomposition called singular value decomposition, which is the preferred method for numerical accuracy [148].

For this analysis, a correlation matrix was initially produced to assess the correlation between the 13 antibiotic groups prescribing rates. After conducting the PCA on all antibiotic drug groups, the cumulative variance was calculated to show how many PCs would be required to explain the majority of the variability within the data. Plots of the principal components were then obtained. A `biplot` was created to show the direction of influence for each antibiotic group.

GP practice demographic information was plotted for PC1 and PC2 to assess for clustering related to age, deprivation and if the practice was dispensing or not. The percentage of the GP practice population aged under 15, over 74 years, most deprived SIMD quintile and dispensing/ non-dispensing were assessed by dividing GP practices into two colour coded groups (red or blue). These groups were split by a threshold determined by either the lower or upper interquartile ranges of these variables.

Generalised Linear Models

Negative binomial Generalised Linear Models (GLMs) were used to assess total antibiotic prescribing rates and the relationship with GP practices' descriptive factors. The residuals from this model were assessed for spatial association, applying Monte Carlo Envelopes to test for spatial dependence. The data for influenza vaccination uptake in patients aged over 65 were only available for 2016. Therefore, the antibiotic data were subset to only include that year in order to compare influenza vaccination rates to GP practice antibiotic prescribing, having adjusted for GP practice demographics.

The antibiotic data were count data which typically assume a Poisson distribution and has the strict property of equal mean and variance. This property does hold if the data are over-dispersed ($E(Y) \neq Var(Y)$), which was found to be the case for the antibiotic data. Negative-Binomial is an alternative approach for dealing with over-dispersion as it does not require the mean and variance to be equal with the addition of a dispersion parameter, and was adopted for this analysis.

Y is defined by the number of antibiotic items prescribed and μ is equal to the expected number of antibiotic items prescribed:

$$E(Y) = \mu$$

$$Var(Y) = V_{NB}(\mu) = \mu + \phi\mu^2 = \mu(1 + \phi\mu)$$

The over-dispersion is represented by the multiplicative factor $(1 + \phi\mu)$ which depends on μ.

Negative-Binomial GLM

The response y_{it} represents the number of antibiotic items per GP practice, i, for each year, t, where δ_t represents the fixed effect for the three years. An offset for GP practice population is denoted by N_{it}. Independent variables include dispensing vs non-dispensing GP practices ($Dispense_{it}$); percentage of GP practice population aged under 15 ($Under15_{it}$), over 74 ($Over74_{it}$) and residing in the least deprived quintile ($LeastDep_{it}$) and in the most deprived quintile ($MostDep_{it}$) where a unit increase corresponds to a 1% increase for each covariate, presented in equation 3.2.

$$E(log(y_{it})) = log(N_{it}) + \beta_0 + \delta_2 Yr2017 + \delta_3 Yr2018 + \beta_1 Dispense_{it} + \beta_2 Under15_{it} + \beta_3 Over75_{it} + \beta_4 LeastDep_{it} + \beta_5 MostDep_{it} \quad (3.2)$$

The residuals from the multivariate model were extracted and added to the spatial point data frame which was then testing for spatial association using Monte Carlo Envelopes (section 2.2.1).

Influenza Vaccination GLM

The antibiotic prescription data were subset for 2016 only, and influenza vaccination uptake in over 65s, $Vaccination_i$, was included as a covariate, adjusting for all other GP practice variables defined in equation 3.2: the percentage of GP practice population aged under 15, over 74, residing in the least deprived and most deprived quintiles.

$$E(log(y_i)) = log(N_i) + \beta_0 + \beta_1 Dispense_i + \beta_2 Under15_i + \beta_3 Over75_i + \beta_4 LeastDep_i + \beta_5 MostDep_i + \beta_5 Vaccination_i \quad (3.3)$$

Here, y_i represents the number of antibiotic items per GP practice i in 2016.

3.3 Results

Initially, summary statistics and plots of antibiotic prescribing by year and health boards were viewed, followed by maps of antibiotic rates by health board and at GP practice level in section 3.3.1. The principal component analysis is then presented in section 3.3.2, which aims to understand variation between prescribing of antibiotic

groups. Finally in section 3.3.3, GLMs were applied to understand relationships between total antibiotic prescribing and descriptive GP practice factors such as practice demographics and influenza uptake.

3.3.1 Exploratory Analysis

GP Practice Demographics

Health boards varied by the number of GP practices and population size. NHS Greater Glasgow and Clyde was the largest health board with more than 200 GP practices and total number of registered patients accounting for more than 20% of Scotland's population. Conversely, NHS Orkney had the smallest number of GP practices (6 practices) and the lowest number of registered patients with less than 20,000 people (table 3.1).

Table 3.1: Total number of GP practices within each NHS Scotland health board by year with mean GP practice population. Ordered by population size.

Health Board Name	Number of GP practice			Mean GP Population
	2016	2017	2018	
NHS Orkney	5	6	6	19369
NHS Shetland	9	10	10	22807
NHS Western Isles	9	9	9	26862
NHS Borders	23	23	23	117922
NHS Dumfries & Galloway	31	32	32	149281
NHS Forth Valley	53	54	52	313504
NHS Highland	85	96	96	322229
NHS Fife	54	54	53	355490
NHS Ayrshire & Arran	54	55	55	382491
NHS Tayside	64	64	64	421113
NHS Grampian	73	75	73	593863
NHS Lanarkshire	99	101	98	637771
NHS Lothian	119	121	117	913954
NHS Greater Glasgow & Clyde	230	239	234	1263305
Total	908	939	922	5539962

The mean number of GP practices across three years was 923, with a total registered practice population of 5,539,962. This was higher than the National Records of Scotland (NRS) of 5.44 million [149]. There are a number of patients who will be registered at two practices and occasionally temporary patients are recorded as permanent patients. The number of GP practices also varied between years due to GP practices opening, closing and mergering (table 3.1).

Table 3.2: GP practice demographic information from 2016 to 2018. Median (IQR) percentage of practice population aged under 15 (%), aged over 74 (%), residing in the most deprived SIMD quintile (%) and residing in the least deprived SIMD quintile (%).

	Median (IQR) percentage of practice population		
	2016	2017	2018
Most Deprived (%)	13.0 (1.0 - 36.0)	12.7 (0.5 - 36.4)	12.7 (0.5 - 36.5)
Least Deprived (%)	9.0 (2.0 - 22.0)	9.2 (2.0 - 21.5)	9.2 (2.0 - 21.6)
Under 15 (%)	15.3 (14.0 - 16.9)	15.4 (13.8 - 16.9)	15.4 (13.8 - 16.9)
Over 74 (%)	8.0 (6.5 - 9.5)	8.0 (6.4 - 9.6)	8.1 (6.4 - 9.7)

GP practice population demographic percentages remain relatively constant between 2016 and 2018. The median percentage of GP practice population residing in the most deprived SIMD quintile was 13% (IQR = 1% - 37%), with a maximum of 91% and minimum of 0% across the three years. The least deprived SIMD quintile percentage was lower (Med = 9%, IQR 2% - 22%) and showed a three year maximum of 94% and minimum of 0%. Both SIMD variables were positively skewed and widely spread. The median percentage of the practice population aged under 15 years old was 15% (IQR = 14 - 17) and the median percentage of the practice population aged over 74 (%) was slightly lower (Med = 8%, IQR = 6% - 10%) (table 3.2). Age covariates were normally distributed.

Antibiotic Prescribing Rates from 2016 to 2018

The median total antibiotic prescribing rate was 1.78 (IQR = 1.52 - 2.08, items/1000/day) in 2016, with a minimum prescribing rate of 0.5 (items/1000/day) and maximum of 4.03 (items/1000/day). There was a 7.3% reduction in median total antibiotic prescribing between 2016 and 2018, with a reduction in IQR and minimum prescribing rates. However, the maximum prescribing rate in 2018 was larger than in 2016 (table 3.3).

Table 3.3: GP practice total antibiotic prescribing rates (items/1000/day) from 2016 to 2018. Median (IQR) with maximum and minimum prescribing rates per year.

	2016	2017	2018
Minimum	0.50	0.22	0.02
Q1	1.52	1.46	1.39
Median	1.78	1.73	1.65
Q3	2.08	2.00	1.91
Maximum	4.03	3.73	4.69

Most of the antibiotic prescribing group rates decreased between 2016 and 2018 although three antibiotic groups increased during this time: antileprotic drugs, clindamycin and lincomycin and some other antibacterials, however these were small changes that may be attributed to noise in the data. Penicillins were the most prescribed antibiotic class and accounted for approximately 46% of total antibiotic prescriptions each year. Furthermore, approximately 30% of all penicillin prescriptions were classed as amoxicillin. The second most prescribed antibiotic class was tetracyclines, accounting for approximately 15% of total antibiotic prescribing each year (table 3.4).

Individual GP practices' prescribing trends over time are presented in figure 3.3 and show that the total number of GP practices classed as "Low" prescribers in 2016 increased in 2017 and 2018. The number of "Medium" prescribers remained relatively similar between 2016 and 2017 but decreased in 2018, and there was a clear reduction in the number of GP practices classed in the "High" prescribing category for each successive year. The overall trend showed the majority of GP practices to either remain

Table 3.4: BNF antibiotic group prescribing rates per 1000 patients per day for 2016, 2017 and 2018 with change over time indicator.

BNF Antibiotic Groups	items/1000/day			Change (+\-)
	2016	2017	2018	
Aminoglycosides	0.0006	0.0005	0.0004	-
Antileprotic Drugs	0.0024	0.0024	0.0025	+
Antituberculosis Drugs	0.0035	0.0034	0.0032	-
Cephalosporins and other Beta-Lactams	0.0340	0.0336	0.0318	-
Clindamycin and Lincomycin	0.0023	0.0025	0.0025	+
Macrolides	0.2001	0.1862	0.1655	-
Metronidazole, Tinidazole & Ornidazole	0.0383	0.0374	0.0359	-
Penicillins	0.8290	0.7896	0.7390	-
Quinolones	0.0475	0.0463	0.0443	-
Some Other Antibacterials	0.0052	0.0067	0.0082	+
Sulfonamides And Trimethoprim	0.2228	0.2098	0.1930	-
Tetracyclines	0.2646	0.2642	0.2603	-
Urinary-Tract Infections	0.1284	0.1281	0.1271	-

in their category each year of move to the lower category however, there was some movement of GP practices from a lower category to a higher category (figure 3.3).

Assessing total antibiotic prescribing by NHS health boards, between 2016 and 2018 showed a decreasing trend for most health boards however, NHS Western Isles and NHS Tayside both showed an increase in prescribing rates from 2017-2018. NHS Orkney and NHS Lothian were the lowest prescribers with NHS Lothian showing the lowest prescribing rate in 2018. NHS Lanarkshire remained the highest prescribing health board throughout all years, however, showed a consistent decrease in antibiotic prescribing between years (figure 3.4).

In 2016, NHS Lanarkshire had the highest total antibiotic prescribing rate of 2.10 items/1000/day and NHS Orkney was the lowest (1.46 items/1000/day). In 2018, NHS Lanarkshire remained the highest prescribing health board, 1.94 items/1000/day, whereas NHS Lothian showed the lowest prescribing rate of 1.38 items/1000/day. NHS health boards in central Scotland and towards the Scottish border appeared to show

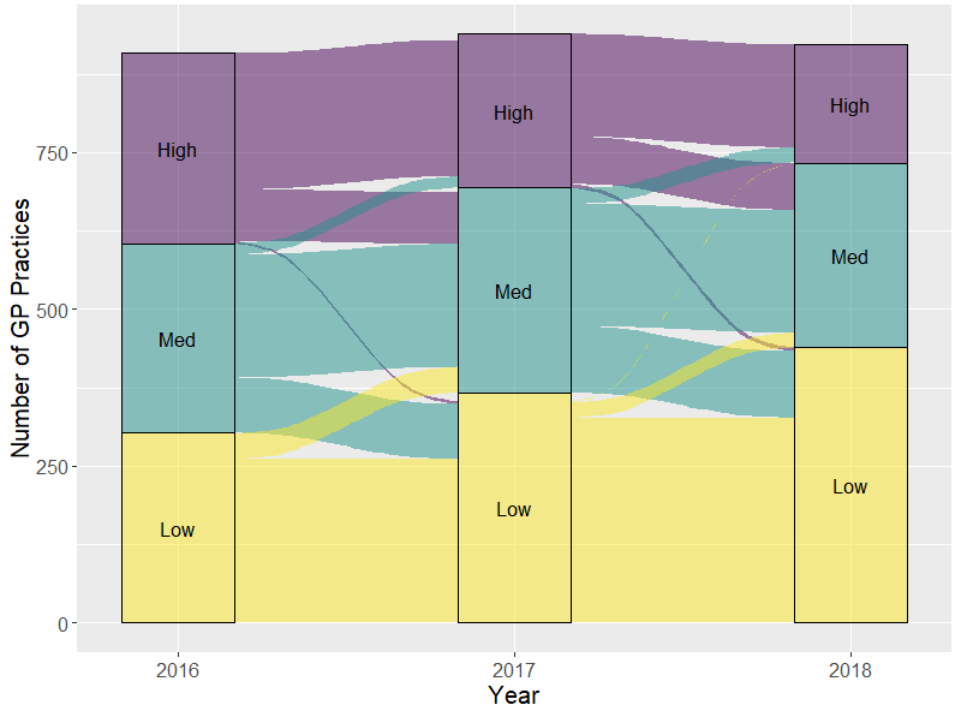


Figure 3.3: Sankey diagram for the change in GP practice antibiotic prescribing rates (items/1000/day) between 2016, 2017 and 2018. High, Medium and Low prescribing categories defined by the tertile thresholds of 2016 antibiotic prescribing rates.

higher prescribing rates compared to the northern health boards across the three years. Antibiotic prescribing shows an overall decrease from 2016 to 2018 for all health boards except NHS Orkney (figure 3.5).

Moran's I test for spatial association gave an I statistic of 0.18 ($p = 0.1598$) in 2016, suggesting positive spatial auto-correlation between health boards, although was not statistically significant. However, the I statistics for 2017 and 2018 showed strong spatial association, at a 10% significance level ($I_{2017} = 0.393$, $p = 0.010$ and $I_{2018} = 0.259$, $p = 0.07$). These results suggest there was positive spatial dependence between GP prescribing rates in health boards, while noting that this is a small number of areal units for performing this test (n=14) and no covariate information has been accounted for.

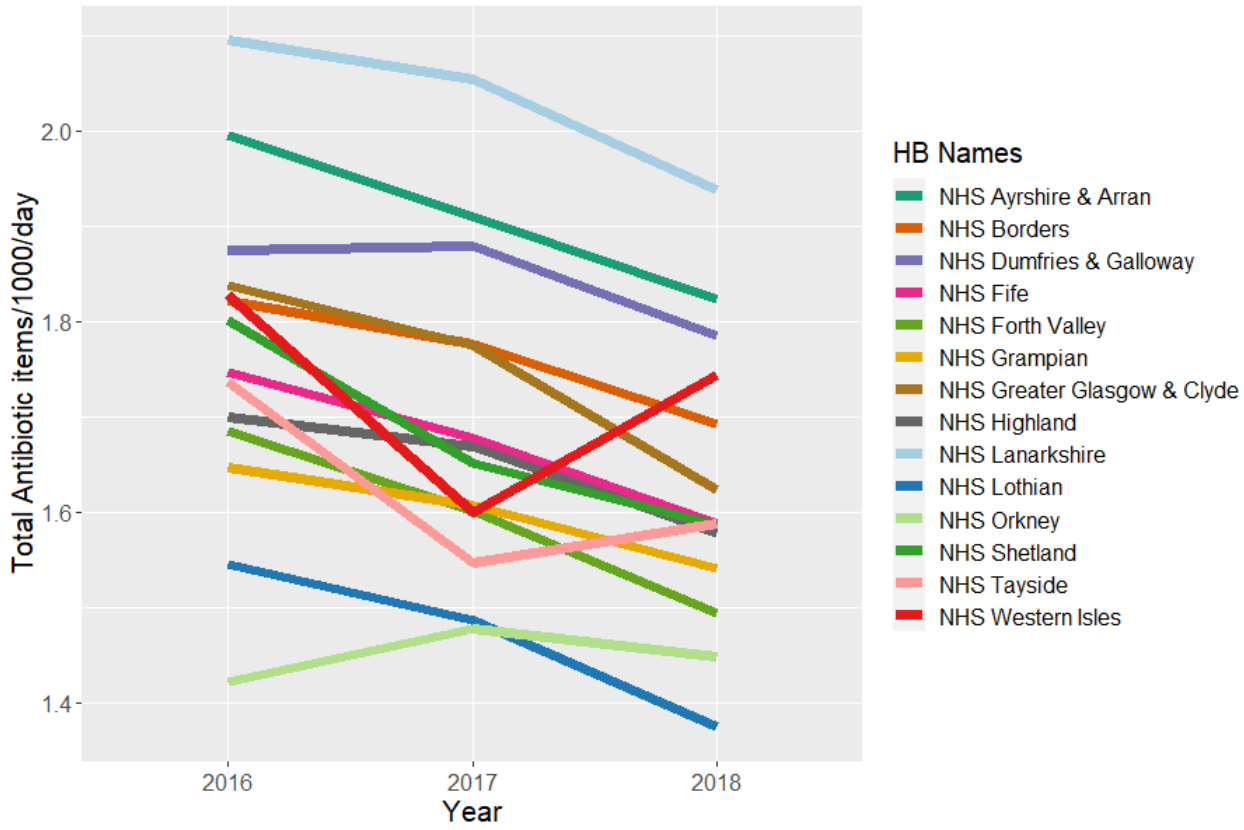


Figure 3.4: Line plot of total antibiotic prescribing rates (items/1000/day) separated by health boards from 2016 to 2018.

High-risk antibiotic groups (cephalosporins, quinolones, co-amoxiclav and clinamycin) were highlighted in the SAPG local surveillance framework to be monitored each year and, therefore, each of these antibiotic groups were mapped by health boards for 2016, 2017 and 2018 in figures 3.6, 3.7, 3.8 and 3.9.

The use of cephlosporins varied greatly between health boards, with NHS Dumfries and Galloway, Greater Glasgow and Clyde, Forth Valley, Fife and Tayside showing low rates of prescribing in comparisons to NHS Western Isles, Highlands, Ayrshire and Arran and Borders which all showed similar higher rates. The prescribing pattern remained similar between years, however some NHS boards increased from 2016 including NHS Shetland and NHS Ayrshire and Arran. Moran's I test for spatial auto-correlation gave no indication of any spatial dependence between cephalosporin prescribing across all years

($I_{2016} = -0.032, p = 0.877$; $I_{2017} = -0.047, p = 0.981$ and $I_{2018} = -0.07, p = 0.8155$) (figure 3.6).

Quinolone prescribing was similar across most health boards apart from NHS Tayside which showed the lowest prescribing rate for 2016, 2017 and 2018. NHS Dumfries and Galloway showed the highest prescribing rate across all years. NHS Shetland, Western Isle and Forth Valley all showed a decrease in prescribing from 2016 to 2018, although the pattern of prescribing largely remained similar across all years. Moran's I showed strong negative spatial correlation coefficients across all years for quinolones. Prescribing, however, was not supported by p-values ($I_{2016} = -0.226, p = 0.401$; $I_{2017} = -0.218, p = 0.416$ and $I_{2018} = -0.247, p = 0.343$). This is due to the difference in prescribing for NHS Tayside compared to all other health boards (figure 3.7).

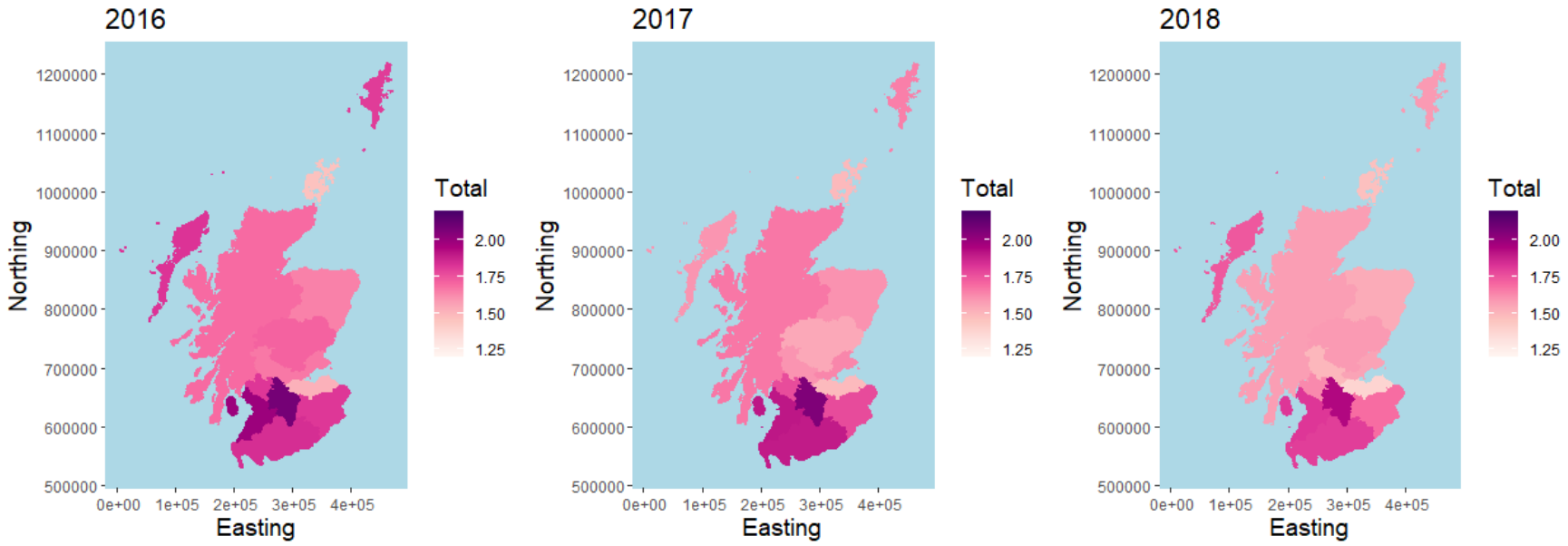


Figure 3.5: Total antibiotic prescribing rate (Total - items/1000/day) for 2016 (left), 2017 (middle) and 2018 (right) by NHS Scottish health boards.

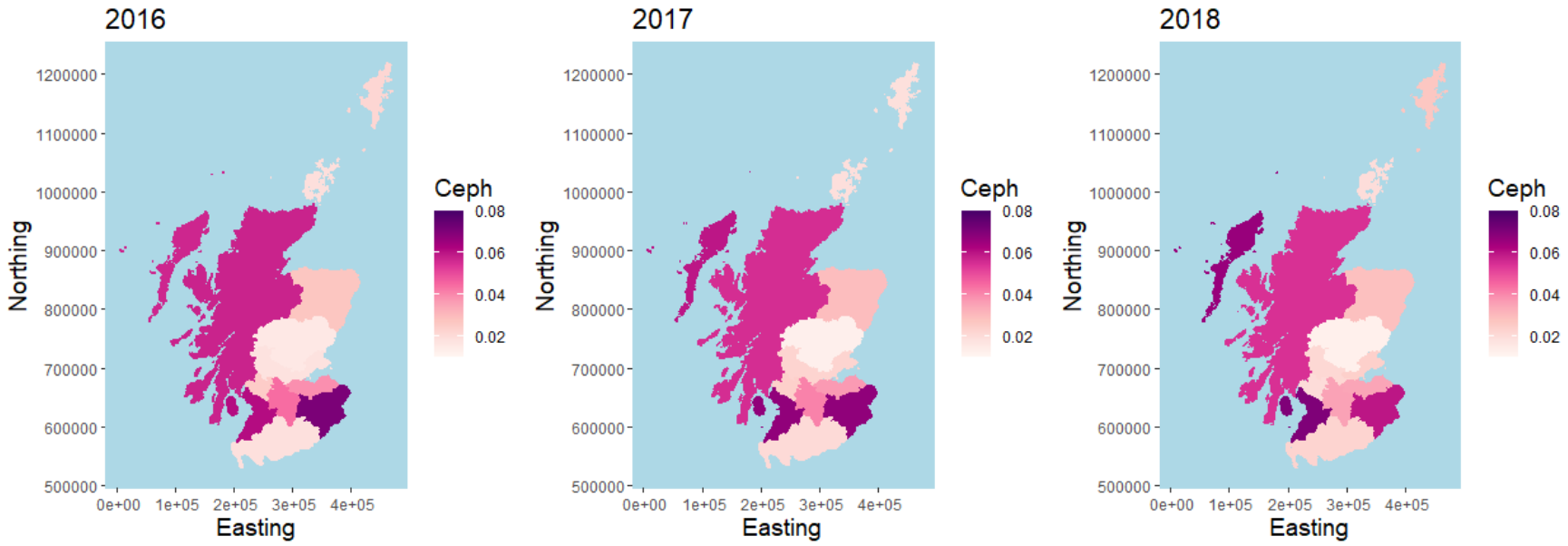


Figure 3.6: Cephlosporins prescribing rate (Ceph - items/1000/day) for 2016 (left), 2017 (middle) and 2018 (right) by NHS Scottish health boards.

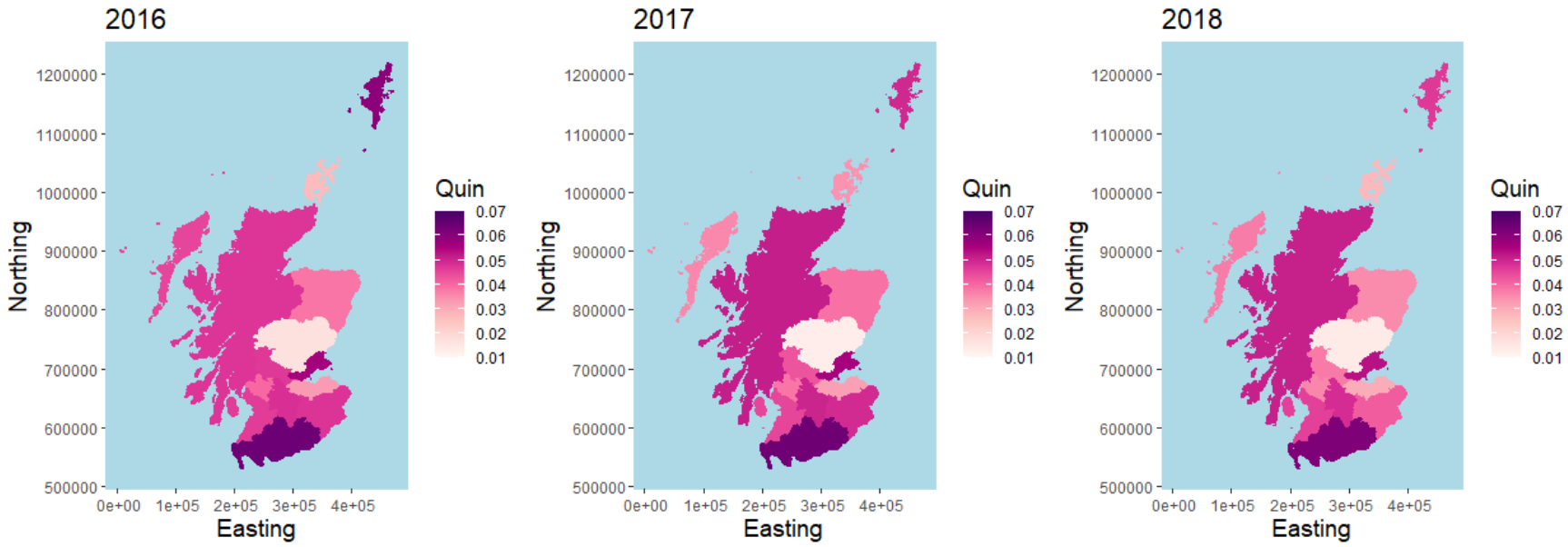


Figure 3.7: Quinolones prescribing rate (Quin - items/1000/day) for 2016 (left), 2017 (middle) and 2018 (right) by NHS Scottish health boards.

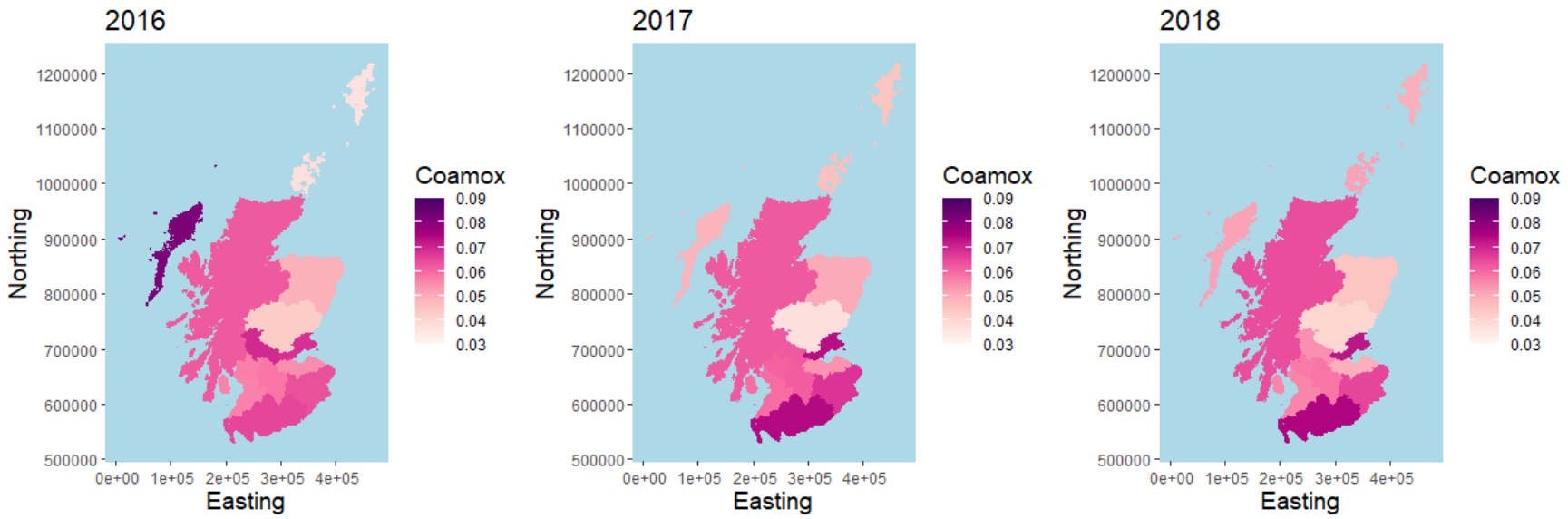


Figure 3.8: Co-amoxiclav prescribing rate (Coamox - items/1000/day) for 2016 (left), 2017 (middle) and 2018 (right) by NHS Scottish health boards.

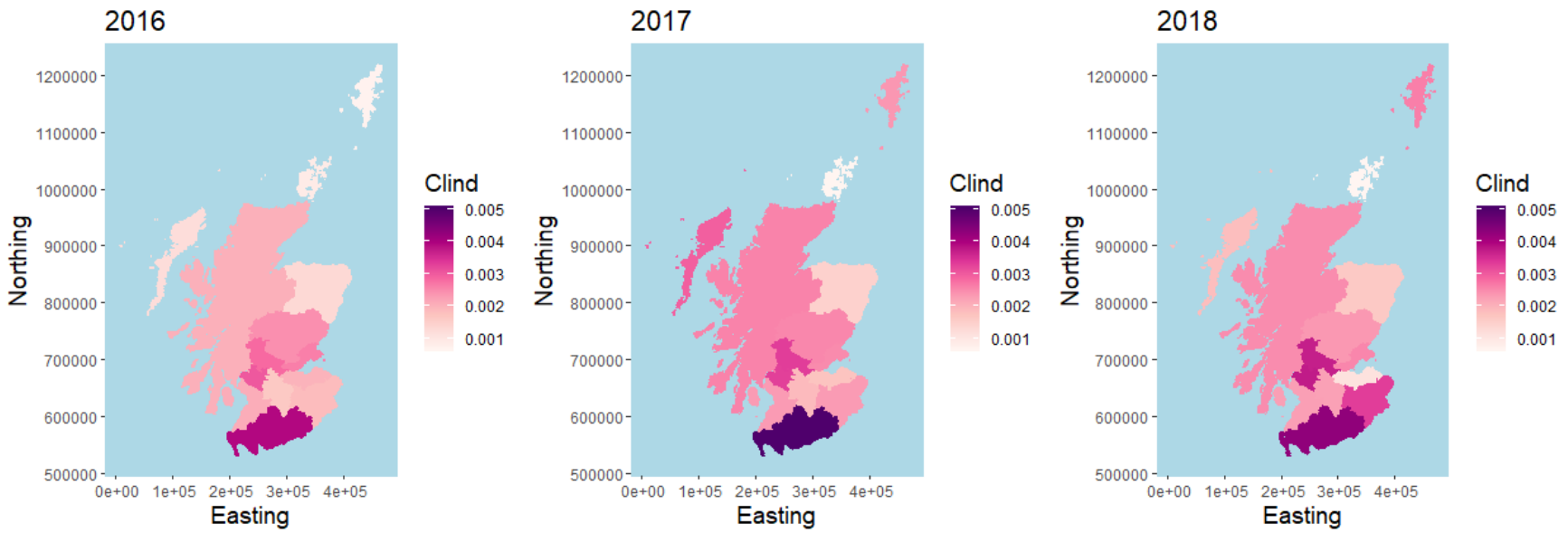


Figure 3.9: Clindamycin prescribing rate (Clind - items/1000/day) for 2016 (left), 2017 (middle) and 2018 (right) by NHS Scottish health boards.

Co-amoxiclav prescribing also showed low rates of prescribing for NHS Tayside across all years. NHS Shetland, Orkney, Borders and Dumfries and Galloway all showed an increase in co-amoxiclav prescribing between 2016 and 2018, however, NHS Western Isles and Forth Valley both showed a decrease in prescribing. Moran's I statistic was inconsistent between years with 2016 showing a positive spatial correlation statistic, 2017 showing no spatial correlation and 2018 showing a negative correlation coefficient however these were not supported by p-values ($I_{2016} = 0.185, p = 0.155$; $I_{2017} = 0.037, p = 0.544$ and $I_{2018} = -0.114, p = 0.837$) (figure 3.8).

The rate of clindamycin prescribing was much lower in comparison the other high risk antibiotic groups. NHS Orkney consistently showed the lowest prescribing rate of < 1 items per 100,000 patients per day, however, most NHS health boards showed an increased clindamycin prescribing rate since 2016, with only NHS Lothian and NHS Orkney showing a consistent decrease in prescribing. Moran's I, again, showed inconsistent spatial correlation coefficients and high p-values (figure 3.9).

Finally, all four high risk antibiotic groups were summed together to present a measure of the "4C" prescribing rates across NHS Scotland health boards. NHS Tayside was consistently the lowest prescriber of high risk antibiotic groups. NHS Highlands, Borders and Ayrshire and Arran were the highest prescribers of 4C antibiotic groups each year. The trend of prescribing was not consistent across all health boards with some showing signs of a decrease in 4C prescribing between years and others showing increased prescribing. Moran's I test for spatial autocorrelation showed no evidence of positive or negative spatial dependence between health boards (figure 3.10).

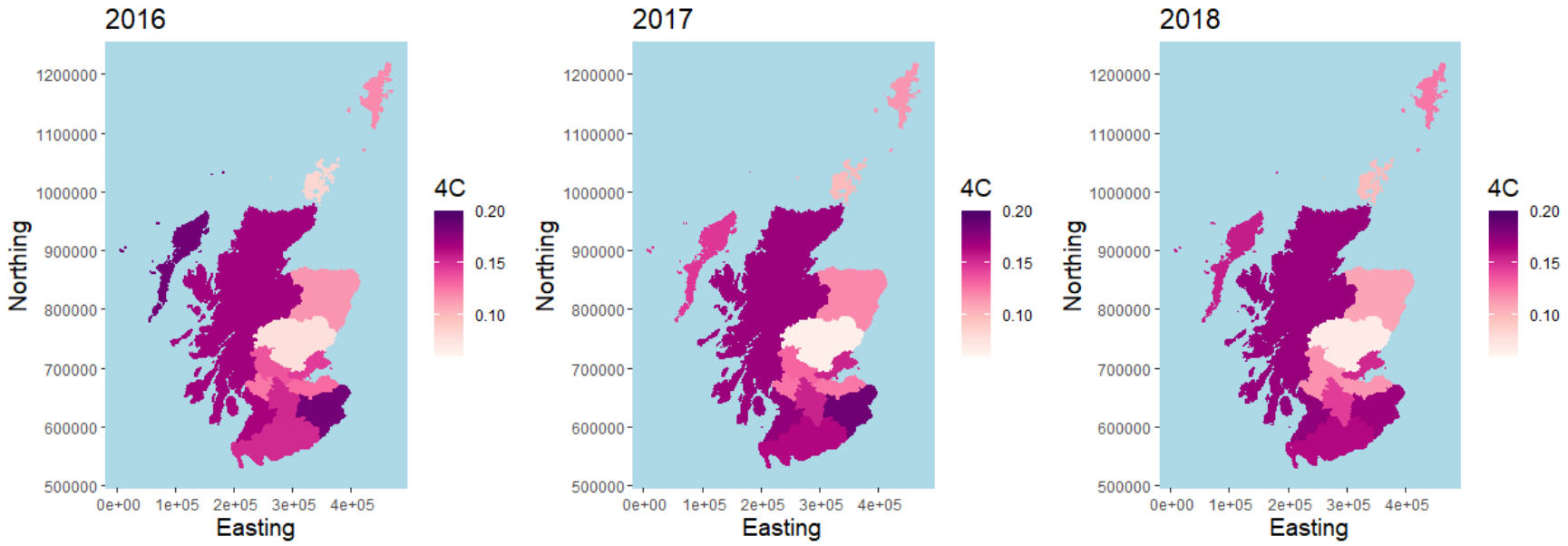


Figure 3.10: 4C prescribing rate (4C - items/1000/day) for 2016 (left), 2017 (middle) and 2018 (right) by NHS Scottish health boards.

Although total antibiotic trends show an overall decrease in prescribing, these maps showed that this was not the case for specific antibiotic groups and highlighted the variation in prescribing between health boards. These results also suggest that overall health board prescribing may be similar to neighbouring health boards, however, the prescribing rates of high risk antibiotics differ between health boards.

GP Practice Antibiotic Prescribing Point Maps and Monte Carlo Test for Spatial Association

This section presents GP antibiotic prescribing rates that were categorised by colour (red, orange, green and yellow) dependent on whether the GP practice met a total antibiotic prescribing target of ≤ 1.9 items per 1000 registered patients per day. Colour codings are defined in figure 3.11:

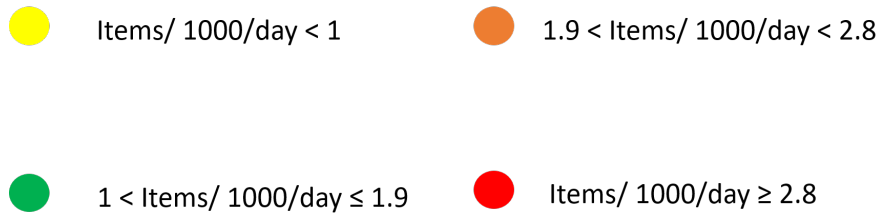
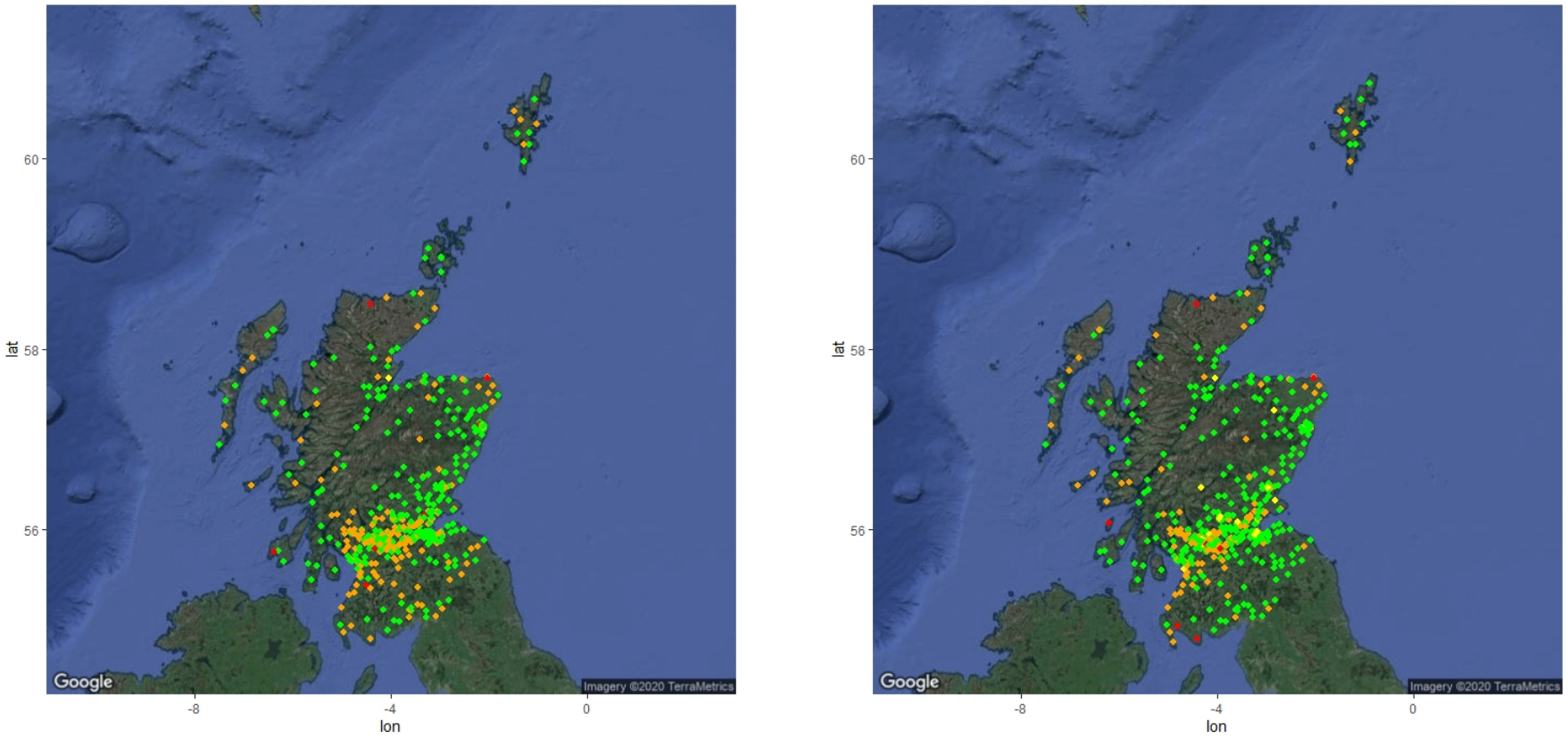


Figure 3.11: Key for antibiotic prescribing (items/1000/day) point maps.

Across Scotland, approximately 60% of GP practices met the target of ≤ 1.90 item/1000/day in 2016 and 74% of GP practices met this in 2018. Comparing 2016 to 2018, there was a visual increase in GP practices changing from *orange* to *green*, particularly in central belt of Scotland and close to the Scottish border. However, there were also a few GP practices in NHS Dumfries and Galloway moving from *orange* to *red* prescribing rates (figure 3.12).



(a) 2016 **(b)** 2018

Figure 3.12: Point maps of Scottish GP practice antibiotic prescribing rates for 2016 (left) and 2018 (right), categorised by definition described in figure 3.11.

The Monte Carlo test for spatial association showed no evidence of spatial auto-correlation between crude GP practice prescribing rates for 2016, 2017 and 2018 as all points on the sample variograms remained within the Monte Carlo Envelopes (figure 3.13).

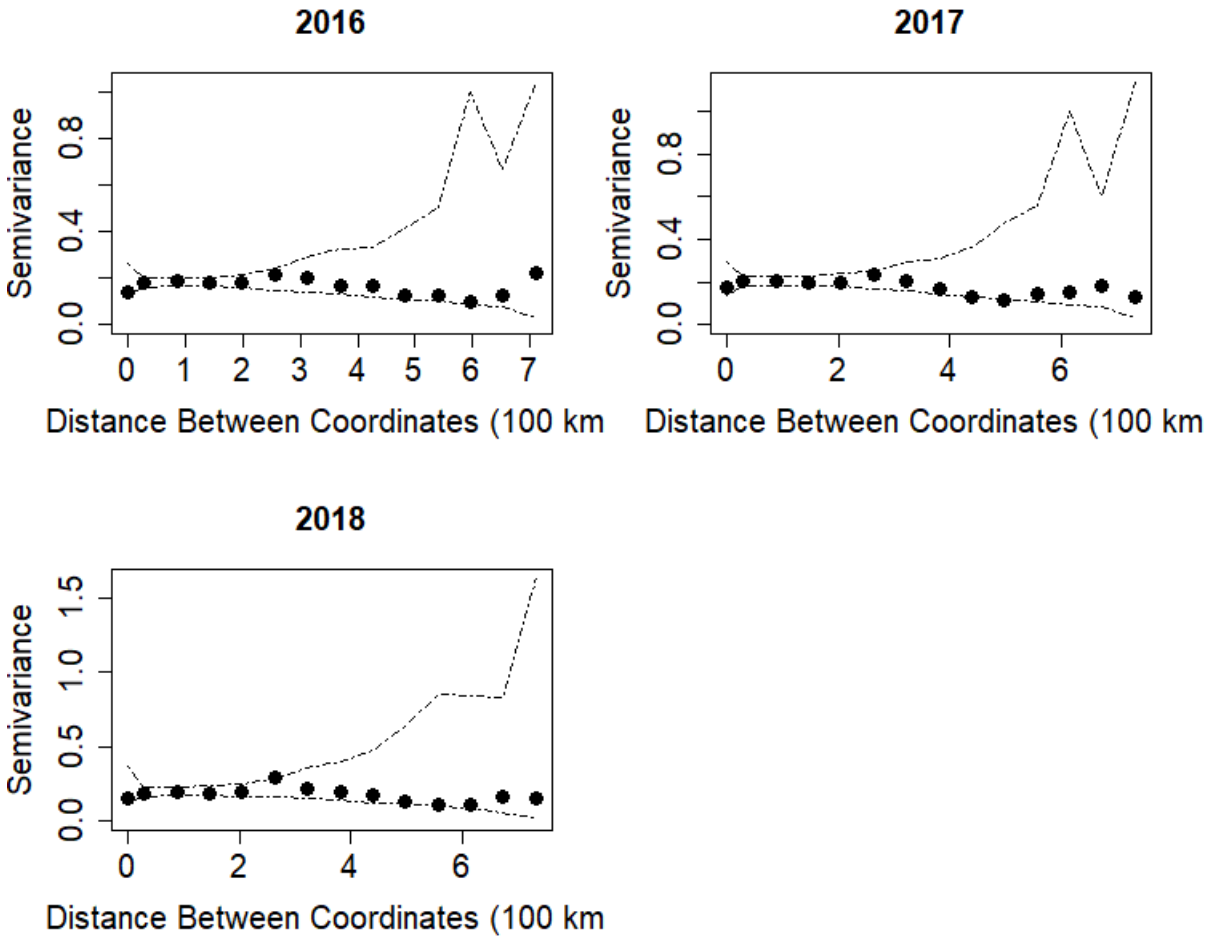


Figure 3.13: Sample variogram and Monte Carlo Envelopes for raw total antibiotic prescribing rates (items/1000/day) for 2016 (top-left), 2017 (top-right) and 2018 (bottom-left).

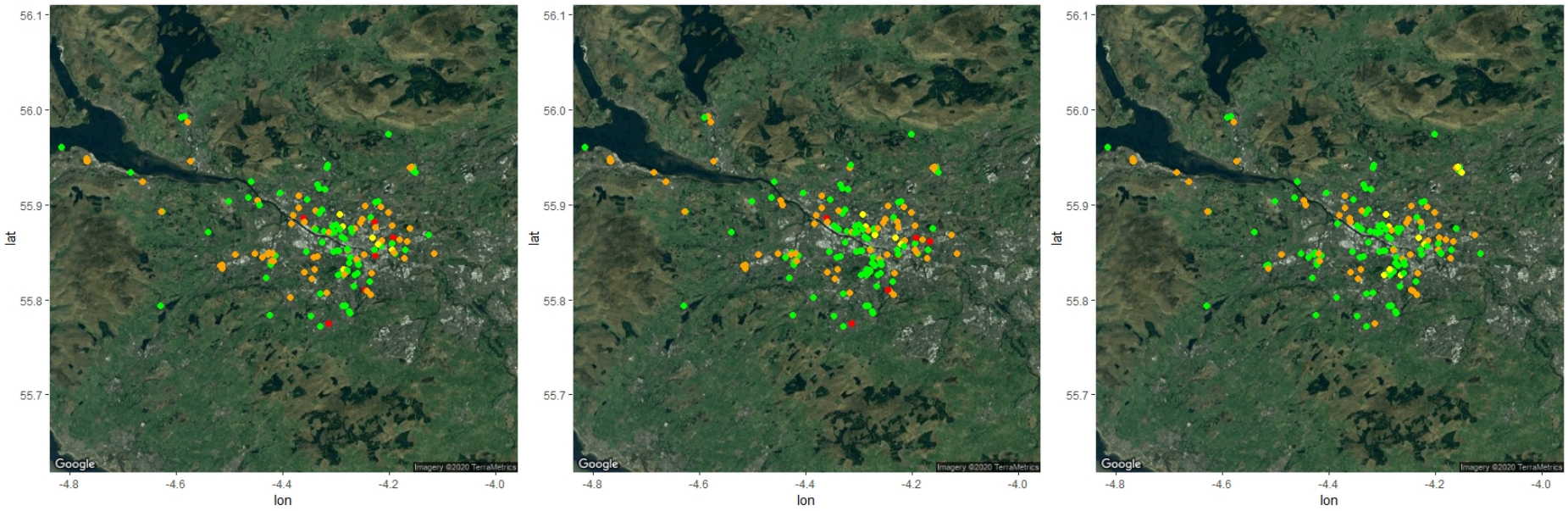
NHS Greater Glasgow and Clyde

There was a change in prescribing from 2016 to 2018 in Glasgow with a reduction of *red* practices in 2018 compared to 2016 and 2017. Multiple practices changed from *orange* to *green* and multiple *yellow* practices appeared in 2018, indicating individual practice reduction in prescribing. Approximately 49% of GP practices met the target in 2016, with 11 GP practices which prescribed $\geq$ 2.80 items/1000/day. In 2018, approximately 68% met the target with 4 GP practices who prescribed $\geq$ 2.80 items/1000/day. The overall distribution of prescribing was higher for GP practices in the north-east and south-west of Glasgow, whereas central Glasgow appeared to maintain lower prescribing rates.

NHS Lothian

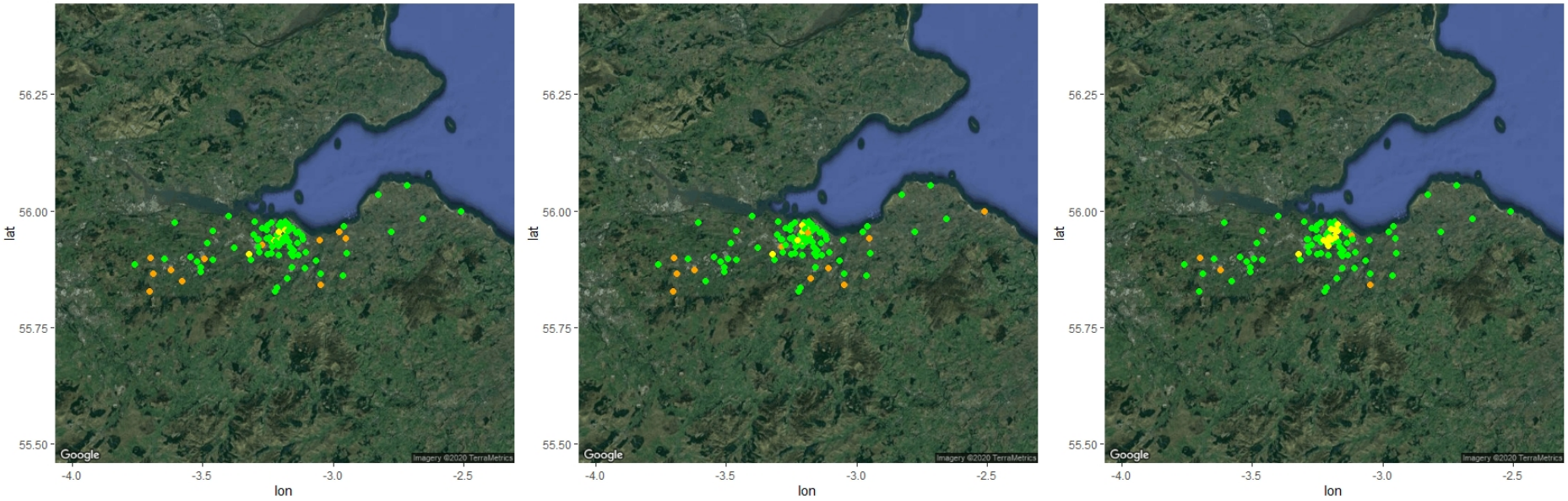
NHS Lothian GP practices predominantly prescribed $\leq$ 1.9 items/1000/day, especially in central Edinburgh where the number of practices prescribing $<$ 1.00 item/1000/day grew over three years. In the west of the region, towards west Lothian, there were a few *orange* practices that became green practices. Higher prescribing GP practices remained on the outskirts of NHS Lothian throughout the time period, with none in the city of Edinburgh itself. Approximately 22% of NHS Lothian GP practices prescribed $\geq$ 1.9 items/1000/day in 2016 and this reduced to 11% by 2018 (figure 3.15).

NHS Greater Glasgow and Clyde showed greater variability between GP Practices compared to NHS Lothian. Fewer practices met the target in NHS Greater Glasgow and Clyde compared to Lothian, however, as mentioned above, the target was an NHS Lothian target and is not necessarily comparable.



(a) 2016 **(b)** 2017 **(c)** 2018

Figure 3.14: Point maps of NHS Great Glasgow & Clyde GP practice antibiotic prescribing rates for 2016 (left), 2017 (middle) and 2018 (right), categorised by definition described in figure 3.11.



(a) 2016 **(b)** 2017 **(c)** 2018

Figure 3.15: Point maps of NHS Lothian GP practice antibiotic prescribing rates for 2016 (left), 2017 (middle) and 2018 (right), categorised by definition described in figure 3.11.

3.3.2 Principal Component Analysis

Antibiotic groups were compared to one another to assess correlations between the level of prescribing across the antibiotic groups. Penicillin prescription levels were strongly positively correlated to macrolide ($\rho = 0.68$), sulfonamide and trimethoprim ($\rho = 0.55$), Tetracycline ($\rho = 0.52$) and quinolone ($\rho = 0.47$) prescriptions. Quinolones prescriptions were also positively correlated with macrolide prescriptions.

Principal Component (PC) 1 accounted for 28% of the total variability within the data and approximately 70% of the total variance was explained by 6 PCs (figure 3.16).

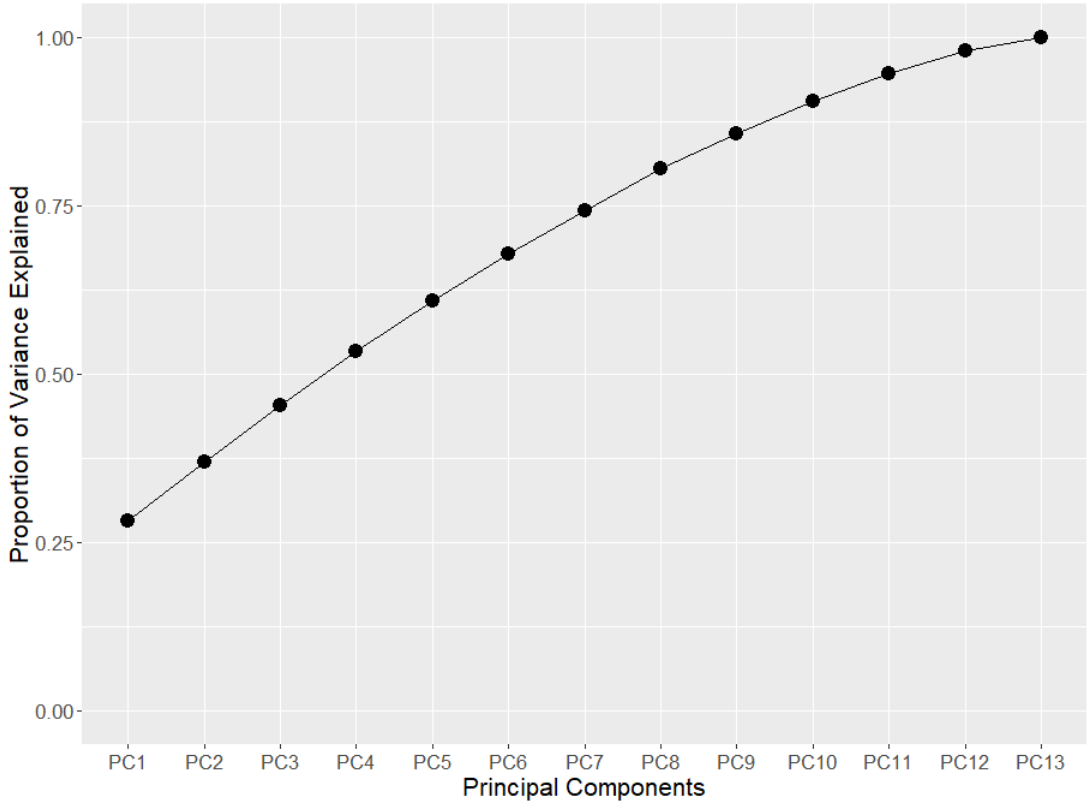


Figure 3.16: Cumulative proportion of variance explained by principal components.

The Rotation Matrix describes the linear combinations of the antibiotic groups. PC1 showed all positive loadings which were also of a similar value (about 0.2-0.4), with the exception of the anti-tuberculosis drugs and aminoglycosides. This implied that

high and low prescribing by GP practices could be described by a weighted average of all antibiotic groups which, in turn, implied that total antibiotics was a reasonable measure of GP practice prescribing behaviour. Penicillin showed the highest loading and accounted for the largest proportion of the antibiotics prescribed by GP practices (table 3.5).

PC2 showed a linear combination of antileprotic drugs, clindamycin and lincomycin, and metronidazole, tinidazole and ornidazole with large positive loadings (≥ 0.2) compared to aminoglycosides, cephalosporins & beta-lactams and quinolones with large negative loadings ($\leq$ -0.2). This implied that a GP practice which prescribed highly in positively loaded antibiotic groups tended to prescribed lower in the negatively loaded group, although PC2 only accounted for approximately 10% of variation in prescribing (table 3.5).

The biplot summarises the antibiotic groups' influence for PC1 and PC2, corresponding to the rotation matrix is shown in table 3.5. All PC plots focus on PC1 and PC2 which cumulatively explain approximately 40% of the total variability. GP practices that fell to the right-hand side of the x-axis represented *higher prescribers* and left-hand side point represented *lower prescribers* across all antibiotic groups (figure 3.17).

A cluster of red points were positioned closer to the negative scores of PC1 for percentage of practice population aged under 15 and over 74 (figure 3.18). These represented GP practices with a low percentage of under 15s and a low percentage of over 74s, suggesting that a low proportion of these age groups may be associated with lower GP practice total antibiotic prescribing rates (figure 3.18).

Table 3.5: Rotation matrix of every antibiotic class with principal components 1 to 7.

Antibiotic Classes	**Principal Components 1-7**						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Aminoglycosides	0.047	-0.229	0.547	0.4176	-0.352	0.532	-0.0099
Antileprotic Drugs	0.108	0.254	0.479	-0.1172	0.625	0.067	-0.5158
Antituberculosis Drugs	0.055	0.033	-0.287	0.8171	0.263	-0.169	-0.0573
Cephalosporins and other Beta-Lactams	0.219	-0.546	-0.171	-0.1178	-0.011	0.179	-0.3462
Clindamycin and Lincomycin	0.134	0.465	-0.300	0.0136	-0.497	0.050	-0.6045
Macrolides	0.375	-0.192	-0.073	0.0032	0.037	-0.245	-0.0038
Metronidazole, Tinidazole & Ornidazole	0.286	0.432	0.014	0.2304	0.061	0.256	0.2125
Penicillins	0.438	-0.058	-0.021	0.0259	0.115	-0.11	0.1309
Quinolones	0.354	-0.26	-0.093	0.0246	0.004	0.07	-0.2483
Some Other Antibacterials	0.142	0.006	0.504	0.1073	-0.345	-0.693	-0.0737
Sulfonamides And Trimethoprim	0.36	-0.059	-0.021	-0.0404	0.042	-0.015	0.1392
Tetracyclines	0.367	0.182	0.029	-0.1473	0.02	0.124	0.2449
Urinary-Tract Infections	0.303	0.193	0.024	-0.2025	-0.175	0.103	0.191

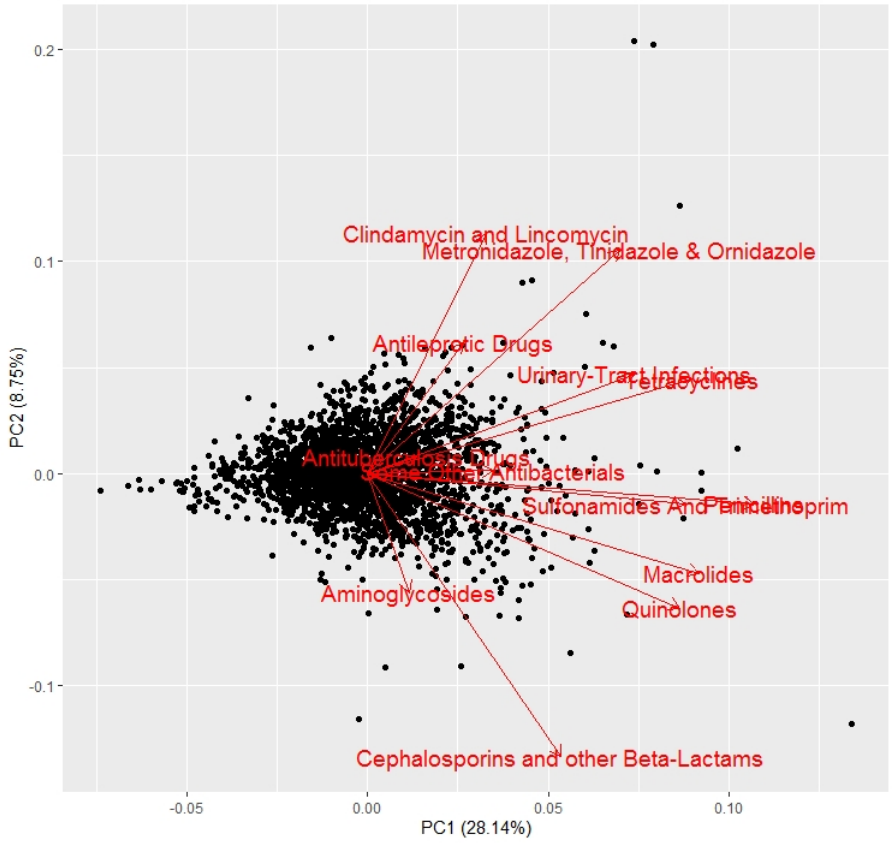


Figure 3.17: Biplot for principal component 1 vs principal component 2 showing direction of each antibiotic group influence.

GP practices with a higher percentage of patients in most deprived SIMD category were seen to cluster towards positive scores of PC1, therefore indicating higher total antibiotic prescribing. Dispensing GP practices appeared to mainly have positive scores on PC1 which suggested that GP practices that are dispensing practices showed higher total antibiotic prescribing (figure 3.19).

The results from the PCA showed that total antibiotic prescribing rate by GP practice to be a reasonable representation of overall prescribing by GP practices in Scotland, and also indicated an association with GP demographic factors including: higher prescribing for increased percentage of GP practice populations aged under 15, over 74 and living in the most deprived quintile. Dispensing GP practices also tended towards higher prescribing on along PC1 compared to the no dispensing GP practices.

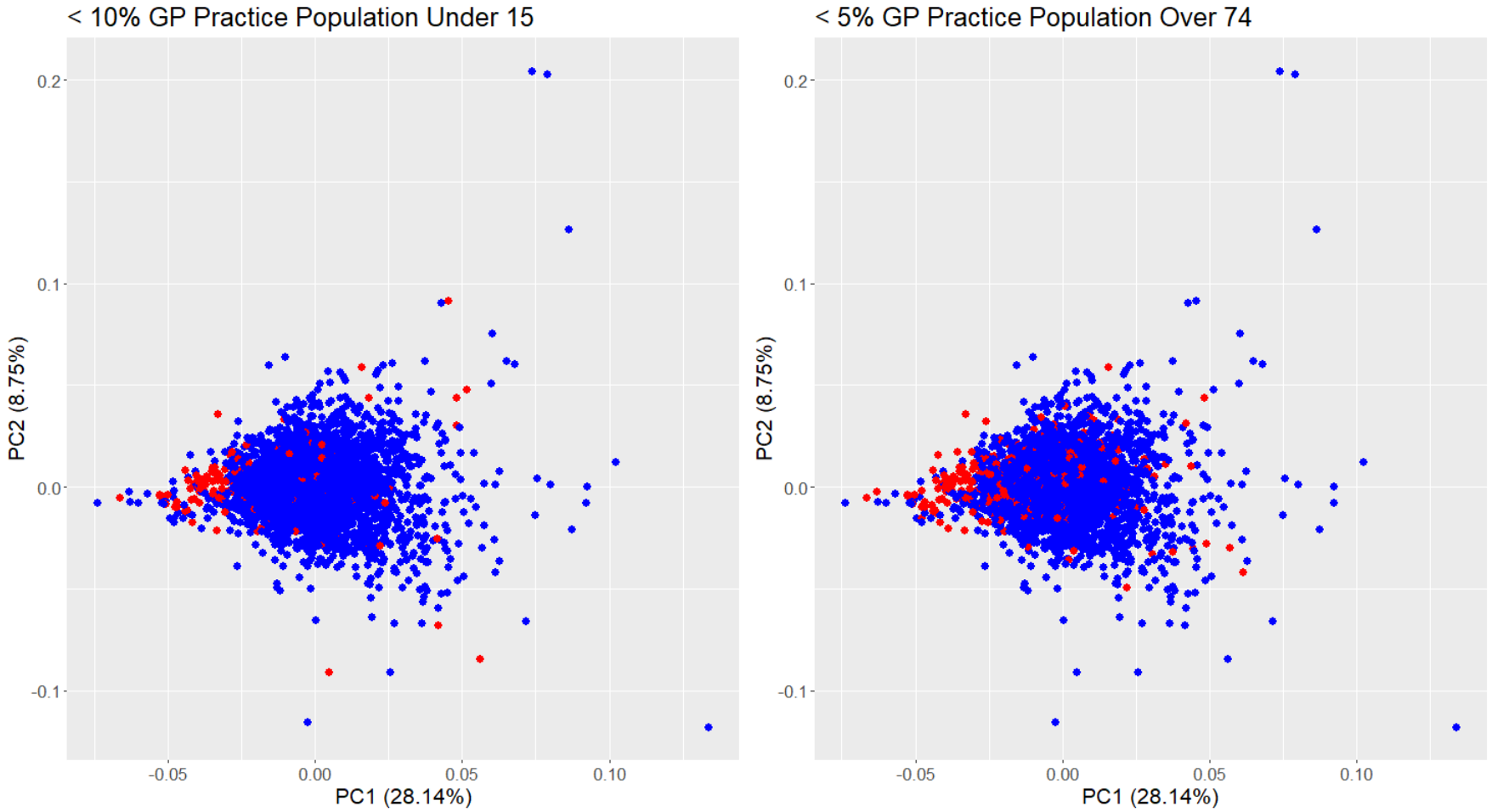


Figure 3.18: Scatter plots of the GP practices on the first two principal components (PC1, PC2). The percentage of GP practice aged under 15 and over 74-year-old are presented on the left and right graphs, respectively. Red represents GP practices with $<$ 10% population under 15 and $<$ 5% population over 74, with blue for those who were $>$ 10% and 5%. Age cut-offs were determined by 1st quartile or exploratory purposes.

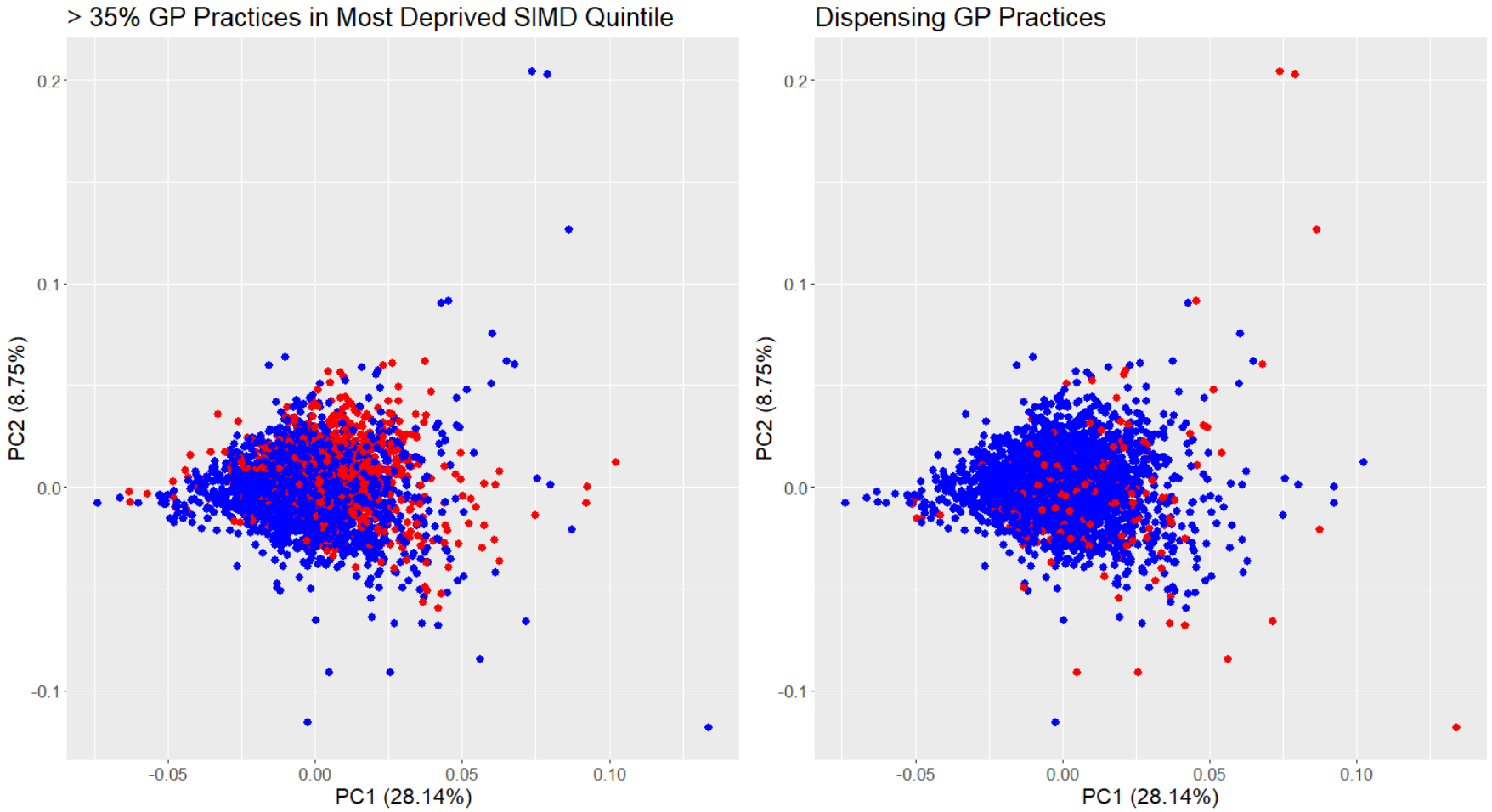


Figure 3.19: Scatter plots of the GP practices on the first two principal components (PC1, PC2). The percentage of GP practice residing in most deprived SIMD quintile and dispensing GP practices are presented in the left and right graphs, respectively. Red represents GP practices with $> 35\%$ population most deprived and dispensing GP practices, with blue for those less than 35% and non-dispensing. Most deprived SIMD cut-off determined by 3rd quartile.

3.3.3 Generalised Linear Model

A positive association was seen between the increasing percentage of the practice population aged under 15 and over 74 with total antibiotic prescribing rates. The percentage of the practice population residing in the most deprived SIMD quintile showed a positive increasing association with the total antibiotic prescribing rate, whereas the percentage of the practice in least deprived areas showed a negative association with total antibiotic prescribing (figure 3.20).

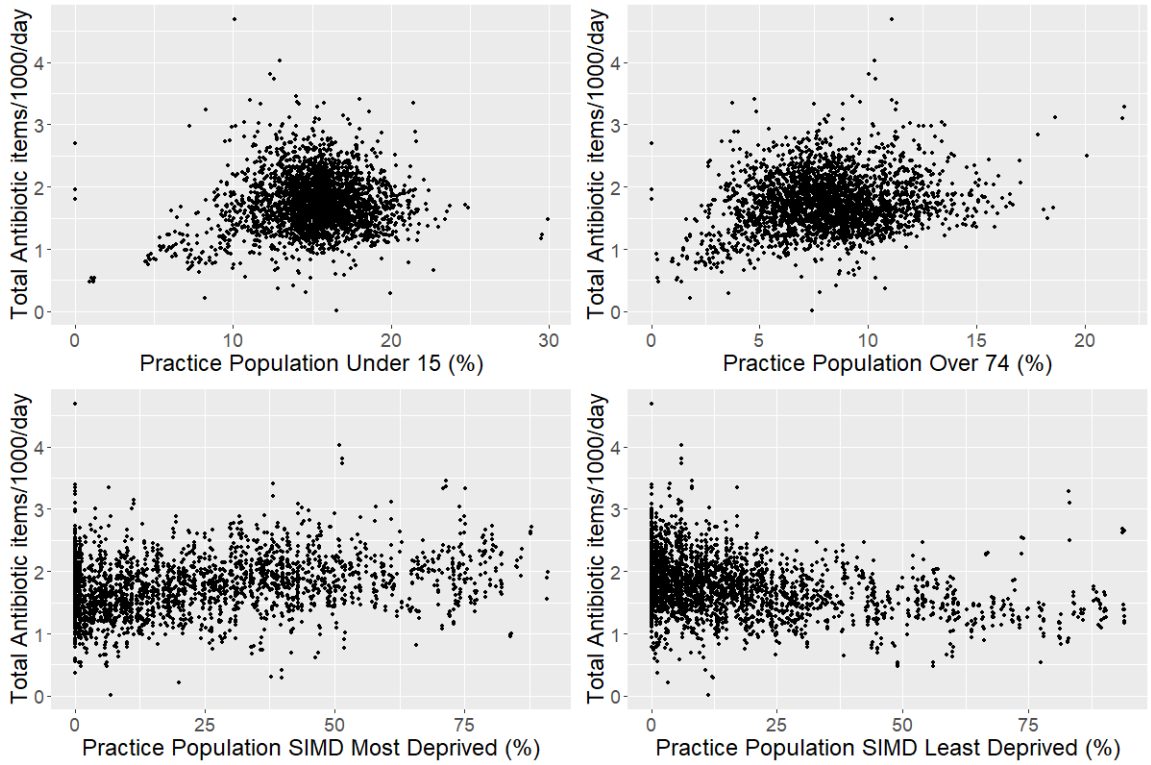


Figure 3.20: Scatter plots of total GP antibiotic prescribing rates (items/1000/day) compared to percentage of practise population aged under 15 years (%) (top-left), over 74 years(%) (top-right), residing in the most deprived quintile (%) (bottom-left) and least deprived quintile (%) (bottom-right)

The median total antibiotic prescribing was lower for non-dispensing practices compared the dispensing practices. The median rate of total antibiotic prescribing progressively decreased from 2016 to 2018 (figure 3.21).

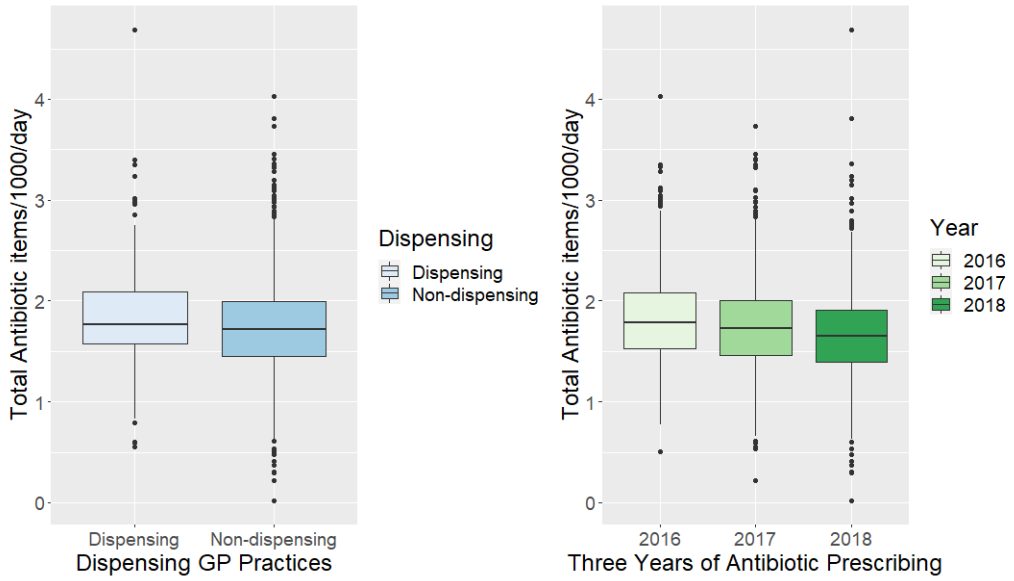


Figure 3.21: Boxplots of total antibiotic prescribing rates (items/1000/day) compared to dispensing/non-dispensing GP practices (left) and year (2016 to 2018) (right).

Total Antibiotic Prescribing GLM

The unadjusted analysis showed a positive association between total antibiotic rate and percentage of under 15s (RR = 1.013, 95% CI 1.009 - 1.016, per 1% increase), percentage of over 74s (RR = 1.020, 95% CI 1.017 - 1.024, per 1% increase) and percentage of patients residing in the most deprived SIMD quintile (RR = 1.003, 95% CI 1.002 - 1.004), whereas high populations residing in the least deprived quintile showed a negative association the with rate of total antibiotic prescribing (RR = 0.996, 95% CI 0.995 - 0.996). There was an increase in prescribing for dispensing practices in comparison to non-dispensing practices (RR = 1.068, 95% CI 1.033 - 1.105), 2017 and 2018 had a decreased rate of total antibiotic prescribing in comparison to 2016 (RR_{2017} = 0.964, 95% CI 0.941 - 0.988 and RR_{2018} = 0.916, 95% CI 0.894 - 0.938). There were also differences between NHS Scotland health boards (p<0.001) (table 3.6).

Table 3.6: Unadjusted and adjusted analyses of total antibiotic prescribing. Negative-binomial GLM with risk ratios and 95% confidence intervals.

	Unadjusted	Adjusted	
	RR (95% CI)	RR (95% CI)	P
Year: 2016	**1**	**1**	**-**
2017	0.964 (0.941 - 0.988)	0.965 (0.946 - 0.984)	<0.001
2018	0.916 (0.894 - 0.938)	0.914 (0.895 - 0.932)	<0.001
Non-dispensing	1	1	-
Dispensing	1.068 (1.033 - 1.105)	1.077 (1.041 - 1.115)	<0.001
NHS Scotland Health Boards	-	-	<0.001
% Practice Population Under 15	1.013 (1.009 - 1.016)	1.015 (1.012 - 1.018)	<0.001
% Practice Population Over 74	1.020 (1.017 - 1.024)	1.037 (1.033 - 1.041)	<0.001
% Practice Population Least Deprived	0.996 (0.995 - 0.996)	0.997 (0.997 - 0.998)	<0.001
% Practice Population Most Deprived	1.003 (1.003 - 1.004)	1.003 (1.002 - 1.004)	<0.001

The global p-value was obtained for the NHS Scotland health boards. Individual comparisons between health boards to a baseline health board did show differences, however these were not displayed to maintain clarity.

The fully adjusted model showed similar estimates to the unadjusted analysis, with total antibiotics showing a positive association with percentage of practice population aged under 15 (%), over 74 (%), residing in the most deprived quintile (%) and a negative association with the least deprived quintile (%). There was an increase in total antibiotic prescribing for dispensing practices compared to non-dispensing practices and 2017 and 2018 were seen to have lower prescribing rates compared to 2016 (table 3.6).

Monte Carlo Test for Spatial Autocorrelation

Model residuals were assessed for spatial dependence separately for 2016, 2017 and 2018. After accounting for GP practice demographics, health boards and year, there was no evidence of spatial association between GP practices for 2016, 2017 or 2018. No points lay outwith the Monte Carlo Envelopes, which suggested there was no spatial dependence between GP practices. This also confirms the independence assumption.

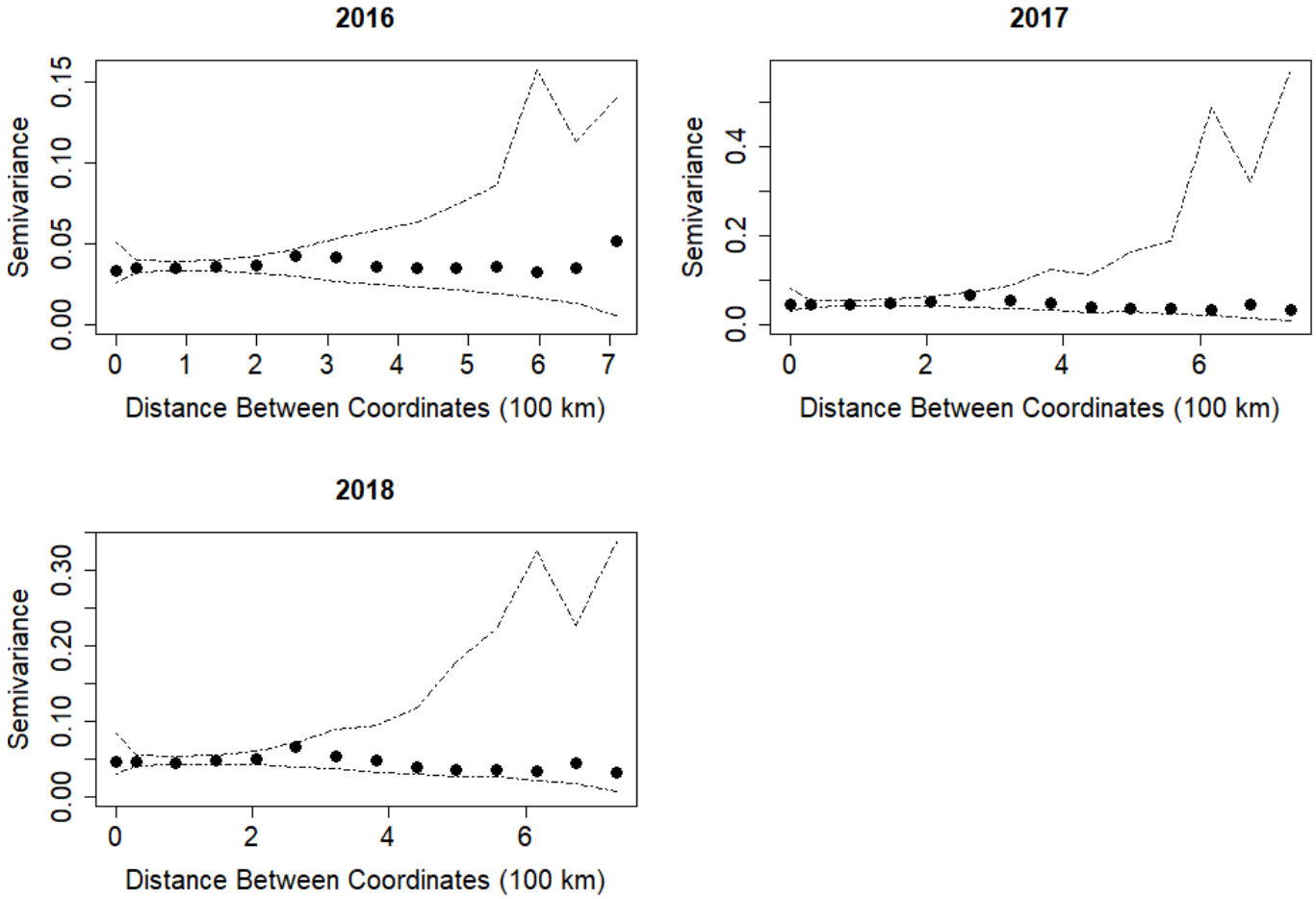


Figure 3.22: Sample variogram for 2016 (top-left), 2017 (top-right) and 2018 (bottom-left) with corresponding Monte Carlo envelopes of residuals from negative-binomial GLM in table 3.6.

GP Practice Influenza Vaccination Uptake

The association between total antibiotic prescribing and proportion of GP practice Influenza vaccination uptake in over 65s was assessed for 2016. The median percentage of GP practice Influenza vaccination in over 65s was 75.3% (IQR = 71.5% - 77.6%), with a minimum of 46.2% and maximum of 92.0%.

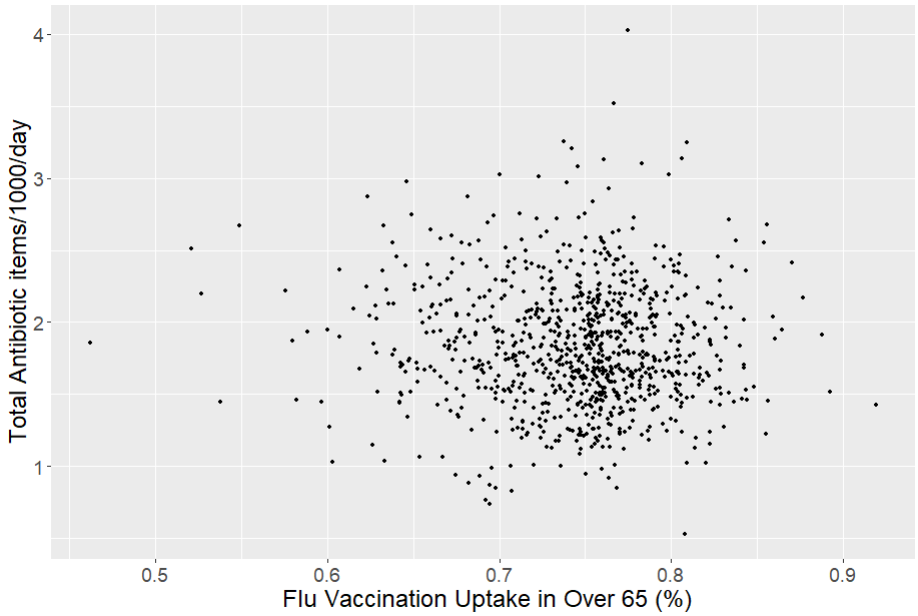


Figure 3.23: Scatter plot of total antibiotic prescribing rate (items/1000/day) compared to GP practice influenza vaccination uptake in registered patients $\geq$ 65 (%).

There was no clear direction of association between total antibiotic prescribing rates and influenza vaccination uptake in over 65's seen in figure 3.23. An unadjusted analysis showed no strong evidence of an association between total antibiotic prescribing and influenza vaccination in over 65s (%), with the confidence interval containing 1 (RR = 0.998, 95% CI 0.995 - 1.001). Influenza vaccination uptake (%) was then presented in a fully adjusted, with the same covariates presented in table 3.7.

An adjusted analysis with the same covariates as table 3.7, showed influenza vaccination uptake to have a weak negative association with GP practice total antibiotic prescribing, however, confidence interval contained 1 (RR = 0.999, 95% CI 0.997 - 1.002). An

Table 3.7: Adjusted analyses of total antibiotic prescribing with influenza vaccination uptake in ≥ 65-years-old (%). Negative-binomial GLM risk ratios and 95% confidence intervals.

	Adjusted	
	RR (95% CI)	P
% Influenza Vaccination Uptake in Over 65 (%)	0.999 (0.997 - 1.002)	0.675
Non-dispensing	1	-
Dispensing	1.071 (1.016 - 1.130)	0.011
NHS Scotland Health Boards	-	<0.001
% Practice Population Under 15	1.018 (1.013 - 1.023)	<0.001
% Practice Population Over 74	1.038 (1.032 - 1.044)	<0.01
% Practice Population Least Deprived	0.997 (0.996 - 0.998)	0.001
% Practice Population Most Deprived	1.002 (1.002 - 1.003)	<0.001

Global p-value was obtained for NHS Scotland health boards. Individual comparisons between health boards showed between health board differences however were not meaningful to these analyses.

positive with GP antibiotic prescribing was seen for dispensing GP practices compared to non-dispensing practices (RR = 1.071, 95% CI 1.016 - 1.130), increased percentage of under 15s (RR = 1.018, 95% CI 1.013 - 1.023), aged over 74s (RR = 1.038, 95% CI 1.032 - 1.044), and most deprived percentage (RR = 1.002, 95% CI 1.002 - 1.003) with a negative association seen for the least deprived perentage (RR = 0.997, 95% 0.996 - 0.998). GP practice demographic estimates remained similar in magnitude and direction to previously presented in table 3.7 which compared three years of GP antibiotic prescribing data, although there was a slightly stronger association with increased prescribing for high percentage of practice population ages over 74 in 2016.

3.4 Discussion

The aim of the analyses in this chapter were to explore the variation between GP practice antibiotic prescribing rates in Scotland with use of spatial statistics and mapping techniques. The data analysed were from 2016 to 2018, and the influence of GP practice demographic information on antibiotic prescribing rates was investigated.

Total antibiotic prescribing rates (items/1000/day) showed a 7.3% reduction between 2016 and 2018, which was comparable to the Scottish One Health Antimicrobial Use and Antimicrobial Resistance 2018 report which showed a 6.2% reduction in human antibiotic use, and a 10.2% reduction in primary care antibiotic use, in comparison to 2014 [131]. Most NHS Scotland health boards showed a decreasing trend for total antibiotic prescribing, however, high-risk antibiotic groups varied between years, with clindamycin showing an increase in prescribing for most health boards over the three years.

The use of total GP antibiotic prescribing to explore variation in prescribing appears to be standard based on other studies [150, 136, 151]. The PCA in this chapter confirmed this as an appropriate measure by showing that 28% of GP antibiotic prescribing variation could be explained by a weighted average across all antibiotic drug groups. However, this leaves 72% of the variation over practices unexplained which provides a reason for the detailed investigation of individual antibiotic groups.

A fully adjusted analysis showed a positive association between total antibiotic prescribing with the percentage of practice population aged under 15 and over 74, such that a 5 percentage-point increase in either was associated with a 7% and 10% increase in antibiotic prescribing, respectively. An increase in the percentage of the practice population residing in the most deprived SIMD quintile was also associated with an increase in total antibiotic prescribing, whereas practices that had a higher percentage of patients who were among the least deprived tended to have lower antibiotic prescribing rates. Total antibiotic prescribing was 1.07 (95% CI = 1.03 - 1.11) times higher for dispensing GP practices in comparison to non-dispensing GP practices. There was evidence of a positive spatial association between antibiotic prescribing rates at a health board level for these data in 2016, 2017 and 2018: Moran's $I_{2016} = 0.18$ (p = 0.1598); Moran's $I_{2017} = 0.393$ (p = 0.010) and $I_{2018} = 0.259$ (p = 0.07), however this was not

seen at an individual GP practice level, with no evidence of spatial association after adjusting for the GP demographic factors. A subset of the analysis for 2016 only, showed no evidence to suggest that total antibiotic prescribing was associated with uptake of influenza vaccinations in registered patients over 65.

A study in Italy commented on the scarcity of spatial studies that report on the distribution of antibiotic use and highlighted the necessity given that antibiotics are commonly prescribed to cure infections that may spread in the community. This study found evidence of spatial dependence between outpatient antibiotic use in Italian regions ($N = 20$), reporting a Moran's I = 0.797 ($p < 0.001$) and a positive association between antibiotic use with income elasticity (economic measure of how responsive the quantity demand for a good or service is to a change in income) and age structure of the population [152], which show similarities with the results in this chapter. Another study of the spatial patterns of GP antibiotic prescribing in England in 2016 showed high and low spatial clusters of antibiotic prescribing, with hot spots of high prescribing predominately in the North of England. These differences were mainly attributed to deprivation factors, particularly income, employment, education and health [150], which again agrees with the results in the chapter. There are few spatial studies that report on spatial variation in antibiotic prescribing [150, 152, 153, 154], with only one study specifically exploring GP antibiotic prescribing spatially [150]. It is understood that this chapter is the first to explore spatial variability of GP antibiotic rates in Scotland.

Previous studies that have explored antibiotic variation at an ecological-level largely report the variation in prescribing associated with deprivation and socioeconomic factors [135, 136, 150, 151]. The results in these studies are comparable with the results from the analysis in this chapter that show an increase in total antibiotic prescribing for Scottish GP practices associated with an increase in the percentage of the practice population residing in most deprived SIMD quintile (RR = 1.003, 95% CI 1.002 - 1.004)

per 1% increase and a decrease associated with least deprived SIMD quintile (%) (RR = 0.997, 95% CI 0.997 - 0.998) per 1% increase. This translates as an approximate 11% increase in total antibiotic prescribing between the 1st and 3rd quartile for the percentage of most deprived practice population (1% - 36%). Patient age has also been highlighted as a contributing factor to GP antibiotic use, with the general trend showing an increase in the use of antibiotics with increasing age, however, younger patients (0 - 9 years old) also show high levels of antibiotic use compared with other age groups at an individual-level [135, 151]. The results from this chapter were, therefore, comparable with individual level study findings of GP antibiotic prescribing, with a positive association between increased GP antibiotic prescribing and increased percentage of practice population aged over 74 and under 15.

A study in Switzerland that found that dispensing GP practices prescribed more antibiotics when compared to non-dispensing practices, after adjusting to spatial demographic differences [155] which agrees with the results in this chapter. The study did not offer reasons for the result and stated that the question of whether dispensing units contributed to the overuse of antibiotics was unclear, however, a study in England that indicated dispensing GP practices prescribed more expensive drugs in comparison to non-dispensing practices [133], highlighting a financial conflict in prescribing which may provide a possible explanation for the difference between dispensing and non-dispensing GP practices. Future work should consider investigating the relationship between the cumulative cost of total antibiotics and the difference between dispensing and non-dispensing GP practices.

An association between higher influenza vaccination uptake and lower antibiotic prescribing rates has been previously reported [138, 139] but has not been previously modelled in Scotland. The analyses in this chapter showed no evidence of an association between total antibiotic prescribing and uptake of influenza vaccines in registered

patients $\geq$ 65 years old. An association has been previously reported in the United Kingdom for patients $\geq$ 65, however, this was at an individual patient-level study which was conducted at the beginning of the millennium (2000) [138]. An association has also been reported at an ecological-level in Ontario, Canada, however, these analyses included vaccinations for the entire population opposed to older age only and was also conducted on data for 1997 - 2000. Therefore, a relationship with elderly vaccinations only may be more difficult to see at an ecological-level and overall antibiotic consumption was higher in the past which may provide a reason for these differences. Another potential explanation is that the earlier studies in the UK were carried out before the universal offer of influenza vaccinations to all over 65, so vaccine rates were much lower and may have been more variable over GP practices. In Scotland uptake of influenza vaccination among over 65's is high, with most GP practices falling between 70% and 80% uptake in 2016, with all $>$ 40%.

The data obtained for these analyses were primarily collected from open source GP practice prescribing data, which were linked to GP demographic information. It is a noted limitation that not all GP practices could be located for this study, with some GP practices' removed due to inconsistent list sizes. The biases of comparing Scottish GP antibiotic rates to a NHS Lothian only target is also noted as a limitation, understanding that the demographic differences between these locations was not accounted for. Any comparisons made were purely for exploratory purposes and should not be used to draw formal conclusions about total antibiotic prescribing between these health boards. However, this study does highlight the capabilities of open source data and present simplistic visuals for comparing GP practices against antibiotic targets which may be beneficial for feedback purposes. These analyses highlight the potential to explore important public health topics easily and inexpensively.

A spatial study of Scottish GP practice respiratory prescriptions, for 2015 and 2016, published in 2018, provided useful comparisons as it closely followed a similar data linkage process to the analyses presented in this chapter [156]. They identified 939 GP practices in 2016, however, did not exclude Open Access GP practices for homeless or University practices which may account for the differences in numbers reported in this chapter. They did, however, report the exact same loss of 27 GP practices due to missing practice list sizes (section 3.2.1). In the analyses reported in this chapter, GP antibiotic prescribing data were geocoded based on their postcode and treated as point-location spatial data. Classing GP prescribing data as this spatial data type is a standard approach [150, 157]. However, the study of respiratory GP prescriptions introduces a conversation regarding the classification of GP practice data as a spatial type. It highlights that GP practice data are not strictly point data as its patient population is related to surrounding areas, and conversely, spatial closeness of GP practices with overlapping practice population implies GP practices are not strictly areal data either. This study presents a novel spatial adjacency framework which may be considered for future work. It would be interesting to compare the results from these analyses using the standard approach to results applying this novel neighbourhood structure [156].

The results in this chapter have showcased the successful results of antibiotic stewardship efforts and highlighted the continuing need for stringent measures in reducing GP practice antibiotic prescribing, particularly in the use of high-risk antibiotics. It has also shown the influence of GP practice demographics on antibiotic prescribing rates and supports the need for antibiotic prescribing targets adjusted for the deprivation profile of the area, to ensure fair comparisons in reductions of antibiotic prescribing.

Chapter 4

Identifying Ecological Risk Factors of *Clostridioides Difficile* Infection: Exploring Spatial and Temporal Effects of CDI in Scotland

4.1 Introduction

Clostridiodies difficile infection (CDI) is a bacterium that effects the human gut, with key symptoms including sickness and diarrhoea. Established individual-level risk factors of CDI include older age; comorbidities such as diabetes, chronic obstructive pulmonary disease (COPD) and exposure to Proton Pump Inhibitors (PPIs) [39, 44]. However, the most established risk factor of CDI is recent exposure to antibiotics, particularly broad spectrum antibiotics clindamycin, cephalosporins, fluroquinolones and co-amoxiclav [59, 58, 60].

Population-based studies have highlighted ecological risk factors of CDI such as population density, percentage of population over 65, deprivation and environmental factors such as proximity to livestock farming, proximity to farming raw materials and proximity to nursing homes [158, 53, 54]. There are a few population-based spatial and spatio-temporal studies of CDI [159, 158, 53, 160] based in Australia and, North Carolina but only one European study to date. These studies show varying levels of spatial clustering of CDI: a 10-year study in Queensland, Australia found no evidence of spatial variation in the proportion of CDI, whereas a study in Australian Capital Territory (ACT) found significant geographical variation, identifying areas at an elevated risk of infection and a positive association with neighbourhood socio-economic disadvantage. A study in the Netherlands found no evidence of spatial, temporal or spatio-temporal clustering of CDI cases and also reported no evidence of an association between CA-CDI and proximity to livestock [160].

Public Health Scotland (PHS) have reported a year on year decreasing trend in CDI incidence, with a 7.5% reduction in total CDI between 2014 and 2018 for patients 15 years and older [161]. The over-arching trend of CDI incidence suggests yearly reductions, however, community-acquired CDI (CA-CDI) has been highlighted as a growing public health concern and a potential burden on public healthcare services, with cases of CA-CDI doubling in the last 30 years [162]. Seasonal effects of CDI introduce temporal aspects of the infection. A systematic review of CDI seasonality showed that CDI had a similar seasonal pattern in northern and southern hemispheres, with a peak in spring followed by lower rates in summer/autumn months and an 8-month lag between the Hemispheres [163]. However, a spatio-temporal study of environmental factors of CDI in Queensland, Australia reported annual trends with peaks of CDI infections in summer months (December - February) [159]. Another study in the United States suggested a seasonal pattern of CDI (23% increase in winter months compared to summer)

and an association with influenza, hypothesized to be due to increased antimicrobial use during influenza seasons [164].

Notwithstanding the possibility of geographical difference in *Clostridiodies difficile (c-difficile)* ecology, these data suggest that spatial and temporal effects of CDI may be common. Knowledge of potential spatial and temporal structures allows for adequate modelling and better understanding about the risk factors for CDI which is essential for monitoring this infection.

This chapter utilises routinely collected quarterly CDI data by intermediate zones from 2014 to 2018 to identify ecological risk factors that are accessible on the same spatial scale and explore spatial and temporal trends of *c-difficile* infection in Scotland.

This chapter aims to address the following research questions:

1. How has CDI incidence changed over time in Scotland and is there any seasonal variation of CDI incidence?

2. Is there any spatial association between areal-level CDI incidence in Scotland?

3. Do routinely collected population-based data of risk factors for CDI have associations with areal-level incidence in Scotland?

4.2 Methods

This chapter summarises the analysis of a retrospective ecological study of routinely collected quarterly CDI cases in Scotland between 2014 and 2018. These data were aggregated by intermediate zone (IZ) and linked to open-source demographic information related to the Scottish population and area level characteristics.

4.2.1 Data

Clostridiodies Difficile Infections

Quarterly counts of laboratory and epidemiologically confirmed CDI between January 2014 and December 2018 in Scotland were linked by the unique patient identifier (CHI) to postcode of residence. CHI was then removed and the data aggregated by intermediate zone (IZ).

Quarters are defined as Quarter 1 January - March, Quarter 2 April - June, Quarter 3 July - September and Quarter 4 October - December. To adjust for age and sex differences within each IZ, the expected CDI counts were calculated using indirect standardisation whereby the age and sex adjusted CDI rates for the whole of Scotland were multiplied by the population breakdown in each age group for males and females in each IZ (figure 4.1 for data example). CDI cases were classified as either community-acquired CDI (CA-CDI) or hospital-acquired CDI (HA-CDI) as specified in the Protocol for the Scottish Surveillance Programme for Clostridium difficile infection [165]. CA-CDI include all cases without hospitalization in the previous 12 weeks and were either tested outside of hospital or tested within 48 hrs (2 days) of hospital admission. All other cases are defined as HA-CDI.

	Sex	Age.Gp	Pop	CDiff.Count	HCAI.Count	CA.Count	CDI.Rate	HCAI.CDI.Rate	CA.CDI.Rate
1	Female	15 - 19	725398	41	31	12	5.652070e-05	4.273516e-05	1.654264e-05
17	Male	15 - 19	760120	22	18	7	2.894280e-05	2.368047e-05	9.209072e-06
2	Female	20 - 24	902661	73	36	47	8.087200e-05	3.988208e-05	5.206827e-05
18	Male	20 - 24	904750	47	25	27	5.194805e-05	2.763194e-05	2.984250e-05

	Sex	Age.Gp	Year	IZ2011	Pop
1	Female	15 - 19	2014	S02001236	101
2	Male	15 - 19	2014	S02001236	149
3	Female	20 - 24	2014	S02001236	151
4	Male	20 - 24	2014	S02001236	116

	IntermediateZone2011Code	YearQuarter	CDI	CA	HCAI
1	S02001236	2014 Q1			
2	S02001241	2014 Q1			
3	S02001242	2014 Q1			
4	S02001244	2014 Q1			

Figure 4.1: Screenshots of the data provided for this study. Sex (Female/Male) and age group (15-19, 20-24, 25-29, ..., 80-84, 85-89, 90 Plus) CDI rates shown in top screen shot, then Intermediate Zone populations by Year in middle screenshot. CDI, CA-CDI and HA-CDI counts by Intermediate Zone, Year and Financial Quarter in bottom screenshot. Age-sex adjusted expected CDI counts were calculated for each year by multiplying CDI rates with age-sex-year populations.

Spatial Covariates

Open-sourced spatial covariate information were linked to the CDI data by IZ code. Covariates were chosen based on accessible open-source population-based factors highlighted in the literature from similar population-based studies [158, 53]. These included SIMD deprivation scores, population and age demographics which were sourced from information services division (ISD) national data catalogue and Scottish Index of Multiple Deprivation (SIMD) [143, 94] (see table 4.1 for covariate definitions).

Table 4.1: Covariate definitions of income deprived, employment deprived, forestry and fishery working population, population density, population aged over 64.

Covariate	Definition
Income Deprived (% IZ population)	Percentage of IZ working age population income deprived, as defined by the Scottish Index of Multiple Deprivation (SIMD), is a measure of the percentage of the population of adults and their dependents in receipt of Income Support, Employment and Support Allowance, Job Seekers Allowance, Guaranteed Pension Credits, Child and Working Tax Credits or Universal Credit (excluding those in the category 'working with no requirements'), or in Tax Credit families on low income.
Employment Deprived (% IZ population)	Percentage of IZ working age population employment deprived , as defined by the Scottish Index of Multiple Deprivation (SIMD), is a measure of the percentage of the working-age population (men aged 16-64 and women aged 16-60) who are on the claimant count, those who receive Incapacity Benefit, Employment and Support Allowance or Severe Disablement Allowance, and Universal Credit claimants who are not in employment.
Forestry and Fishery (% IZ Population)	2011 census information on agricultural working in Scotland. Percentage of working age population per IZ working in the Fishery and Forestry Industry.
Population Density per IZ	Population density is defined by the number of people per km-squared per IZ.
Aged over 64 (% IZ Population)	Percentage of IZ population aged over 64.

4.2.2 Statistical Methods

The data were initially explored using descriptive statistics and visualising barplots of the raw CDI counts by year and quarter. Temporal trends within the data were then assessed using Poisson GLMs of the crude CDI incidence aggregated by quarter and year. Spatial patterns within the CDI data were then explored, aggregating the data over time (2014 - 2018) by IZs and year, then model selection was performed to identify risk factors of CDI using a Poisson GLM. The data were then modelled using a CAR Leroux spatial model and adjusting for previously defined covariates. Finally, a spatio-temporal AR(1) was then assessed on the space-time CDI data, adjusting for same covariates with the addition of year as a fixed effect.

The spatial and spatio-temporal analyses modelled CDI Standardised Incidence Ratio (SIR). SIR is defined by the number of observed cases in an area i divided by the number of expected cases (E_i): a SIR > 1 implies more cases than expected in that geographical area and a SIR < 1 implies fewer cases than expected. The expected values calculated for these analyses were age-sex adjusted and calculated:

$$E_i = \sum_{strataj} r_j N_{ij} \tag{4.1}$$

for $j = (1, .., J)$ strata's. In this analysis, data were split by 16 age groups and by sex (male and female) ($J = 32$), where r_j represents the risk (or rate) of disease in strata j and N_{ij} represents the population size in area i for strata j. The SIR is then be calculated:

$$SIR_i = Y_i / E_i \tag{4.2}$$

Descriptive Statistics and Temporal Trend Analysis

Barplots of the raw counts of CDI, separated by HA-CDI and CA-CDI, were created to visualise the distributions over years. Total CDI counts were also plotted by financial quarters over all years. Temporal aspects of the data were then explored by aggregating over the spatial areas for each combination of year and quarter. CDI incidence (per 100,000 population) over the quarter-year timepoints was plotted together with 95% Byar's confidence intervals and a GAM was fitted to assess potential non-linear trends (Chapter 2, equation 2.16). This gave an initial view of the overall trend of total CDI over time, and the quarterly varying effects each year. Univariate Poisson GLM models were then fitted for CDI incidence including temporal points, year and financial quarters as fixed effects. Modelling year as an ordered factor provided insight into the linearity of the yearly trend.

Spatial Analysis

Maps of SIR for total CDI, CA-CDI and HA-CDI were plotted to visualise the spatial distribution of *c-difficile* infections. Moran's I test for spatial association was then applied to assess for spatial auto-correlation in total, community and hospital acquired CDI SIR. A 95% confidence interval was then constructed around the total CDI SIR calculation for each IZ: all 1279 IZ were classified either into "Lower", "Greater" or "Wide" groups dependent on whether: the upper confidence limit was < 1, lower confidence limit > 1 or whether the CI spanned 1. After categorising all IZs, they were colour coded and plotted onto a map which provided initial insight into clustered areas of consistently high and low CDI incidence between 2014 and 2018.

The correlation between pairs of covariates were checked to explore a potential collinearity problem: deprivation variables were expected to show some level of positive correlation. Scatter-plots of CDI SIR were then created to visualise the relationship between covariates with fitted GAMs to examine potential non-linear associations. Univariate Poisson GLMs explored the association between CDI SIR with Employment Deprivation (%), Income Deprivation (%), log population density and Forestry and Fishery Jobs (%) which gave an initial understanding of the associations between the covariates and CDI SIR. Age over 64 was not included in these analyses as the expected values used to calculated SIR were age-sex adjusted.

A multivariable Poisson GLM was then constructed for total CDI SIR to identify risk factors, applying backward selection by comparing model AIC using the function drop1: the final SIR model is shown in equation 4.3). Model assumptions were checked using Moran's I test for spatial association on model residuals, assessing independence and the dispersion parameter checked for equal mean and variance. The R-package **DHARMa** was used to visualise and test for dispersion using the functions testDispersion and plotSimulatedResiduals.

The SIR model is presented with the observed CDI counts represented by y_i for each IZ i with an *offset* of the expected CDI counts, E_i, per IZ. Model covariates included Employment Deprivation (%) ($Employ_i$) and Log_{10} Forestry and Fishery (%) ($ForestFish_i$) at each IZ i.

$$E(log(y_i)) = log(E_i) + \beta_0 + \beta_1 Employ_i + \beta_2 ForestFish_i \tag{4.3}$$

The GLM model covariates were then taken forward and CDI SIR was modelled spatially with a CAR Leroux prior, to account for spatial autocorrelation (section 2.3.2 for description of Leroux CAR Model). Stratified models were then run with the same covariates for HA-CDI and CA-CDI SIR to assess any differences in model estimates between classes of CDI infection (equation 4.4).

Where, y_i represents CDI counts (total CDI, HA-CDI and CA-CDI) for each IZ i with *offset* of expected CDI E_i.

$$E(log(y_i)) = log(E_i) + \beta_0 + \beta_1 Employ_i \tag{4.4}$$

Employment Deprivation (%), $Employ_i$ for each IZ i, was the only covariate carried forward to the spatial model as forestry and fishery (%) no longer contributed to the model fit after adjusting to spatial random effects.

Spatio-Temporal Analyses

An AR(1) spatio-temporal model with CAR Leroux prior was then constructed (Chapter 2, section 2.3.3 for description of the *Rushworth* AR(1) Leroux CAR Model). This analysis carried forward the risk-factors identified in the GLM model selection, this time introducing year as a fixed effect into the model. A spatio-temporal AR(1) model was run on five-years of temporally varying data by IZs to assess the ecological risk factors of CDI. Similar to the spatial modelling, the CA-CDI and HA-CDI cases were modelled separately with the selected final covariates (equation 4.5).

The observed and expected CDI counts (total CDI, HA-CDI and CA-CDI) are represented by y_{it} and E_{it} for each IZ i and year t.

$$E(log(y_{it})) = log(E_{it}) + \beta_0 + \delta_t Yr + \beta_1 Employ_{it} \tag{4.5}$$

Employment Deprivation (%) remained in the model as a fixed effect, $Employ_{it}$ for each IZ i and year t, with years (2014 to 2018) as a categorical fixed effect with contribution $\delta_t Yr$.

Finally, a spatio-temporal Clustered Trends Model was run to assessed individual IZ trends of total CDI over the years. This model is represented by two components: an overall spatial structure and multiple temporally varying trends including linearly decreasing, linearly increasing and constant. This model does not include covariates but aims to identify clusters of areas with similar temporal trends and highlight any differences between areas (Chapter 2, equation 2.25 for Clustered Trends model as proposed by *Napier et al.*).

4.3 Descriptive Statistics

There were a total of 7442 cases of CDI in Scotland from 2014 - 2018, with 2029 (27%) cases defined as CA-CDI and 5413 (73%) defined as HA-CDI. The number of CA-CDI and HA-CDI cases showed an overall decrease from 2014 to 2018, however, HA-CDI showed more variability between years (figure 4.2).

The quarterly difference in the number of CDI cases showed greater variability in 2014 and 2015 compared to other years. Throughout all years, Q3 showed a higher number of CDI cases, although Q2 and Q4 were also high in 2014 and 2015. Nevertheless, overall quarterly differences were small in later years (2016, 2017 and 2018).

4.3.1 Temporal Effects

Total CDI incidence (cases per 100,000 IZ population) decreased between years, with a 24.6% decrease in total CDI between 2014 and 2018. A decrease in CDI incidence was also seen for HA-CDI and CA-CDI. The lowest CA-CDI incidence was in 2017 (1.92, 95% CI 1.54 - 2.37) but increased slightly by 2018 (2.02, 95% CI 1.63 - 2.47) (table 4.2).

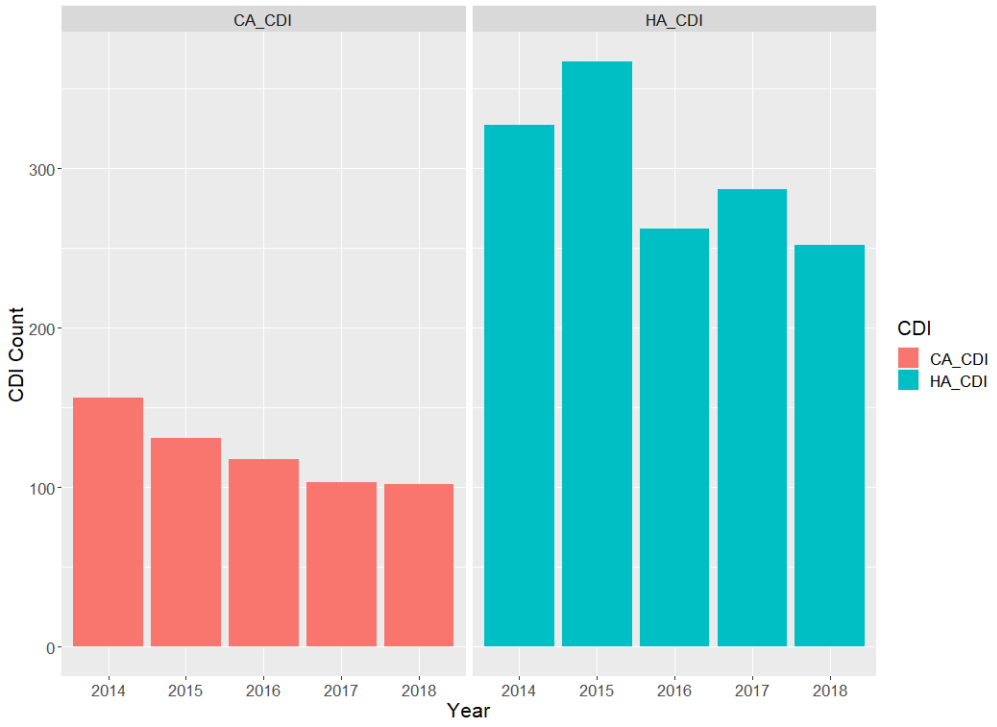


Figure 4.2: Barplot of the number of CDI cases split by CA-CDI and HA-CDI for 2014 to 2018.

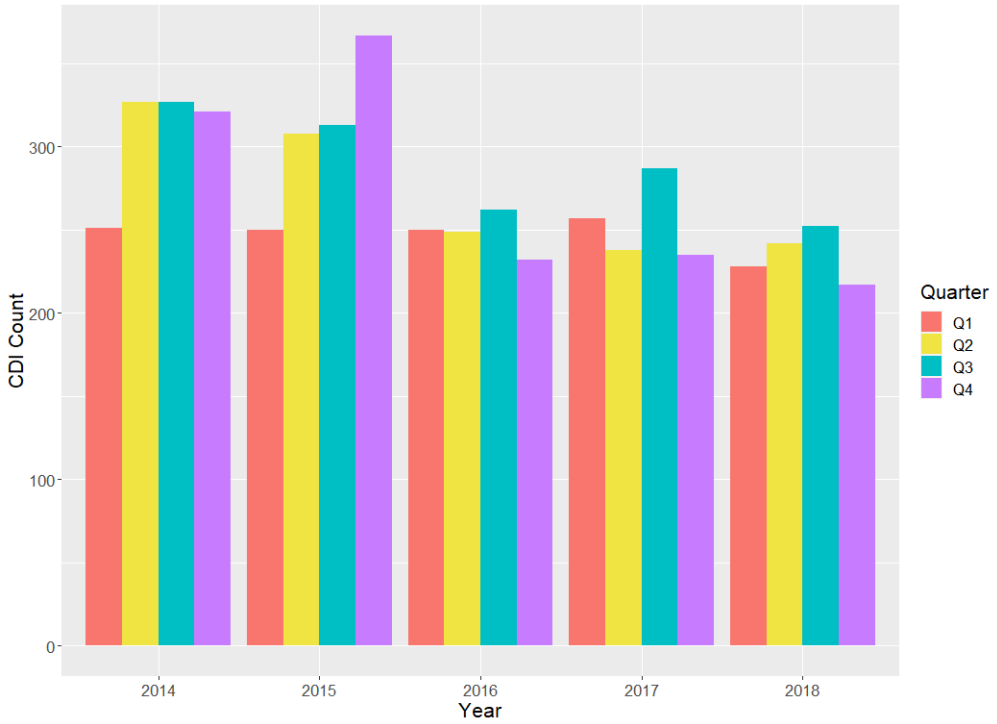


Figure 4.3: Barplot of the number of CDI cases by financial quarters (Q1 - Q4) for 2014 to 2018.

Table 4.2: CDI incidence (Total CDI, HA-CDI and CA-CDI, cases per 100,000 population) by year with 95% Byars's CI.

	Total CDI	HA-CDI	CA-CDI
2014	9.48 (8.60 - 10.42)	6.82 (6.08 - 7.63)	2.66 (2.21 - 3.18)
2015	9.25 (8.39 - 10.18)	6.85 (6.11 - 7.66)	2.40 (1.97 - 2.90)
2016	7.65 (6.86 - 8.49)	5.46 (4.80 - 6.19)	2.18 (1.78 - 2.66)
2017	7.50 (6.73 - 8.34)	5.57 (4.91 - 6.30)	1.92 (1.54 - 2.37)
2018	7.15 (6.40 - 7.97)	5.13 (4.50 - 5.83)	2.02 (1.63 - 2.47)

There was a visual quarterly effect with majority of years showing higher CDI incidence in Q2, Q3 and Q4 compared to Q1, with Q3 (July - September) showing the highest CDI incidence rate for all years except 2015. The difference between quarters reduced in later years. The fitted GAM showed an overall decreasing linear trend (figure 4.4).

Modelling the temporally aggregated data showed a decreasing trend in CDI incidence with each data point (RR = 0.983, 95% CI 0.979 - 0.987). CDI incidence modelled with financial quarters as a fixed effect showed an increase in CDI incidence for Q2, Q3

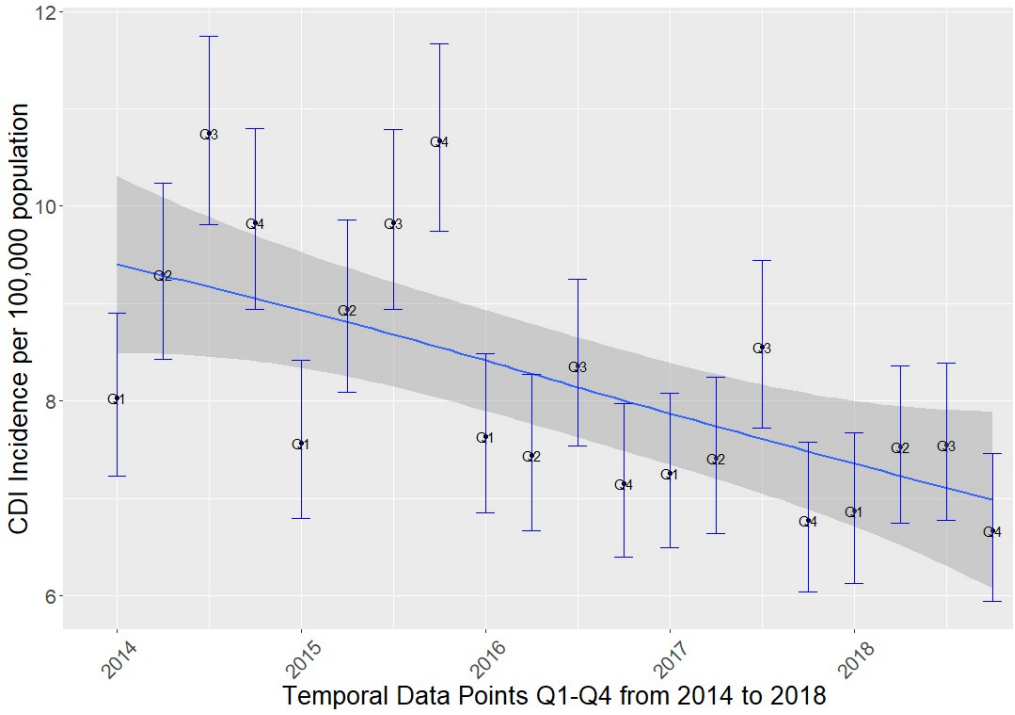


Figure 4.4: Error plot of yearly and quarterly CDI incidence trend with 95% confidence intervals by quarter and fitted GAM

and Q4 in comparison to Q1 (RR = 1.087, 95% CI 1.017 - 1.161; RR = 1.205, 95% CI 1.130 - 1.285 and RR = 1.099, 95% CI 1.029 - 1.174). Including year as a fixed effect showed no difference between 2014 and 2015, however, 2016, 2017 and 2018 all showed a reduced CDI incidence rate compared to 2014 (RR = 0.976, 95% CI 0.913 - 1.045; RR = 0.807, 95% CI 0.752 - 0.866; RR = 0.791, 95% CI 0.737 - 0.850 and RR = 0.755, 95% CI 0.702 - 0.811) (table 4.3).

Combining year and quarter into the same model showed the same estimates to univariable models. Year was modelled as an ordered factor which showed a linearly decreasing trend for year ($p < 0.001$) .

Table 4.3: Poisson GLM for temporally aggregated CDI incidence by year and financial quarters to assess linear trend (points 1-20), financial quarters and years with risk ratios (95% CI).

	Unadjusted RR (95% CI)	p-value
Linear Trend (1-20)	0.983 (0.979 - 0987)	<0.001
Financial Quarters: Q1	-	-
Q2	1.087 (1.017 - 1.161)	0.013
Q3	1.205 (1.130 - 1.285)	<0.001
Q4	1.099 (1.029 - 1.174)	0.004
Year: 2014	-	-
2015	0.976 (0.913 - 1.045)	0.488
2016	0.807 (0.752 - 0.866)	<0.001
2017	0.791 (0.737 - 0.850)	<0.001
2018	0.755 (0.702 - 0.811)	<0.001

4.4 Spatial Analyses Results

4.4.1 Exploratory Spatial Analysis

The SIR showed areas of high and low CDI incidence. The distribution of total CDI SIR, aggregated from 2014-2018, showed areas of high CDI incidence at the northern point of Scotland, with SIR > 2 (figure 4.5). There were a number of IZs with higher values of SIR in the central belt of Scotland, particularly Glasgow, whereas areas such as the Highlands generally appeared to have lower CDI than expected for the population national average.

Moran's I results showed evidence of moderate positive spatial association amongst IZs for total CDI SIR ($I = 0.19$, p-value < 0.001).

HA-CDI and CA-CDI both showed evidence of positive spatial association, however, this was stronger for HA-CDI compared to CA-CDI, ($I_{HA} = 0.18$, $p < 0.001$ and $I_{CA} = 0.08$, $p < 0.001$). HA-CDI predominately showed high SIR values in the central belt

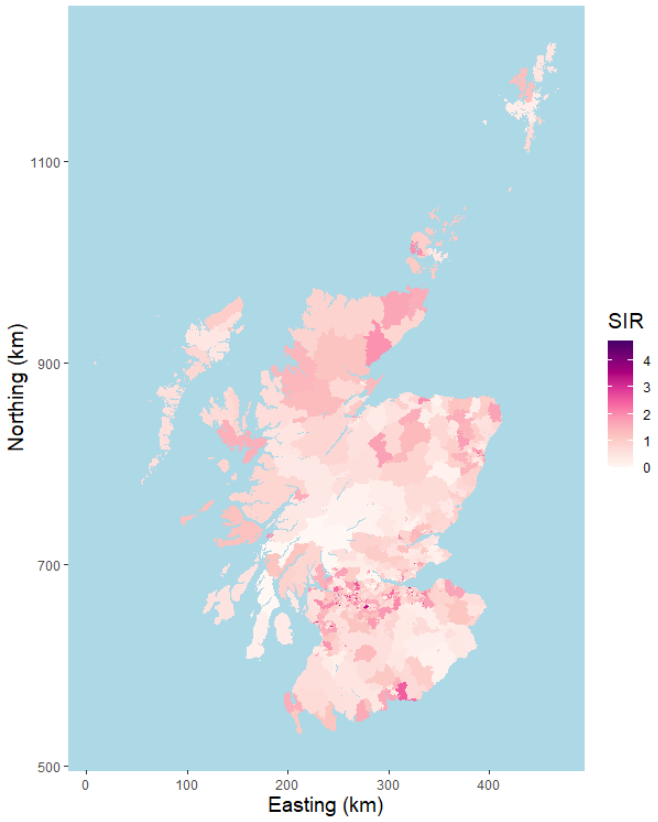
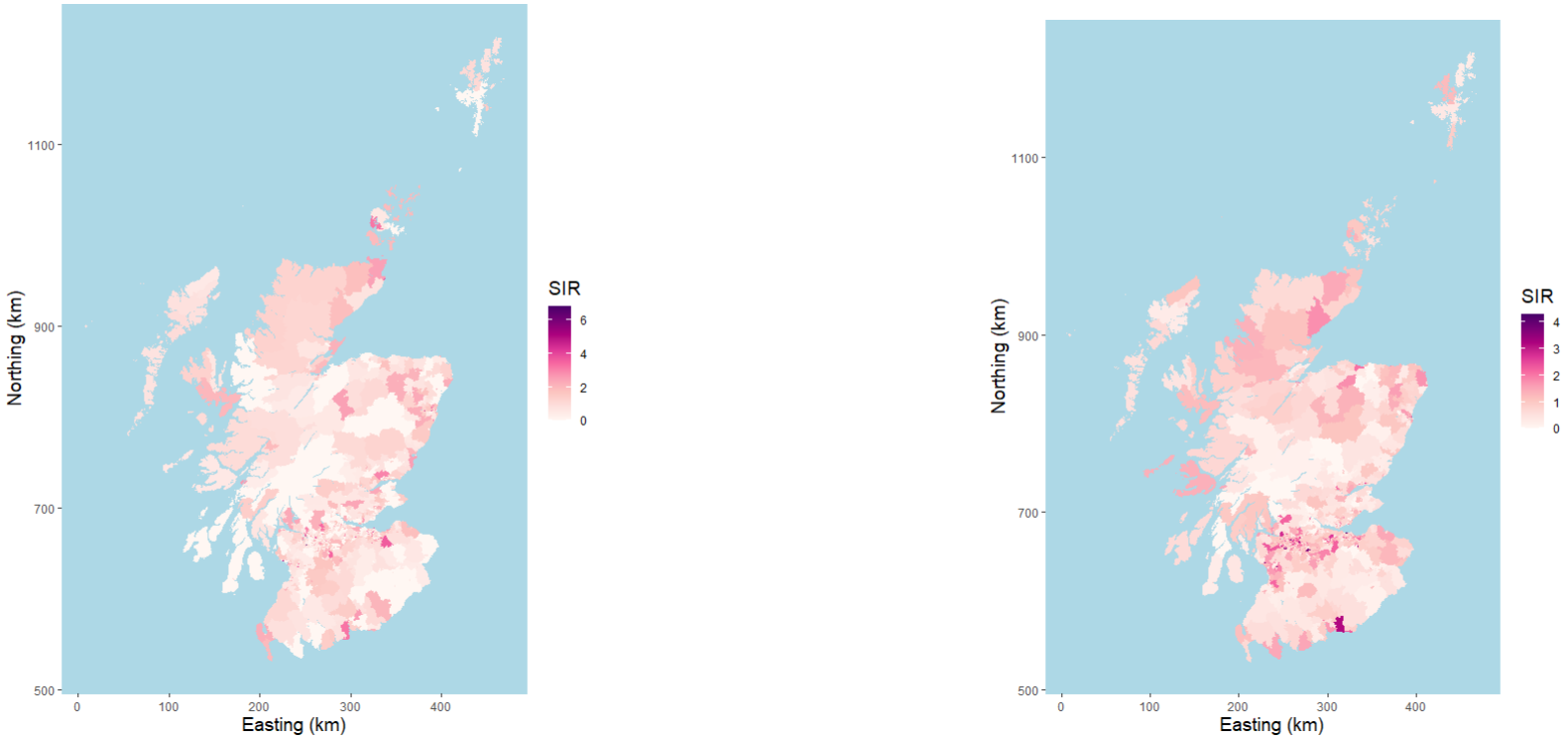


Figure 4.5: SIR plot of CDI incidence by IZ, aggregated from 2014 to 2018.

of Scotland and north of Scotland. CA-CDI SIR values were also higher in the North of Scotland (figure 4.6).



(a) CA-CDI

(b) HA-CDI

Figure 4.6: SIR plots of CDI incidence aggregated from 2014 to 2018 by CA-CDI (left) and HA-CDI (right).

To highlight areas of high and low CDI, 95% CIs were constructed around CDI SIR and colour coded into Greater, Lower or Wide. There were a small number of IZs with a 95% CI strictly greater than 1 (red), implying there were more infections than expected: areas in the Highlands, Gretna Green in Dumfries and Galloway and a small number of IZs in health boards NHS Greater Glasgow and Clyde and Lothian. Similarly, there were a few IZs with lower than expected CDI (yellow) in health boards including NHS Borders, Highlands and Tayside. However, the majority of Scotland showed CDI SIR with 95% confidence interval containing to 1 (figure 4.7).

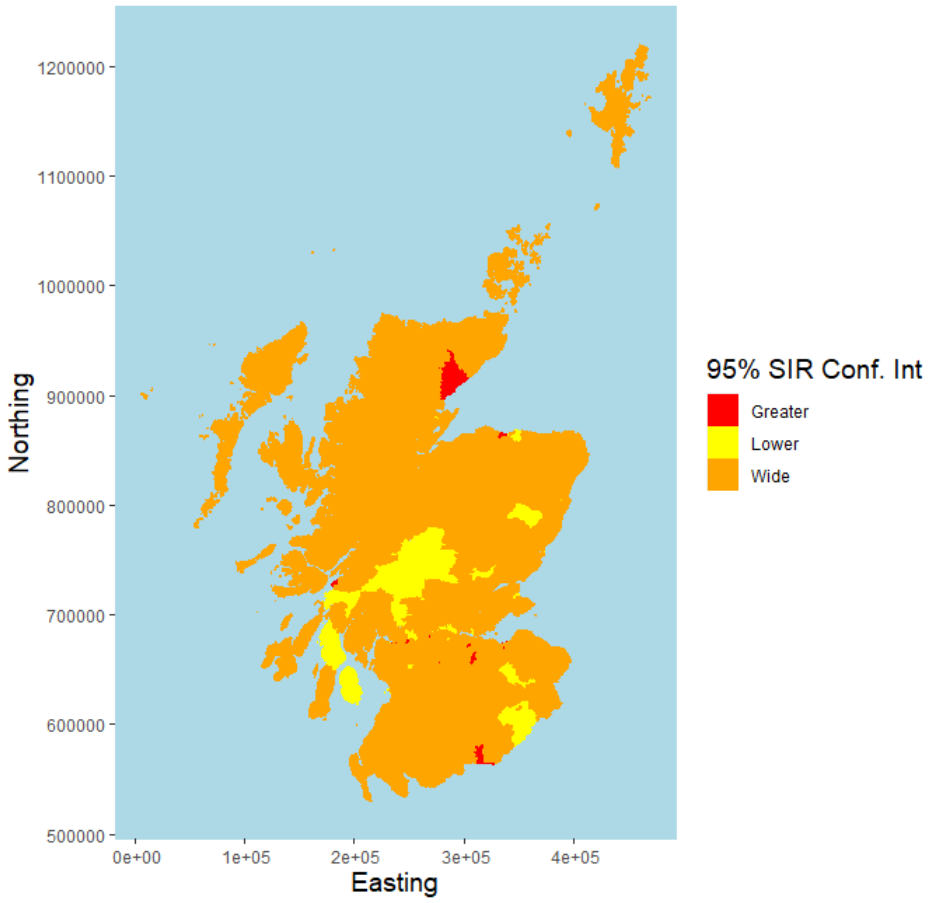


Figure 4.7: CDI Standardised Incidence Ratio categorised by 95% confidence Interval by IZ, aggregated from 2014 to 2018. A 95% CI was constructed around total CDI SIR and colour coded by intervals of strictly greater than 1 (red), strictly less than 1 (yellow) or wide (orange) for 95% CI containing 1.

4.4.2 Poisson GLM Results

Initially, a correlation matrix of all covariates was assessed. Percentage of IZ population income deprived and employment deprived were strongly positively correlated (ρ = 0.98). This was expected as they are both a measure of deprivation and would therefore lead to collinearity issues if included in a model together. Population Density and % IZ population aged over 64 years old were negatively correlated (ρ = -0.48). Similarly, population density was negatively correlated with % population working in forestry and fishery industries (ρ = -0.43).

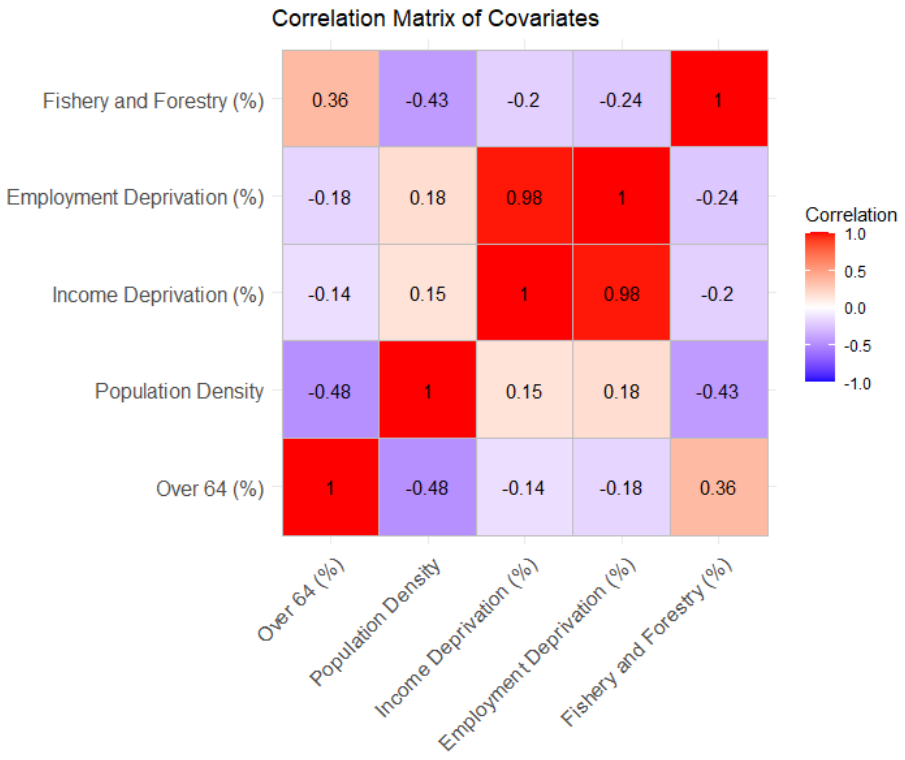


Figure 4.8: Correlation Plot of Spatial Covariates: Fishery and Forestry Jobs (%), Employment Deprivation (%), Income Deprivation (%), Population Density, Over 64 (%). Log base 10 transformations were taken for population density and forestry and fishery (%). Natural log transformation was taken for SIR to be comparable with modelling.

Figure 4.9: Covariate scatterplots for log total CDI SIR compared to Employment Deprivation (%), Income Deprivation (%), Log Fishery and Forestry Jobs (%) and Log Population Density with fitted GAMs.

Exploration of the potential non-linear nature of the associations between total CDI SIR and the covariates were assessed by fitting GAMs. Income deprivation and employment deprivation both showed similar positive linear associations with log total CDI SIR. The relationships between log SIR and log population density showed a slight increasing trend whereas the relationship with log percentage of IZ population involved in forestry and fishery industries showed a slight decreasing trend (figure 4.9).

An unadjusted GLM analysis showed that the risk of CDI increased with increasing percentage of IZ population employment deprived and income deprived (RR = 1.029, 95% CI 1.025 - 1.033 and RR = 1.013, 95% CI 1.012 - 1.015). Log population density was positively associated with an increased risk of CDI (RR = 1.108, 95% CI 1.080 - 1.138) and log percentage of population working in Forestry and Fishery industries showed a negative association with CDI SIR (RR = 0.874, 95% CI 0.846 - 0.904). This implied a reduced risk for higher percentage of IZ population working in forestry and fishery industries: this variable was most likely representing a measure of rurality.

Table 4.4: Unadjusted and adjusted GLMs for CDI SIR compared to spatial covariates: Employment Deprivation (%), Income Deprivation (%), Log Population Density and Log Forestry and Fishery Industry (%)

	Total CDI SIR	
	Unadjusted RR (95% CI)	Adjusted RR (95% CI)
Employment Deprivation (%)	1.034 (1.030 - 1.038)	1.031 (1.026 - 1.037)
Income Deprivation (%)	1.015 (1.014 - 1.017)	-
Log Population Density	1.108 (1.080 - 1.138)	-
Log Forestry and Fishery Industry (%)	0.874 (0.846 - 0.904)	0.935 (0.903 - 0.968)

A fully adjusted model was created and backward selection was performed. Income deprivation (%) was the first coefficient to be dropped from the multivariable model. Comparing model AIC showed no difference with and without the inclusion of income deprivation (AIC_{with} = 6539.4 vs $AIC_{without}$ = 6537.6). Income deprivation was also strongly correlated with employment deprivation (%) which showed a very

slightly stronger correlation with total CDI SIR ($\rho_E = 0.33$ vs $\rho_I = 0.32$). Following this, log population density was dropped from the model, ($AIC_{with} = 6537.6$ vs $AIC_{without} = 6535.5$) leaving employment deprivation (%) and log forestry and fishery (%).

In the adjusted model, the risk of CDI increased with increasing percentage of population employment deprived (RR = 1.031, 95% CI 1.026 - 1.037), whereas there was a negative association with percentage of IZ population working in forestry and fishery industries (RR = 0.935, 95% CI 0.903 - 0.968). Model assumptions were checked and there was no evidence of over-dispersion (see figure 4.10), however, the residuals from the model indicated positive spatial association when checked using Moran's I test for spatial association (I = 0.15, p <0.001), violating the Poisson model independence assumption. This model was then carried forward to be modelled spatially with a CAR Leroux prior.

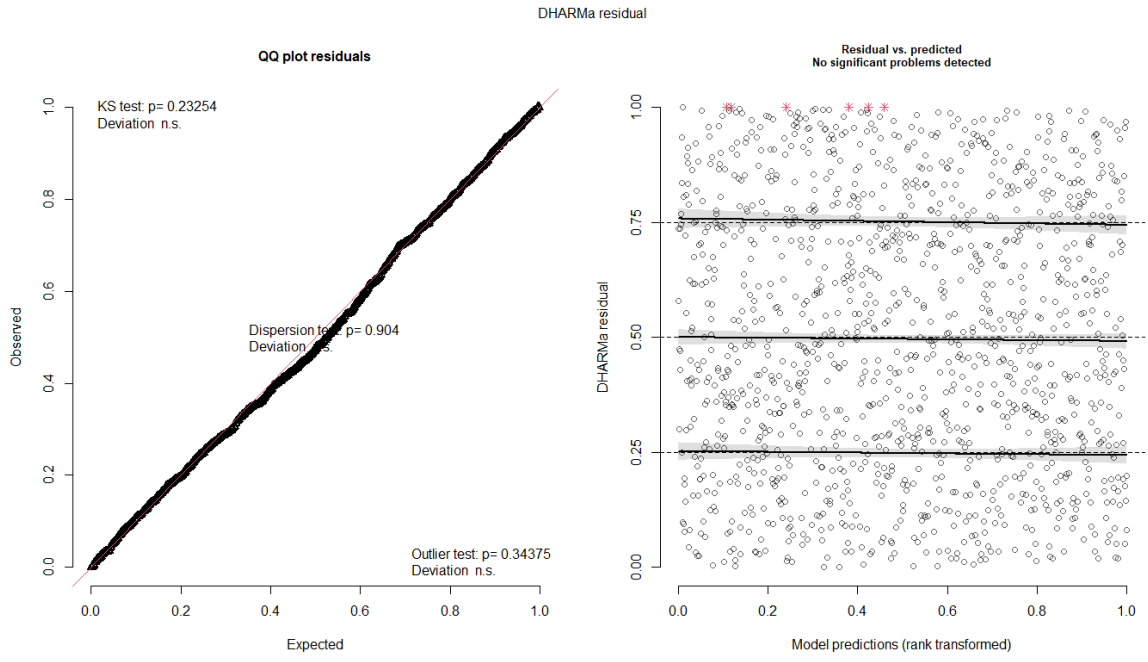


Figure 4.10: Test for over-dispersion for final fully adjusted poisson GLM

4.4.3 Spatial CAR Leroux Model Results

An intercept only model was assessed to estimate the amount of spatial variation and spatial dependence within total CDI data ($\tau = 0.298$ and $\rho = 0.823$). This could then be compared to the covariate models (table 4.5). A model with % of employment deprivation and log forestry and fishery (%) for total CDI was assessed, however, forestry and fishery (%) did not present as a strong predictor of total CDI with 95% CI spanning 1 (RR = 0.990, 95% CI 0.929 - 1.055) and was therefore dropped from the model. This result was also seen for CA-CDI and HA-CDI when modelled separately.

Spatial variability (τ) reduced from $\tau = 0.298$ in the total CDI intercept model, to $\tau = 0.188$ with the inclusion of fishery and forestry (%) and employment deprivation (%). Although, once fishery and forestry (%) was dropped, the spatial variability showed very little difference ($\tau = 0.189$). This showed that forestry and fishery (%) accounted for very little spatial variability in the model and the final model would only include employment deprivation (%) (**Case 3 Models**). Again, this was similar for CA-CDI and HA-CDI when modelled separately.

For **Case 3 Models**, the RR estimates were comparable with GLM results in table 4.4, employment deprivation (%) showed an increased association with total CDI SIR (RR = 1.032, 95% CrI 1.027 - 1.038). Comparing with the total CDI intercept model, employment deprivation (%) accounted for 36% of the spatial variability (table 4.5).

CA-CDI and HA-CDI showed similar estimates for employment deprivation (%) to total CDI SIR. Increased percentage of population employment deprived was associated with an increased risk of CA-CDI and HA-CDI, however, effect size was reduced with CA-CDI (RR = 1.021, 95% CrI 1.012 - 1.031 and RR = 1.036, 95% CrI 1.030 - 1.042). The spatial dependence (ρ) was estimated highest in the HA-CDI SIR models when

compared to all other models, however, all remained within 95% credible intervals of one and other (table 4.5).

Table 4.5: Spatial CAR Leroux GLM for total CDI, CA-CDI and HA-CDI SIR with Employment Deprived (%) and log Forestry and Fishery (%) with risk ratios (RR) and 95% credible intervals (CrI).

	Total CDI	CA-CDI	HA-CDI
	Intercept Models		
τ	0.298 (0.231 - 0.365)	0.308 (0.180 - 0.462)	0.304 (0.230 - 0.399)
ρ	0.822 (0.664 - 0.933)	0.828 (0.525 - 0.949)	0.887 (0.748 - 0.959)
	Case 2 Models		
Log Forestry and Fishery (%)	0.990 (0.929 - 1.055)	0.991 (0.895 - 1.094)	0.989 (0.919 - 1.063)
Employment Deprived (%)	1.032 (1.027 - 1.038)	1.021 (1.012 - 1.031)	1.036 (1.030 - 1.042)
τ	0.188 (0.135 - 0.251)	0.249 (0.137 - 0.397)	0.179 (0.119 - 0.241)
ρ	0.895 (0.751 - 0.963)	0.906 (0.719 - 0.975)	0.919 (0.772 - 0.974)
	Case 3 Models		
Employment Deprived (%)	1.032 (1.027 - 1.038)	1.021 (1.012 - 1.031)	1.036 (1.030 - 1.042)
τ	0.189 (0.134 - 0.252)	0.253 (0.131 - 0.405)	0.174 (0.110 - 0.246)
ρ	0.895 (0.751 - 0.963)	0.903 (0.697 - 0.974)	0.926 (0.801 - 0.977)

Model residuals were checked for spatial dependence with Moran's I test for spatial dependence which showed I = -0.08, p<0.001. This indicated evidence of relatively weak negative spatial dependence between IZ. This was an improvement from the positive residual spatial autocorrelation seen from the GLM results (I = 0.15, p<0.001), however, ideally no autocorrelation would be seen.

4.5 Spatio-Temporal Analyses Results

4.5.1 Spatio-Temporal AR (1) Model

An intercept only spatio-temporal AR(1) model was run to assess the spatio-temporal variability ($\tau = 0.406$) and autocorrelation parameters ($\rho_S = 0.852$ and $\rho_T = 0.597$). Employment deprivation (%) was then included into the model with the addition of year as a fixed effect to assess model estimates and effect on spatio-temporal variability.

For the total, hospital-acquired and community-acquired CDI SIR models, employment deprivation percentage (%) showed similar positive estimates seen in the spatial CAR Leroux model in table 4.5. Year showed similar results to the estimates seen in the temporal analyses in table 4.3: 2016, 2017 and 2018 showed a decrease in total CDI, HA-CDI and CA-CDI SIR in comparison to 2014, however, this was not seen for 2015 (table 4.5.1).

The spatio-temporal variability in the total CDI SIR model only slightly reduced when compared to the intercept model ($tau_{intercept} = 0.406$ and $tau_{covariates} = 0.395$).

Table 4.6: Multivariable Spatio-Temporal AR(1) GLM for Total CDI, HA-CDI and CA-CDI SIR with risk ratios (RR) and 95% credible intervals (CrI).

	RR (95% CrI)		
	Total CDI	CA-CDI	HA-CDI
Year: 2014	-	-	-
2015	0.953 (0.865 - 1.051)	0.877 (0.732 - 1.047)	0.974 (0.889 - 1.066)
2016	0.786 (0.703 - 0.878)	0.790 (0.647 - 0.969)	0.773 (0.700 - 0.852)
2017	0.758 (0.670 - 0.846)	0.686 (0.554 - 0.852)	0.781 (0.704 - 0.860)
2018	0.712 (0.628 - 0.797)	0.704 (0.558 - 0.887)	0.713 (0.640 - 0.788)
Employment Deprived (%)	1.030 (1.024 - 1.035)	1.020 (1.010 - 1.029)	1.037 (1.030 - 1.043)
tau	0.395 (0.300 - 0.487)	0.250 (0.132 - 0.469)	0.297 (0.178 - 0.434)
rho.S	0.666 (0.373 - 0.849)	0.939 (0.833 - 0.979)	0.840 (0.201 - 0.938)
rho.T	0.489 (0.360 - 0.613)	0.668 (0.446 - 0.876)	0.603 (0.443 - 0.754)

Although not directly comparable, the spatio-temporal variability in table 4.5.1 for total CDI was greater than in the purely spatial analysis (table 4.5). This is because there was extra Poisson variability in the spatio-temporal analyses as the counts of CDI were much lower when divided by year and IZ compared to aggregated CDI counts by IZ in the spatial analyses. The spatial dependence (ρ_S) was lower in the spatio-temporal analyses than in the spatial analysis. Total CDI, however, remained high when split separately from HA-CDI and CA-CDI. The temporal dependence (ρ_T) was lower for the total CDI model in comparison to the stratified HA-CDI and CA-CDI spatio-temporal models.

4.5.2 Spatio-Temporal Cluster Model

Figure 4.7 indicated IZs with different total CDI SIRs for IZ's, with some IZ's showing CDI cases strictly greater than and less than expected CDI, however, the majority of IZs showed wide 95% CIs surrounding total CDI SIR. It was of interest to assess the possibility of varying temporal trends of CDI over IZs. An intercept only model of total SIR was created to assess the data for clusters of linearly decreasing, linearly increasing and constant trends by year.

Table 4.7: The allocation of IZs to temporal trends from a spatio-temporal clustered trend model over the time period of 2014 to 2018.

Constant	Linearly Decreasing	Linearly Increasing
635	150	494

The 1279 IZs were classified into either constant, linearly increasing or linearly decreasing trends over the 5-year time period. From table 4.7, 39% of the IZs were classified to have linearly increasing; 11% were linearly decreasing and 50% were classed as constant. However, assessing probability plots of each IZ questioned the strength of these results (figure 4.11).

Probability boxplots of each classification showed that the probability of the trend in an IZ being classified as 'Constant' were all very close. Linearly decreasing trend in an IZ and linearly increasing trend in an IZ both showed close probability to being classified as constant, although there was a slightly lower chance of being classified a completely opposing trend (linearly decreasing vs linearly increasing). Regardless, none of these probabilities were particularly strong. These results should, therefore, be handled sensitively (figure 5.10).

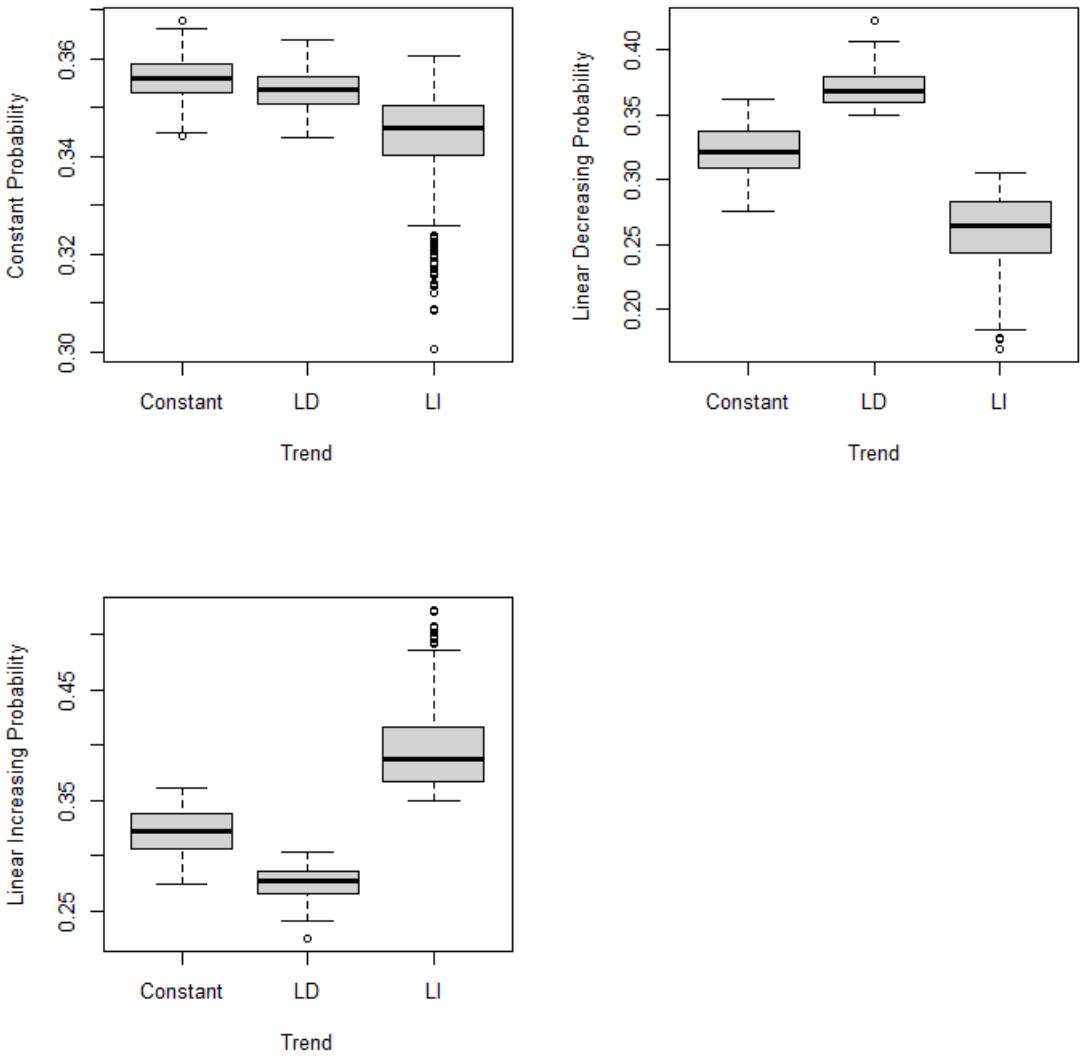


Figure 4.11: Boxplots of probabilities for Constant (top-left), Linearly Decreasing (LD) (top-right) and Linearly Increasing (LI) (bottom-left) temporal trends.

Plotting the trends onto a map showed high variability across Scotland, with the over a third of IZ showing an increasing trend in CDI. However, the probabilities for this analysis were not strong and, therefore, are not conclusive. As previously shown in this chapter, the overall trend of CDI is decreasing over time which contradicts of the results from this model which show a large proportion of IZs linearly increasing (figure 4.12).

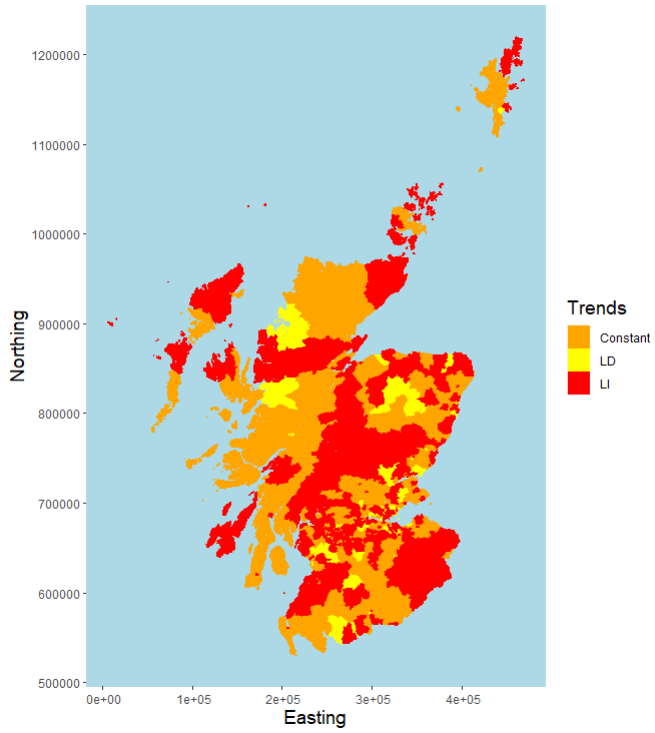


Figure 4.12: Constant, Linearly Increasing (LI) and Linearly Decreasing (LD) clustered trends in Scotland.

4.6 Discussion

This chapter has shown evidence of spatial and temporal patterns of CDI incidence in Scotland. There was positive spatial autocorrelation seen between intermediate zones for crude CDI incidence. Adjusting for population-based risk factors accounted for some spatial variation, however, spatial dependence remained. A quarterly effect was seen across years, showing quarter 1 (January - March) to have the lowest CDI incidence in comparison to all other quarters and a linearly decreasing trend was seen over the five-year study period. Percentage of IZ population employment deprived was identified as an ecological risk factor showing a positive association with an increase in risk of CDI (RR = 1.032, 95% CI 1.027 - 1.038, per 1% increase).

These analyses showed relatively strong spatial auto-correlation for total CDI incidence (Moran's I = 0.19, $p < 0.001$) and stratified HA-CDI ($I_{HA} = 0.18$, $p < 0.001$) however was weak for CA-CDI I_{CA}= 0.08, $p < 0.001$). The multivariable GLM analyses showed percentage of IZ population employment deprived (%) to have a positive association with CDI incidence, whereas working in forestry and fishery industries (%) showed a negative association, a presumed measure of rurality. After adjusting for these population-based risk factors there was remaining evidence of spatial autocorrelation amongst residuals when tested using Moran's I (I = 0.15, $p < 0.001$). To account for residual spatial auto-correlation, spatial random effects were included into the model with a Leroux CAR spatial model and population-risk factors were reassessed. Employment deprivation remained a strong predictor of CDI incidence and accounted for 36% of the estimated spatial variability.

Comparing with other spatial studies of CDI incidence, a study in Queensland reported no evidence of spatial autocorrelation amongst CDI cases, with Moran's I statistic (0.002, 95% CI 0.005 to 0.001) indicating spatial randomness [159]. Another study in Australian Capital Territory reported significant spatial correlation ($p < 0.004$) at a 5% significance level, however, they did not report the Moran's I correlation statistics [158]. A study of CA-CDI in North Carolina reported significant clusters, or hot-spots, of infection (Getis-Ord $p < 0.001$: a test for high/low spatial clustering) [53], whereas a study in the Netherlands reported no evidence of spatial clustering amongst CA-CDI [160]. It is understood that these are the only studies to assess spatial autocorrelation and clustering of CDI. It is understood that this chapter is the first to explore spatial patterns of CDI in Scotland, and has shown strong spatial associations for HA-CDI that could not be accounted for by population-based risk factors and census information, however there was no strong evidence of spatial association for CA-CDI comparable with the study in Netherlands [53]. Extending these analyses to include spatio-temporal random effects, using an AR(1) spatio-temporal model, again showed

employment deprivation showed a positive association with CDI incidence, with each year of study showing a decreased risk of CDI incidence when compared to 2014.

The temporal analysis showed the winter months (quarter 1) to have the lowest risk of CDI compared to all other quarters, with a 20% increase in risk of CDI in the summer months (July - September). This contradicts other studies of CDI seasonality [163, 164], except for a study in Queensland which also reported increased CDI in summer months which was a noted disparity [159].

Unfortunately, seasonal effects could not be modelled for these data due to low counts of CDI. The CDI data were accessible by intermediate zones (N = 1279). This is a relatively small geography, therefore, when divided by years and quarters the counts of CDI become zero inflated and unrealistic to model. The next accessible spatial scale for these data would be local authority (N = 32 in Scotland), which would allow for the quarterly variations to be modelled suitably. However, the reduced spatial scale is likely too large to be able to detect beneficial population-based risk factors. It is, therefore, a trade-off when modelling these data dependent on the goal of the analysis. To the best of our knowledge, there has been only one other spatio-temporal analysis of CDI, however, there was no evidence of spatial auto-correlation found and, therefore, spatial random effects were not accounted for in the model [159]. Spatial and spatio-temporal modelling are strong tools for removing spatial and temporal sampling biases. When applied to high incidence infectious disease such as COVID-19, spatio-temporal modelling has been shown to be highly effective in the development of real-time surveillance tools in highlighting areas of preventable disease outbreak [166, 167].

In conclusion, this study has shown novel evidence of spatial patterns amongst CDI incidence which could not be explained by population and census information, with the additional novel application of spatio-temporal modelling of CDI in Scotland. Min-

imising spatial and temporal biases when identifying risk factors of CDI is crucial to the management of infection. Chapter 5 will carry forward the spatio-temporal models defined in this chapter and explore other population-based risk factors that are on the causal pathway for *c-difficile* infection, including broad spectrum antibiotic prescribing in the community and environmental factors such as proximity to livestock

Chapter 5

Clostridiods Difficile Infection Associated with Primary Care Antibiotic Prescribing and Cattle Density in Scotland for Multilevel Spatial Data

5.1 Introduction

This chapter is a continuation of the analysis conducted in Chapter 4, which explored spatial and temporal trends of CDI data at intermediate zone (IZ) level between 2014 and 2018, and investigated ecological risk factors available at the same spatial scale. The percentage of IZ population employment deprived was identified as risk factors of CDI at an IZ level. Spatial and temporal autocorrelation were highlighted between IZs, employment deprivation explaining a proportion of the spatial variation. However,

some spatial dependence remained and therefore the final model accounted for both temporal and spatial random effects.

As previously mentioned in Chapter 4, exposure to the 4C high-risk broad-spectrum antibiotics (cephalosporin's, coamoxiclav, fluroquinolones or clindamycin) is a well-defined causal risk factor associated with an increased risk of CDI [58, 168, 57]. The impact of antimicrobial prescribing in the community on the risk of CDI has been highly reported at an individual-level, for both HA-CDI and CA-CDI [58, 57, 59]. Recent exposure to a high-risk antibiotic has been shown to be associated with an increased risk of CA-CDI, however, an increased risk is still present for an antibiotic prescription 4-6 months prior to infection. The ecological effects of community prescribing on the risk of CDI has been less reported, however, a study of Welsh GP practices' antibiotic prescribing was seen to be associated with an increased risk of CDI, particularly for clindamycin prescribing (Chapter 6, [1]).

Environmental factors such as farming have also been highlighted as risk factors on the causal pathway for an increased risk of CDI [160, 169]. This is particularly highlighted for livestock farming [169] and the impact on CA-CDI [53], however, an association with HA-CDI has also been highlighted [54]. Common human *c-difficile* isolates have been found amongst livestock animals and on farm workers [170]. A study in Spain showed the most common strain of human *c-difficile* infections in the country was commonly found amongst pigs, suggesting a potential source of epidemic multidrug resistant strains [171].

Data for these established risk factors were available spatially in Scotland: GP practice antibiotic prescription point-location data were previously created for Chapter 3 from NHS open data platform [140], and Scottish cattle density data were available from the Scottish Government at areal-level agricultural Parish Council (N = 891) [172, 173].

However, these data were not available on the same spatial scale as the CDI data (IZs) and were incompatible. This presented a multilevel spatial problem.

One method for handling multi-level spatial data is interpolation. Interpolation is a method of spatial prediction and smoothing that can be used for point-level or areal level data, with areal-level interpolation being a particularly useful method for combining data whose boundaries do not coincide [174]. Interpolation is frequently used in environmental sciences for transforming point-to-area and area-to-point spatial data and making spatial predictions, with global kriging interpolation highlighted as a powerful method of spatial prediction [175, 106]. However, more simplistic methods such as inverse-distance weighted (IDW) interpolation are useful in the absence of strong spatial autocorrelation and have been shown to out-perform kriging in some instances [105]. There is currently no "rule of thumb" for choosing an interpolation method as effectiveness varies situationally and, therefore, it is recommended to try multiple methods where possible and quantitatively compare [107].

This chapter aims to investigate antibiotic and environmental exposures as ecological risk factors of CDI with data from multiple spatial scales. GP antibiotic prescribing data were collated as GP practice level spatial point-location data (see Chapter 3) for 2016 to 2018 and Cattle Density data were provided at agricultural Parish Council for 2019. This chapter will apply methods of spatial interpolations as a means for transforming multi-level spatial data to aid analysis.

5.2 Methods

This chapter initially focuses on addressing a multi-level spatial data problem using interpolation methods, with an aim of converting GP practice antibiotic prescribing data and cattle density data by Agricultural Parish into a form that can be represented at an intermediate zone (IZ) level. These data could then be combined with existing CDI data by IZs, previously introduced in Chapter 4 to assess antibiotic prescribing and cattle density as risk factors of CDI using spatio-temporal modelling.

5.2.1 Data

GP Antibiotic Prescribing Data

Open-sourced GP practice antibiotic prescribing data [140], as described in Chapter 3, were manipulated to include counts of total antibiotics, 4C antibiotics (high risk antibiotic groups combined) and individual high-risk antibiotic groups (clindamycin, cephalosporins, quinolones and co-amoxiclav) by GP practice for 2016 to 2018 (data were unavailable prior to 2016). Rates of antibiotic prescribing (items per 1000 registered patients) were calculated by dividing antibiotic counts by GP practice populations [94]. Data were collected for all GP practices in Scotland each year (mean = 923), however, the number of practices varied between years due to practice closures and merges and missing GP practice population data (Chapter 3). Antibiotic prescribing data were merged to GP practice population information, including postcode, by unique GP practice identifier. Easting and northing coordinates were then obtained from the GP practice postcodes to allow the data to be transformed into a Spatial Points Data Frame.

Cattle Density

Cattle density data for Scotland were obtained from the Scottish government by agricultural parish for 2019. These data are routinely collected each year, however, were not available to download from the Scottish Government agricultural platform [176], but were provided upon enquiry. Data were obtained for 2019 only, however, it was assumed that these data would not vary greatly between years. In Scotland there are 891 agricultural parishes. The unique parish number was used to merge data to an agricultural parish shapefile, downloaded from the Scottish Government spatial data platform [172]. The data were merged to create a Spatial Polygon Data Frame which included boundary information and the total number of cattle per parish, parish name, district name, area of agricultural parish by hectare (ha) and cattle density (number of cattle per ha) (figure 5.1).

	parish	parname	district	Area.of.agricultural.parish..ha.	total_cattle	Holdings.with.cattle	cattle_pha
1	1	ABERDEEN	City of Aberdeen	11080.887	2036	24	0.18373981
2	2	BELHELVIE	Gordon	4921.031	1982	18	0.40276113
3	3	DRUMOAK	Kincardine & Deeside	2993.147	1190	6	0.39757484
4	4	DYCE	City of Aberdeen	7494.001	3881	22	0.51788094

Figure 5.1: Data structure screenshot of cattle density by agricultural parish (AP).

Cattle density data were positively skewed with multiple parishes showing low cattle density, however there were only seven parishes that showed holdings with zero cattle (cattle density = 0) (figure 5.2).

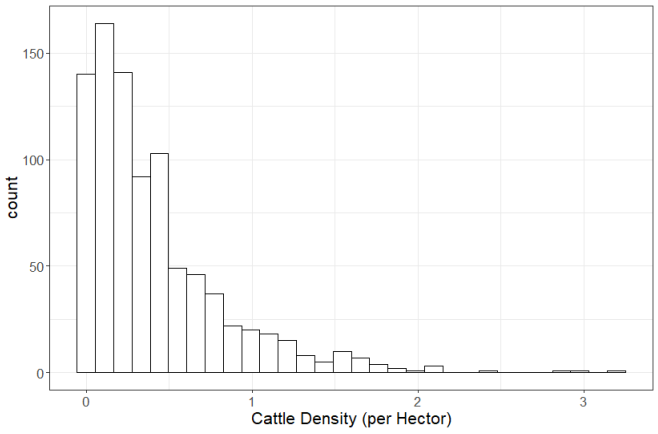


Figure 5.2: Histogram of Cattle Density by Agricultural Parish in 2019

Clostridiodies-difficile Infections by Intermediate Zones

CDI data by IZ were previously introduced in Chapter 4, sections 4.2.1. These data include counts of total, hospital-acquired and community-acquired CDI by IZs with corresponding expected CDI counts for 2014 - 2018. For the purpose of this analysis the data were subset for 2016 to 2018 to match the GP antibiotic prescribing data.

The antibiotic prescription and cattle density data sets were handled separately with the aim to transform them from GP point-location data (N $\approx$ 923) and areal-level agricultural parishes (N = 891) to areal-level IZs (N = 1279). These data could then be merged to the CDI data to show a measure of antibiotic prescribing for years 2016 to 2018 and cattle density (which will be assumed the same for all years) for each IZ. These variables could then be assessed as risk factors of CDI in spatio-temporal models.

5.2.2 Spatial Interpolation

Two methods of spatial interpolation were considered for these analyses: inverse-weighted distance (IDW) and kriging. Interpolation is a method of spatial prediction, commonly used in environmental sciences as a method of transforming areal-to-point and point-to-areal spatial data [175]. See Chapter 2, section 2.2 for a detailed description of these methods.

Interpolation methods are used to make spatial predictions at new locations based on neighbouring information at known locations. To conduct interpolation a prediction grid is required which for these analyses consisted of the centroids of each intermediate zone (IZ) in Scotland. This would allow a measure of antibiotic prescribing and cattle density to be predicted at the IZ centroids. Cross-validation was applied to measure prediction accuracy and aid method comparisons. The observed data were split into training and test data, then the method of interpolation was applied to obtain spatial

predictions. The root mean squared error (RMSE) was then calculated for each set of predictions and the method of interpolation that minimised the RMSE was then accepted to calculate the final predictions at IZ centroids. These predictions were then carried forward to be assessed in the final spatio-temporal models.

Interpolating Antibiotic Prescribing Data

Kriging interpolation was not possible for the antibiotics data. This method of interpolation requires data to be spatially correlated as it relies on a fitted variogram model to obtain predictions. As seen in Chapter 3, figure 3.13 presented Monte Carlo's test for spatial autocorrelation on the crude antibiotic rates which showed no evidence of spatial autocorrelation at GP practice point level for total antibiotic prescribing. In conjunction, there was no strong evidence of spatial autocorrelation for high-risk antibiotic groups at health-board level when assessed individually (Chapter 3, section 3.3.1).

Therefore, the GP antibiotic prescribing data were interpolated using IDW. Total antibiotic rates for 2016 were initially interpolated using multiple power values, p, for various levels of soothing ($p = 0.1, 1, 5, 10$), and then visualised by mapping the antibiotic spatial predictions by IZ. The optimum power value p was then chosen using cross-validation of the observed data, applying IDW interpolation on training and test data for $p = 0.5$ to 30 and calculating the RMSE for each set of spatial predictions. The power value p that minimised the RMSE was then chosen to make final predictions of total antibiotic prescribing at IZ centroids. The same power value p was applied for all other antibiotic variables (4C combined, cephalosporins, coamoxiclav, quinolones and clindamycin individually) and repeated for every year.

Interpolating Cattle Density Data

A point-location version of the cattle density data set was created by assuming values of cattle density (per hectare) at each agricultural parish centroids. These data were strongly spatially correlated (Moran's I = 0.37, p < 0.001) and therefore could be interpolated using kriging interpolation. As no covariate information was accounted for, ordinary kriging was applied. A variogram was fit to the data and the multiple variogram models were assessed to see which fit best. Exponential, linear and spherical variogram models were fit, then the chosen model was applied to obtain kriging prediction. Cross-validation of kriging predictions from training and test subsets of the observed data were used to calculate the RMSE.

IDW interpolation was then also applied to obtain predictions of the cattle density data using the same process as the antibiotic data. Multiple values of p were used to visually compare smoothing of cattle density predictions ($p = 0.1$, 1, 5 and 10) before the RMSE was calculated for multiple IDW power values ($p = 0.5$ to 30) from cross-validation of cattle density data.

The method, and corresponding conditions, of interpolation carried forward to calculate the final cattle density predictions by IZs was then chosen based on the minimised the RMSE.

5.2.3 Sensitivity Analysis

Impact on Model Estimates

Both of the interpolation methods used in these analyses had subjective elements and limitations, therefore it was important to compare the impact of different methods of spatial prediction on analyses. A sensitivity analysis was performed using univariate Poisson GLMs of total CDI SIR to compare antibiotic prescribing and cattle density IZ

predictions from multiple power values p for IDW and kriging interpolation to assess the impact on model estimates and p-values. GLMs were considered appropriate for these analyses as the aim of was not to interpret the relationship between CDI and covariates, but to compare between model estimates of multiple spatial predictions of the same covariates.

Antibiotic IDW Interpolation Power Value

The same power value p for IDW interpolation was applied for all antibiotic groups (total, 4C combined, cephalosporins, coamoxiclav, quinolone and clindamycin individually), determined from total antibiotics in 2016. Cross-validation was applied to each individual high-risk antibiotic group, calculating the RMSEs for multiple power values ($p = 0.5$ - 30) to assess whether the chosen power value p was suitable in comparison to the RMSE from other values of p.

5.2.4 Spatio-temporal Models

Once a measure of GP antibiotic prescribing exposure (total antibiotics, 4C antibiotics and individual high-risk antibiotics) and cattle density were determined by IZ, each variable was assessed as a potential risk factor for CDI. Multiple fully adjusted spatio-temporal models, including the previously defined covariates from Chapter 4, section 4.5, were assessed separately for total CDI, HA-CDI and CA-CDI, and compared to each of the antibiotic groups (total, 4C combined , cephalosporins, coamoxiclav, quinolone and clindamycin).

As in Chapter 4, y_{it} represents CDI (total, HA- and CA-CDI for separate models) for IZ i and year t, with *offset* of expected total, HA- and CA-CDI (E_{it}). The covariates $Antibiotic_{it}, Cattle_{it}$ and $Employ_{it}$ represent each Antibiotic Variable (100 items per 1000 registered patients), Cattle Density (per ha) and Employment Deprivation (%)

by IZ i and years (2016, 2017 and 2018) as a categorical fixed effect with contribution $\delta_t Yr$ (equation 5.1).

$$E(log(y_{it})) = log(E_{it}) + \beta_0 + \delta_t Yr + \beta_1 Antibiotic_{it} + \beta_2 Cattle_{it} + \beta_3 Employ_{it} \quad (5.1)$$

Antibiotic groups (total, 4C, cephalosporins, coamoxiclav, quinolone and clindamycin) were assessed separately for total CDI, HA-CDI and CA-CDI represented adjusting for the same model covariates as equation 4.5, replacing X_{1ij} in each model to for the different antibiotics groups.

A total of 18 spatio-temporal models were run. Each model was set-up with the same model specifications: burnin (N=50,000), thinning (N=20) and samples (N = 150,000) - implying that a total of 5000 posterior samples remained for each parameter. Uninformative priors were applied for all models.

5.3 Results

5.3.1 Interpolation Results

GP Antibiotic Prescribing

Median total antibiotic prescribing rates (items/1000 registered patients) have decreased by 7% from 2016 to 2018 (Med_{2016} = 645.9 to Med_{2018} = 600.5, items per 1000 registered patients). High-risk antibiotic prescribing also showed a decrease in prescribing with a 5% decrease in 4C prescribing (items/1000) between 2016 and 2018 (Med_{2016} = 47.1 to Med_{2018} = 45.0, items per 1000 registered patients). Cephalosporins and coamoxiclav both showed a lower prescribing rate in 2018 compared to 2016, however, there was also a slight increase between 2016 and 2017. Quinolones showed a

decrease in prescribing each successive year, whereas clindamycin showed a slight increase each year ($Med_{2016} = 0.57$ to $Med_{2018} = 0.62$, items per 1000 registered patients). Although it is noted that clindamycin prescribing rates were very low in comparison to others antibiotic groups (table 5.1).

Table 5.1: Median (IQR) GP antibiotic prescribing rates from 2016 to 2018 for total antibiotics, 4C combined antibiotics, and individual high-risk antibiotic groups (cephalosporins, coamoxiclav, quinolone, and clindamycin) (items per 1000 registered patients).

	Median (IQR) GP Antibiotic Prescribing Rates (items per 1000 registered patients)		
	2016	2017	2018
Total Antibiotics	645.8 (552.0 - 757.7)	629.8 (531.1 - 731.0)	600.5 (507.8 - 696.4)
4C Antibiotics	47.1 (35.4 - 63.0)	47.3 (35.3 - 62.3)	45.0 (33.0 - 60.1)
Cephalosporins	9.6 (5.3 - 17.2)	9.8 (5.5 - 16.5)	9.2 (5.2 - 15.7)
Coamoxiclav	19.7 (14.4 - 26.5)	20.0 (14.1 - 26.7)	19.1 (13.9 - 25.5)
Quinolone	14.3 (9.8 - 19.7)	14.1 (9.5 - 19.6)	13.6 (9.3 - 18.3)
Clindamycin	0.57 (0.20 - 1.13)	0.61 (0.21 - 1.28)	0.62 (0.20 - 1.36)

There was no clear visual trend in the spatial distribution between years for total antibiotic prescribing rates for mainland Scotland, with higher rates of antibiotic prescribing focused in the central belt of Scotland and towards the west coast (Greater Glasgow and Clyde health board). There were a few high prescribing GP practices in the Scottish Borders and north/ north east points of Scotland, however, this was consistent across all three years of prescribing data (figure 5.3).

GP Antibiotic Prescribing Interpolation

The same interpolation power value p was applied to each antibiotics group (total antibiotics, cephalosporins, coamoxiclav, quinolones and clindamycin) for every year and was determined from an initial assessment of multiple power values p for total antibiotic rates in 2016. Maps of interpolated total antibiotic rates (items/1000) were initially assessed for varying degrees of smoothing (p = 0.1, 1, 5, 10) (figure 5.4).

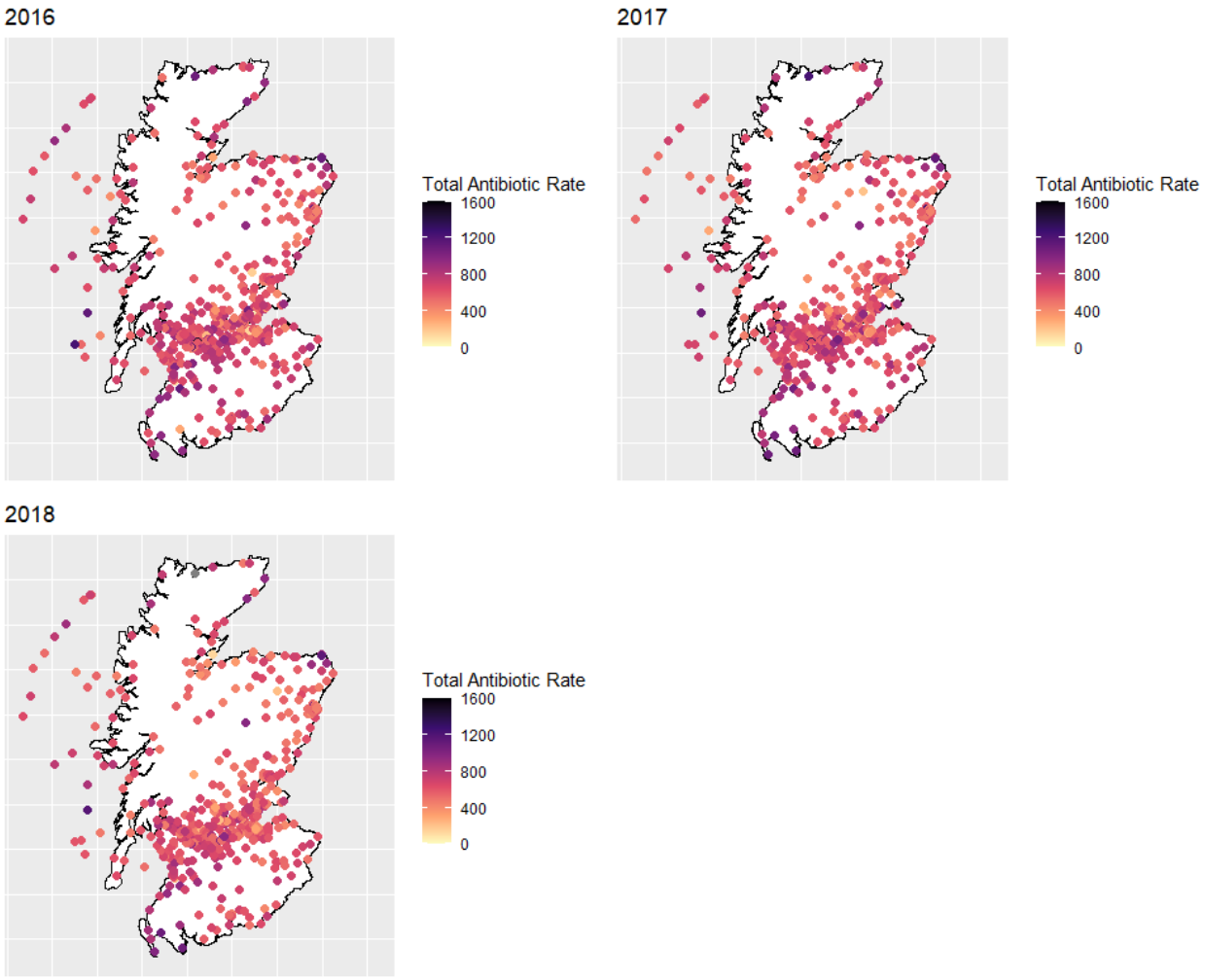


Figure 5.3: Point maps of GP antibiotic prescribing rates (items per 1000 registered patients) for 2016 (top-left), 2017 (top-right) and 2018 (bottom-left).

A very low power value, $p = 0.1$, over-smoothed the antibiotic prescribing data, with a low spread of antibiotic values (604 items/1000 to 612 items/1000), which is not reflective of the spread seen in the original data: median (IQR) = 645.8 (552.0 - 757.7) in 2016, with maximum = 1475.8 and minimum = 16.0 items per 1000 registered patients (figure 5.4). Increasing the power value to $p = 1$ showed more variability and began to highlight a spatial pattern similar to that seen in figure 5.3. As values of p were increased further to $p = 5$ and $p = 10$, there did not appear to be much difference in IZ total antibiotic prescribing rates (figure 5.4) with both showing similar spatial distributions. The interpolated values for each IZ were heavily influenced by

Figure 5.4: Inverse Distance Weighted Interpolation: p = 0.1 (top-left), p = 1 (top-right), p = 5 (bottom-left), p = 10 (bottom-right) for GP antibiotic prescribing rates 2016

the closest possible GP practices at these power values and, therefore, were more representative of the spatial patterns in the point data compared to the lower power values.

Cross-validation was applied on multiple power values (p = 0.5 - 20) and compared to see which would minimise the RMSE. A power value of $p = 1.5$ was found to minimise RMSE for total antibiotic prescribing rate in 2016 (figure 5.5).

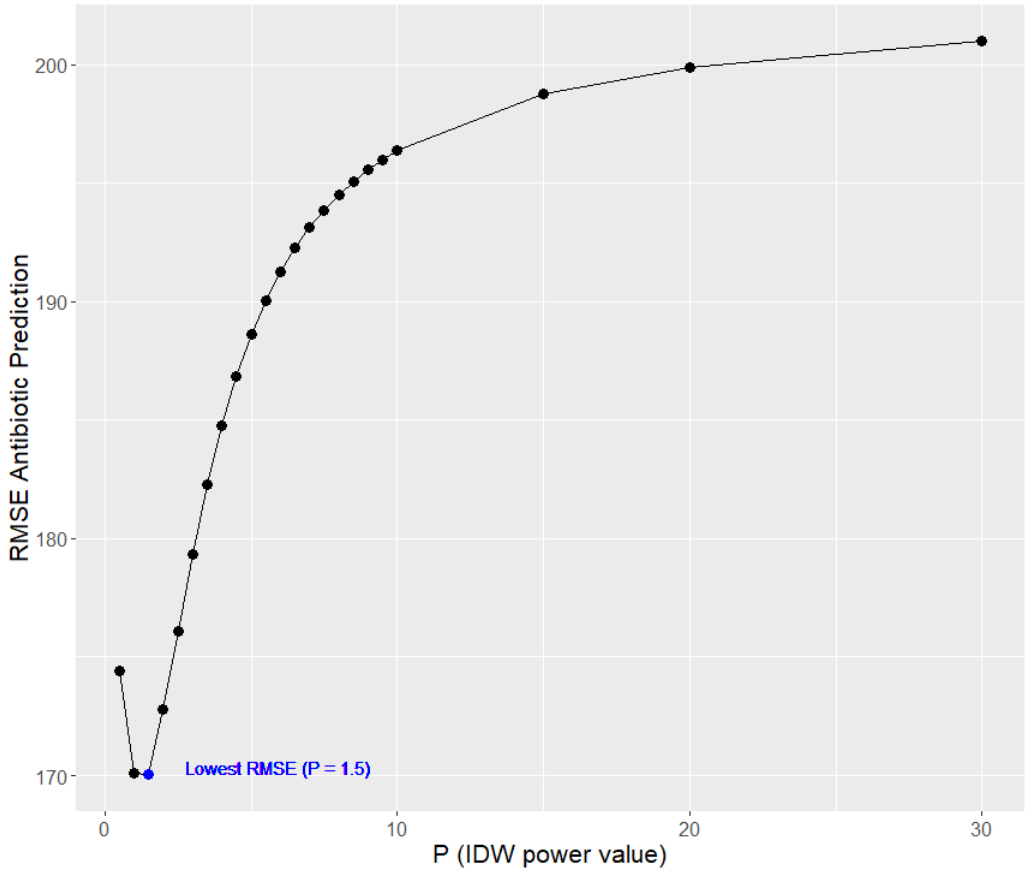


Figure 5.5: RMSE for varying IDW power values (p = 0.5 - 30) for Total GP antibiotic prescribing 2016 (items/1000 registered patients).

Maps are shown for interpolated total, 4C and individual high-risk antibiotic rates in 2016 using a power value of p = 1.5, (figures 5.6, 5.7 and 5.8). Interpolated total antibiotic rates highlight areas of high antibiotic prescribing when compared to the point map, such as Peterhead in Aberdeenshire and some high prescribing areas in the central belt of Scotland (figure 5.6).

Interpolated 4C antibiotics prescribing rates showed some areas (particularly Scottish islands) with higher rates of combined high-risk antibiotic prescribing. A similar spatial pattern was seen for high-risk antibiotic prescribing by IZs when compared to the observed point maps, with low prescribing on the east coast of Scotland, near Dundee, which was also seen in GP prescribing rates (figure 5.7).

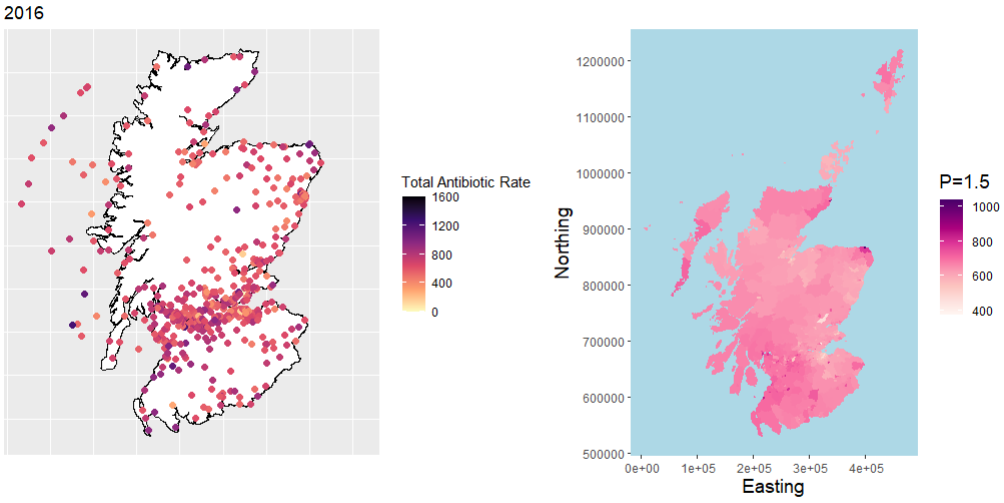


Figure 5.6: Areal map (right) of IDW with p = 1.5 compared to point map of observed GP antibiotic prescribing data in 2016 (left).

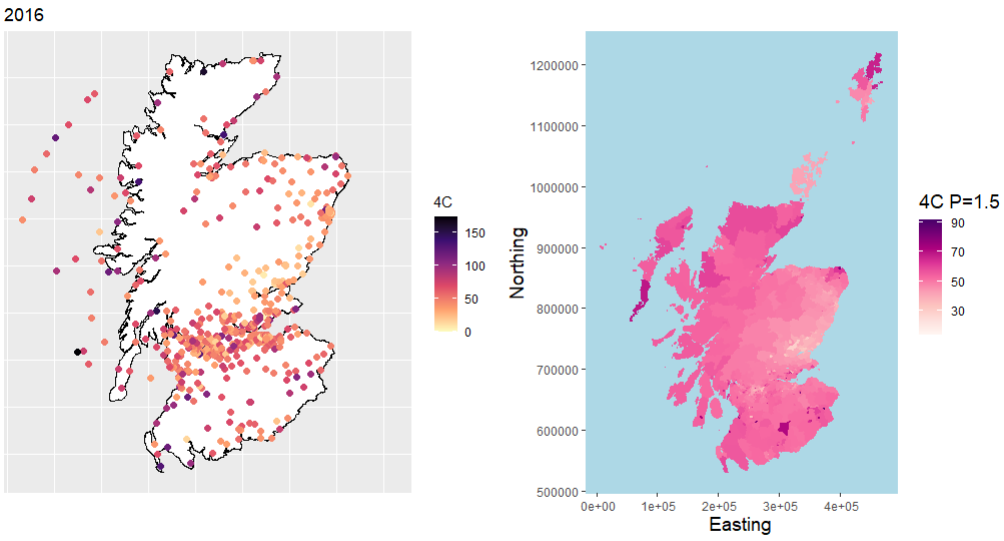


Figure 5.7: Areal map (right) of IDW with p = 1.5 compared to point map of observed combined high-risk GP antibiotic prescribing data in 2016 (left).

Figure 5.8 shows interpolated high-risk antibiotic separately for 2016 compared to the observed GP practice rates. The spatial patterns appear to be similar between predicted and observed data: Cephalosporins (top-left) shows areas of low prescribing in the Highlands and east coast, with higher prescribing rates in the north of Scotland. A similar spatial distribution was seen for quinolone (bottom-left) prescribing, with coamoxiclav (top-right) showing an overall higher rate of prescribing across Scotland and clindamycin (bottom-right) clearly showing as least prescribed antibiotic group.

The same process was applied for all antibiotic prescribing groups in 2017 and 2018 until a complete data set was created for all GP antibiotic prescribing rates by IZs. Comparing summary statistics (Median and IQR) of the interpolated values for all antibiotic groups for 2016, 2017 and 2018, in table 5.2, to the observed data (table 5.1), there was less variability in the interpolated data with narrower IQRs.

Table 5.2: Interpolated median (IQR) antibiotic prescribing rates from 2016 to 2018 for total antibiotics, 4C combined antibiotics, and individual high-risk antibiotic groups (cephalosporins, coamoxiclav, quinolone, and clindamycin) (items per 1000 registered patients) by Intermediate Zones

	Median (IQR) GP Antibiotic Prescribing Rates (items per 1000 registered patients)		
	2016	2017	2018
Total Antibiotics	654.8 (631.9 - 684.8)	603.0 (581.3 - 630.5)	603.0 (581.3 - 630.5)
4C Antibiotics	48.61 (43.82 - 52.25)	48.93 (46.67 - 51.18)	46.76 (44.25 - 48.99)
Cephalosporins	11.04 (8.65 - 13.62)	11.30 (10.17 - 12.36)	10.42 (9.32 - 11.50)
Coamoxiclav	20.80 (19.10 - 22.27)	21.58 (20.64 - 22.17)	20.64 (19.60 - 21.28)
Quinolone	14.90 (13.18 - 16.35)	15.09 (14.10 - 15.67)	14.52 (13.45 - 15.23)
Clindamycin	0.89 (0.72 - 1.03)	0.95 (0.86 - 1.09)	1.051 (0.88 - 1.27)

This was due to maximum and minimum prescribing values from the GP practice point data increasing the range, as seen in the point maps (figure 5.8). However, these values were smoothed out during interpolation, therefore narrowing IQR. The median values were similar across all antibiotic groups and overall the predicted antibiotic rates were within an expected range.

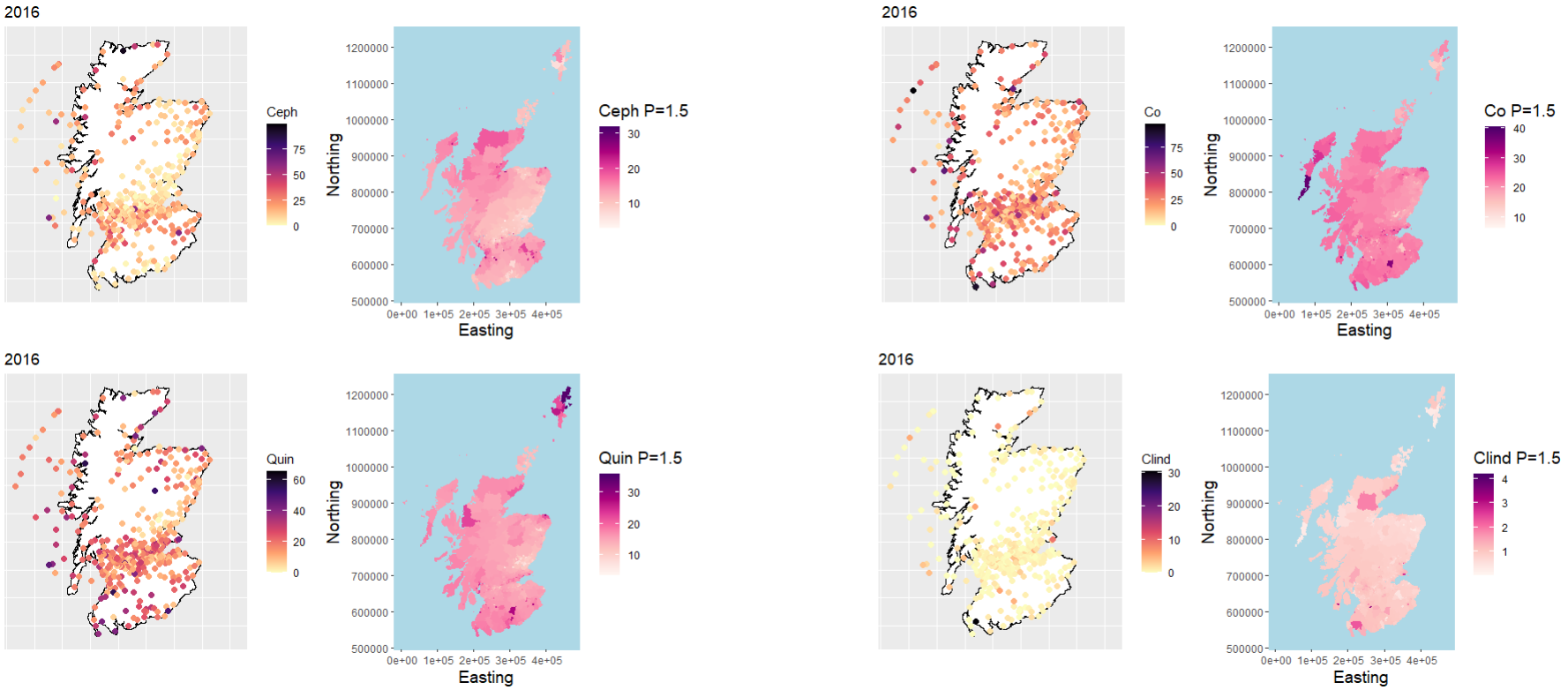


Figure 5.8: Areal maps of IDW with p = 1.5 compared to observed individual high-risk GP antibiotic prescribing data in 2016: Cephalosporins (top-left), Co-amoxiclav (top-right), Quinolones (bottom-left) and Clindamycin (bottom-right)

Cattle Density

Cattle Density was only collected for one year due to data accessibility, however, these analyses assumed that the spatial pattern and rates of agricultural farming areas would not have changed greatly over the time period in question.

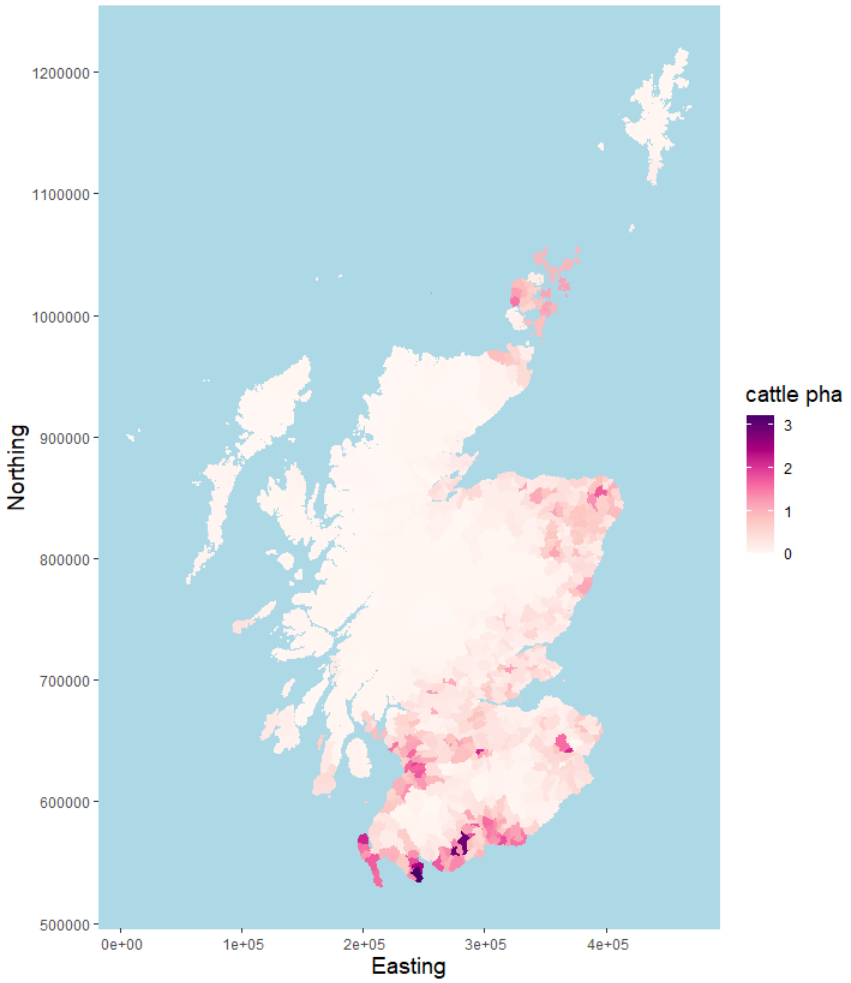


Figure 5.9: Areal map of cattle density by agricultural parish per ha.

There was a high density of cattle farming close to the Scottish Borders, the south-west of Scotland, and east coast (figure 5.9). Moran's test for spatial association showed cattle farming density to have strong positive autocorrelation with $I = 0.37$, $p < 0.001$.

The centroids of each agricultural parish were extracted and the data were converted to a spatial points data frame (figure 5.10).

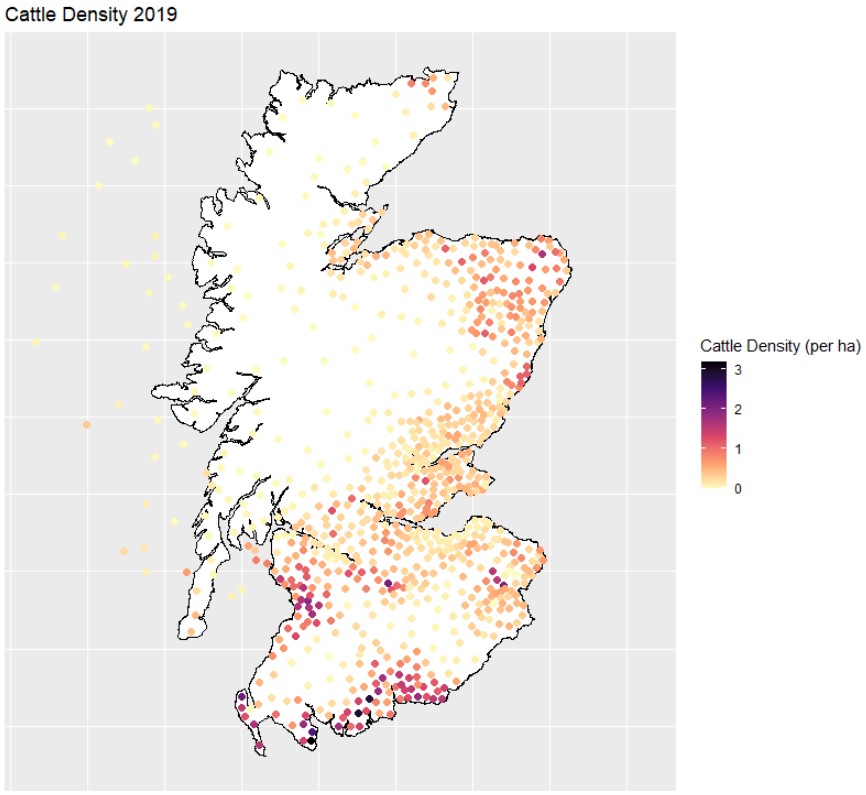


Figure 5.10: Point map of cattle density (per ha) transformed to Spatial Point Location Data from centroids of agricultural parishes.

Kriging Interpolation

As these data displayed evidence of strong spatial autocorrelation it was possible to apply kriging interpolation by fitting variogram models. Multiple variogram models were fitted to the data and assessed for best fit. This was a subjective process: the fit of the exponential, linear and spherical models were very similar, however, the spherical variogram model was selected as it appeared to fit the points at the beginning of the variogram slightly better than the others. This was important as the beginning of

the variogram represents the spatial auto-correlation between the closest points (figure 6.10).

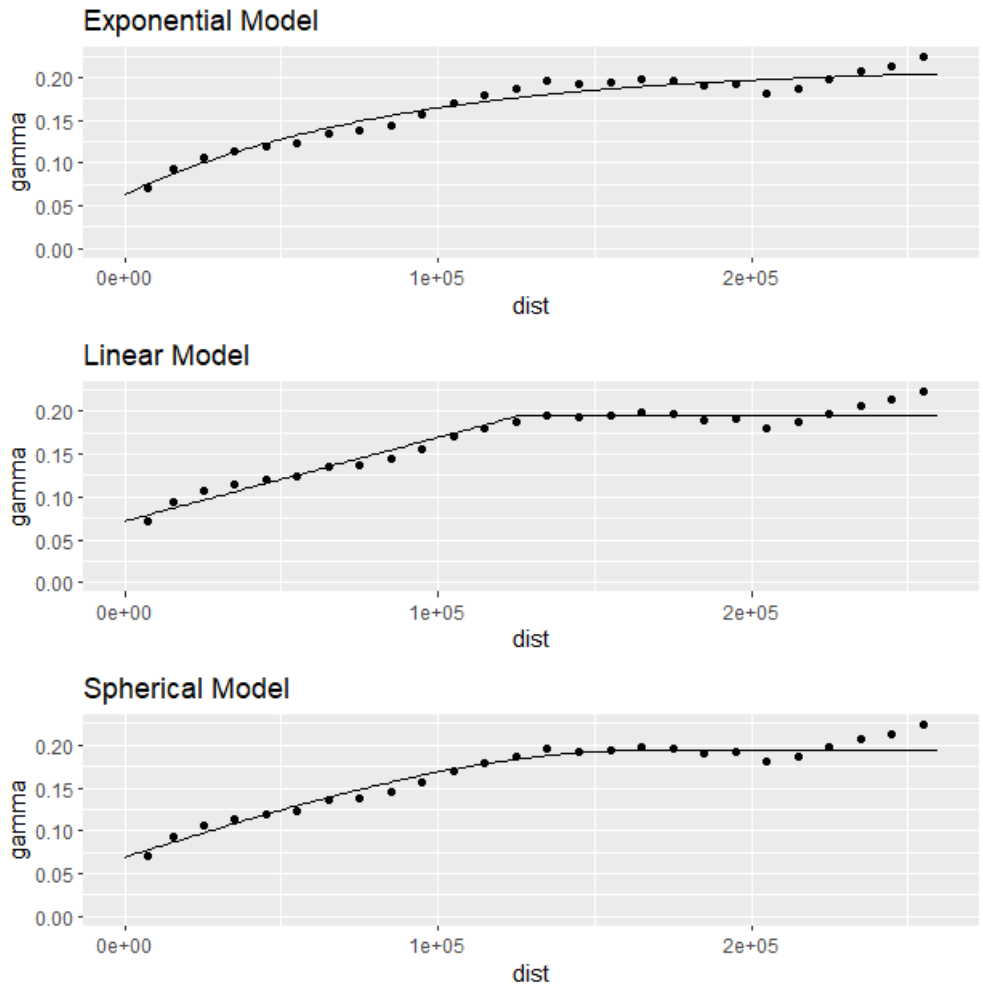


Figure 5.11: Cattle Density fitted variogram models: Exponential, Linear and Spherical

Spatial predictions were created using a kriging interpolation process and fitted using the spherical variogram to give a measure of cattle density (per hectare) for each IZ. Figure 5.12 compares the observed cattle density data by agricultural Parish (left) to the kriging predictions by IZs (right). The predictions showed a lower range of values and reduced variability compared to observed data.

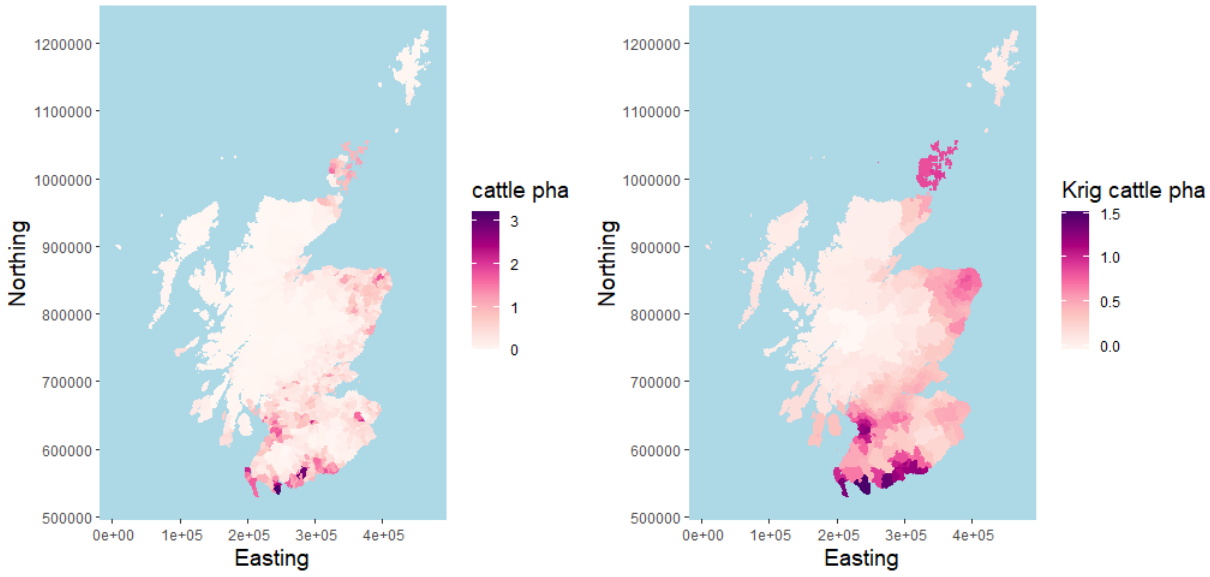


Figure 5.12: Areal maps of cattle density (per ha) Kriging predictions by Intermediate Zones (right) compared to observed data by Agricultural Parish (left).

Inverse-Distance Weighted Interpolation

Cattle density predictions for each IZ were then calculated using IDW interpolation to allow for the comparison of predictions. Multiple power values p were used to generate cattle density predictions and then plotted to compare with observed cattle density by parish (figure 5.13). Similar to the antibiotics data, low power values ($p = 0.1$) heavily smoothed the cattle density data over IZs and high values showed more variability ($p = 10$), however, the distribution appeared similar between plots for $p > 5$, with equal ranges (0.5 - 1.5) (figure 5.13).

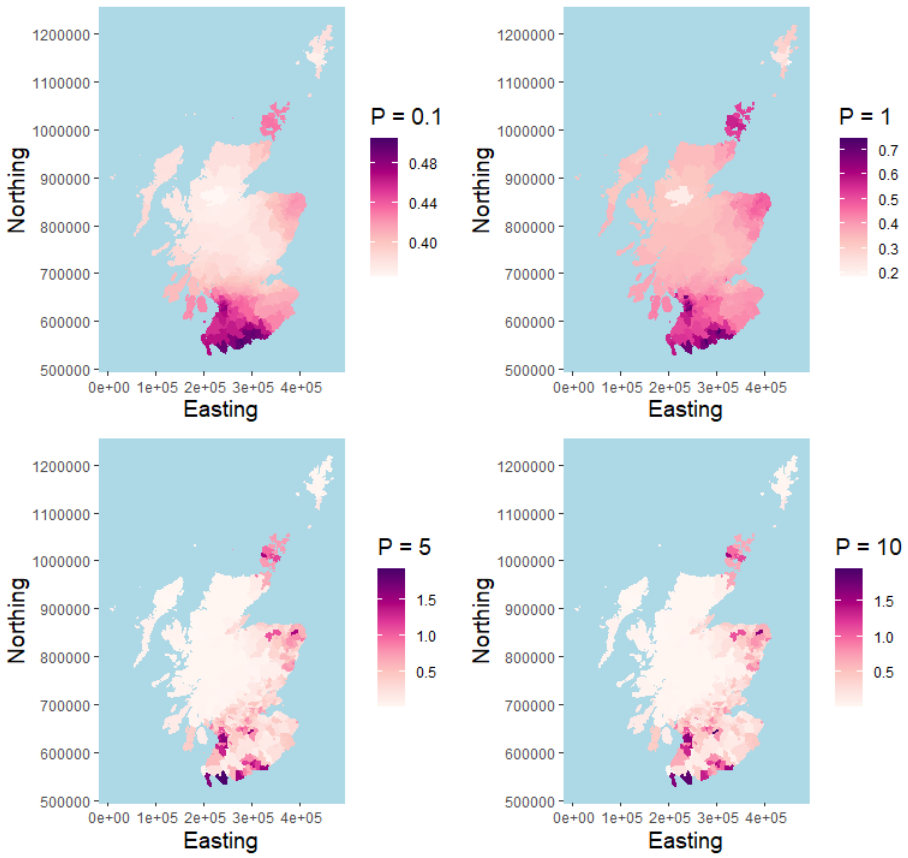


Figure 5.13: Areal maps of cattle density (per ha) IDW predictions for different powers: p = 0.1 (top-left), p = 1 (top-right), p = 5 (bottom-left), p = 10 (bottom-left).

Cross-validation results, comparing multiple power values (p = 0.5 - 30), showed $p = 3.5$ to minimise the RMSE (RMSE = 0.274). The RMSE was calculated using kriging predictions and showed a slightly higher RMSE (RMSE = 0.286). The IDW predictions were therefore taken forward to be run in the final spatio-temporal model.

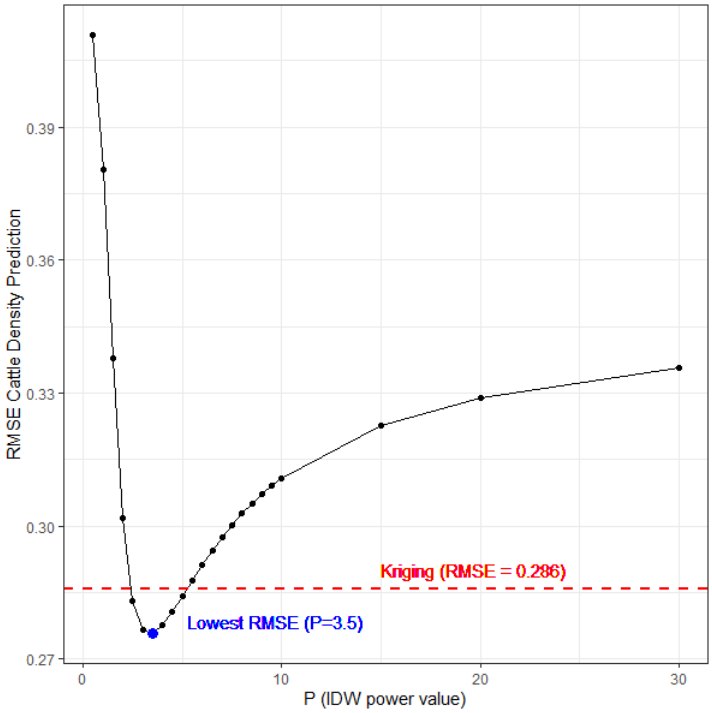


Figure 5.14: RMSE for varying IDW power values for Cattle Density (per ha).

5.3.2 Sensitivity Analysis

Comparing Methods of Interpolation

A sensitivity using univariate Poisson GLM's of CDI SIR was conducted for antibiotics variable predictions and cattle density predictions to compare the impact on model estimates of different values of IDW power value p, and kriging predictions (for cattle density only) (tables 5.3 and 5.4).

Table 5.3: GP total antibiotic prescribing sensitivity analysis comparing increasing power for Inverse-distance weighted interpolation. Univariate Poisson GLMs for total CDI SIR with RR and 95% confidence intervals.

GP Practice Total Antibiotic Prescribing		
	RR (95% CrI)	P-Value
IDW (P = 0.1)	1.0017 (1.0004 - 1.0031)	0.00957
IDW (P = 1)	1.0020 (1.0013 - 1.0026)	<0.001
IDW (P = 5)	1.0006 (1.0004 - 1.0009)	<0.001
IDW (P = 10)	1.0006 (1.0004 - 1.0008)	<0.001

Increased total antibiotic prescribing (items per 1000 registered patients) by IZ was associated with an increased risk of CDI for all values of p. The strength of association decreased with higher values of p (e.g. more variation), however, the p-values remained low and the confidence intervals narrowed. The direction of association remains the same across all models (table 5.3).

Cattle density also showed a positive association with risk of CDI. The strength of association was higher with lower power values (smoothed), whereas higher power values reduce the strength of associations and increased p-values with confidence intervals spanning 1. However, across all predictions (IDW and kriging), the direction of association remained the same for all models (table 5.4).

Table 5.4: Cattle density sensitivity analysis comparing kriging predictions against increasing power for inverse-distance weighted interpolation. Univariate Poisson GLMs for total CDI SIR with RR and 95% confidence intervals

Cattle Density Sensitivity Analyses		
	RR (95% CrI)	P-Value
Kriging Predictions	1.188 (1.087 - 1.297)	<0.001
IDW (P = 1)	2.060 (1.536 - 2.753)	<0.001
IDW (P = 3.5)	1.066 (0.992 - 1.145)	0.0797
IDW (P = 5)	1.058 (0.986 - 1.133)	0.116
IDW (P = 10)	1.053 (0.983 - 1.126)	0.136

This highlighted that the choice of p does affect the associations within models, with smoothed data tending to show stronger associations with CDI and more granular data showing reduce effect sizes, however the direction of association remains constant across all predictions therefore providing some confidence in the interpretation of risk factors.

Power Value for Antibiotic IDW Interpolation

The same power value, $P = 1.5$, was assumed for all antibiotic groups, which was determined by the power value that minimised the RMSE for total antibiotic prescribing in 2016. Cross-validation plots were created for each high-risk antibiotic to assess whether a value of $p = 1.5$ would also be appropriate to minimise the RMSE for cephalosporins, coamoxiclav, quinolones and clindamycin (figure 5.15).

As seen in figure 5.15, the power value of p = 1.5 does not exactly minimise the RMSE for all high-risk antibiotic groups, however, p = 1.5 does produces low RMSE's with the minimising power values all close to p = 1.5. The RMSE was minimised for quinolone prescribing for p = 1.5 with cephalosporins and coamoxiclav both showing very close values. Clindamycin prescribing showed the greatest deviation from the optimal minimising power value, however the RMSE still remained low for p = 1.5.

This sensitivity analysis concluded that the antibiotic and cattle density predictions made from the interpolation power values p = 1.5 and p = 3.5, respectively, would be taken forward to assess in multivariable spatio-temporal models.

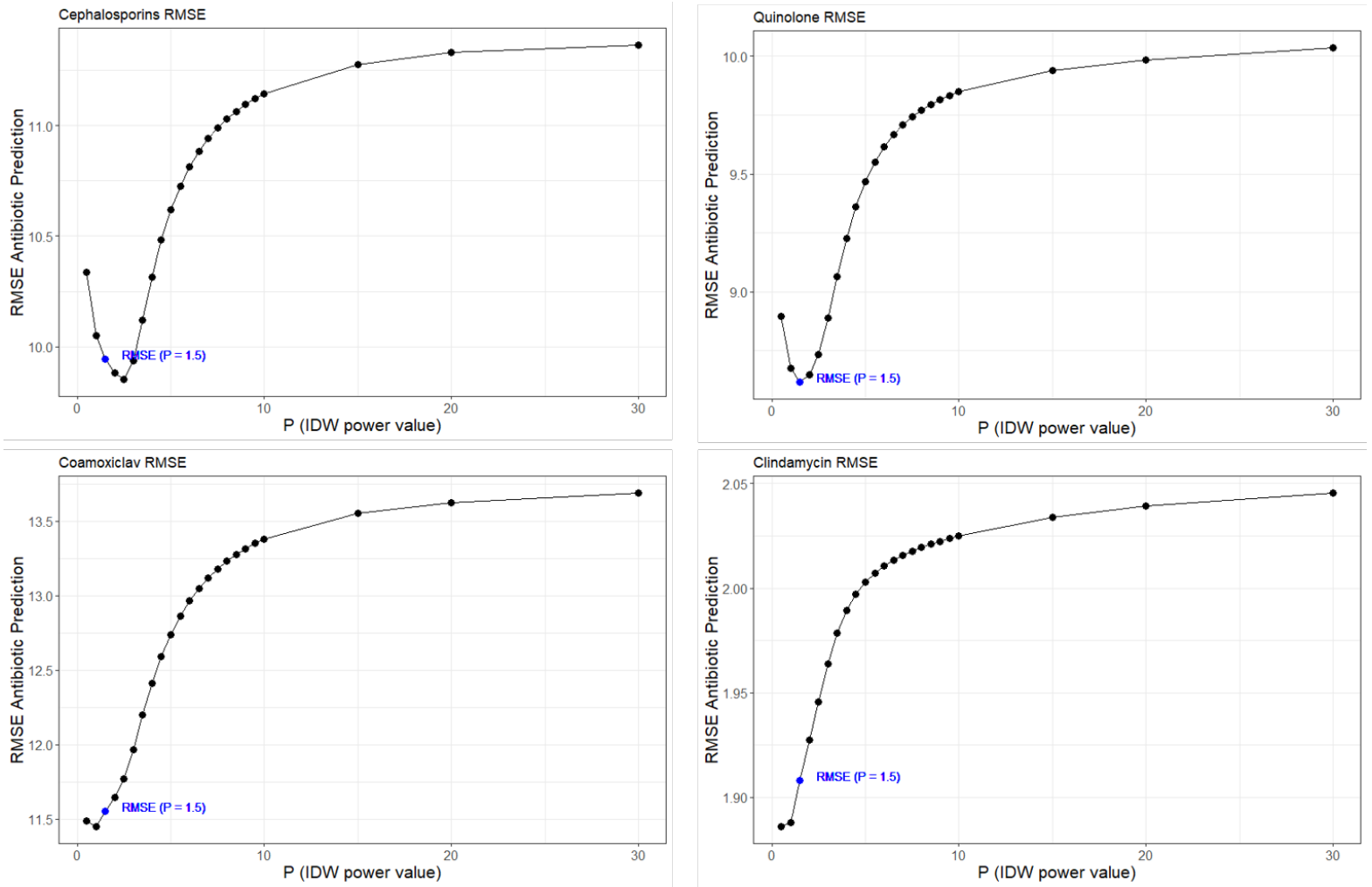


Figure 5.15: RMSE for varying IDW power values (0.5 - 30) for high-risk antibiotic groups compared to chosen power value (p = 1.5): cephalosporins (top-left), coamoxiclav (bottom-left), quinolones (top-right) and clindamycin (bottom-right).

5.3.3 Spatio-temporal AR(1) Model

GP Antibiotic Prescribing and Cattle Density as Risk Factors of CDI

Multivariable spatio-temporal models for total, community-acquired and hospital-acquired CDI were initially assessed including cattle density and total antibiotic prescribing, adjusting for percentage of IZ employment deprived and year. Combined 4C and individual high-risk antibiotics were then assessed, again, adjusting for cattle density, employment deprivation (%) and year.

The fully adjusted model showed cattle density to have a positive association with total CDI (RR = 1.105, 95% CrI 0.972 - 1.264), however, the credible interval (CrI) spanned 1. Total antibiotic prescribing (100 items per 1000 registered patients) also showed a positive association with increased risk of total CDI (RR = 1.020, 95% CrI 0.914 - 1.139), which again showed CrI spanning 1. There were no strong temporal differences between 2017 and 2018 compared to 2016. Employment deprivation (%) was seen to be positively associated with risk of total CDI with RR = 1.036, 95% CrI 1.029 - 1.043).

Stratified spatio-temporal models for CA-CDI and HA-CDI showed a strong positive association between cattle density and risk of CA-CDI (RR = 1.455, 95% CrI 1.188 - 1.774). HA-CDI had a negative association with cattle density (RR = 0.989, 95% CrI 0.848 - 1.147), however, CrI spanned 1. Total antibiotic prescribing showed a strong negative association with CA-CDI (RR = 0.819, 95% CrI 0.698 - 0.961) implying IZ's with high antibiotic prescribing were at a lower risk of CA-CDI, whereas HA-CDI tended towards a positive association (RR = 1.128, 95% CI 0.990 - 1.271) although CrI spanned 1. CA-CDI and HA-CDI both showed a positive association with employment deprivation (%), however, the effect size was stronger with HA-CDI. There were no

significant differences between 2017 and 2018 when compared to 2016 for CA-CDI and HA-CDI.

An intercept only model for total CDI estimated the spatio-temporal variability as $\tau = 0.667$. The spatial dependence parameter was estimated as $\rho_S = 0.543$, which implied moderate spatial dependence whereas the temporal dependence parameter was very low with $\rho_T = 0.043$. This essentially implied no temporal autocorrelation was present, however, there were only three temporally varying points in these models. The spatio-temporal variance parameter for the fully adjusted model was estimated at $\tau = 0.541$, therefore the covariates in the fully adjusted model accounted for approximately 19% of the spatio-temporal variability. The spatial dependence reduced to ($\rho_S = 0.385$) in comparison to the intercept model, whereas the temporal dependence remained very low and relatively unchanged $\rho_T = 0.042$ (Table 5.5). A spatio-temporal model only including employment deprivation (%) and years (2016 to 2018) showed $\tau = 0.544$, implying that cattle density and total antibiotic account for very little of the spatio-temporal variability.

Table 5.5: Multivariable Spatio-Temporal AR(1) GLM for Total CDI, HA-CDI and CA-CDI SIR compared to total antibiotic prescribing with risk ratios (RR) and 95% credible intervals (CrI).

	RR (95% CrI)		
	Total CDI	CA-CDI	HA-CDI
Cattle Density (per ha)	1.105 (0.972 - 1.264)	1.455 (1.188 - 1.774)	0.989 (0.848 - 1.147)
Total GP Antibiotics (100 items/1000)	1.020 (0.914 - 1.139)	0.819 (0.698 - 0.961)	1.128 (0.990 - 1.271)
Year: 2016	-	-	-
2017	0.980 (0.891 - 1.083)	0.786 (0.659 - 0.937)	1.074 (0.959 - 1.205)
2018	0.920 (0.828 - 1.029)	0.815 (0.688 - 0.971)	0.973 (0.868 - 1.091)
Employment Deprived (%)	1.036 (1.029 - 1.043)	1.022 (1.011 - 1.034)	1.041 (1.033 - 1.050)
tau	0.541 (0.428 - 0.658)	0.450 (0.172 - 0.724)	0.557 (0.416 - 0.711)
rho.S	0.385 (0.202 - 0.599)	0.175 (0.008 - 0.703)	0.341 (0.143 - 0.609)
rho.T	0.042 (0.002 - 0.165)	0.069 (0.003 - 0.336)	0.129 (0.010 - 0.333)

Fully adjusted models for 4c and high-risk antibiotic groups were then assessed, adjusting for cattle density, employment deprivation (%) and year. Combined 4C prescribing showed a negative association with total, community-acquired and hospital-acquired CDI, however, all CrI's spanned 1 (table 5.6).

Table 5.6: Adjusted RR and 95% Credible Intervals (CrI) for total antibiotics, 4C combined antibiotics, and individual high-risk antibiotic groups (cephalosporins, coamoxiclav, quinolone, and clindamycin) compared to Total CDI, CA-CDI and HA-CDI.

	RR (95% CrI)		
	Total CDI	CA-CDI	HA-CDI
4C Antibiotics (items per 1000)	0.996 (0.987 - 1.007)	0.990 (0.975 - 1.005)	0.999 (0.989 - 1.011)
Cephalosporins (items per 1000)	0.994 (0.974 - 1.015)	1.007 (0.974 - 1.038)	0.988 (0.965 - 1.011)
Coamoxiclav (items per 1000)	0.985 (0.960 - 1.010)	0.950 (0.911 - 0.987)	0.998 (0.970 - 1.026)
Quinolone (items per 1000)	0.999 (0.976 - 1.024)	0.982 (0.945 - 1.020)	1.010 (0.981 - 1.038)
Clindamycin (items per 1000)	1.101 (0.926 - 1.305)	0.539 (0.397 - 0.722)	1.361 (1.128 - 1.639)

For individual high-risk antibiotic prescribing: cephalosporins, coamoxiclav and quinolones all showed a negative association with total CDI. However, all CrI's contained 1, whereas clindamycin showed a positive association with total CDI (RR = 1.101, 95% 0.926 - 1.305). CA-CDI showed a strong negative association with coamoxiclav (RR = 0.950, 95% CrI 0.911 - 0.987) and in particular, clindamycin prescribing (RR = 0.539, 95% CrI 0.397 - 0.722), whereas HA-CDI showed a strong positive association with clindamycin prescribing (RR = 1.361, 95% 1.128 - 1.639) (table 5.6).

5.4 Discussion

This chapter explored the use of spatial interpolation methods to address a multilevel spatial data problem of routinely collected data, before conducting a spatio-temporal analysis of CDI incidence related to causal risk factors including primary care antibiotic prescribing and cattle density in Scotland.

These analyses showed a strong positive association between CA-CDI and cattle density by IZ in Scotland (RR = 1.455, 95% CrI 1.188 - 1.774). CA-CDI showed a decreased risk for 2017 and 2018 compared to 2016, whereas total CDI and HA-CDI showed negative estimates between years. However, credible intervals spanned 1. Total antibiotic prescribing was seen to have a negative association with CA-CDI (RR = 0.819, 95% CrI 0.698 - 0.961) which was also seen for some high-risk antibiotic groups: coamoxiclav (RR = 0.950, 95% CrI 0.911 - 0.987) and clindamycin (RR = 0.539, 95% CrI 0.397 - 0.722). This implied that increased prescribing in the community for total antibiotics, coamoxiclav and clindamycin were associated with a decreased risk of community-acquired CDI. Conversely, HA-CDI showed a strong positive association with clindamycin prescribing (RR = 1.361, 95% CrI 1.128 - 1.639). A positive association was also seen between total antibiotic prescribing (RR = 1.020, 95% CrI 0.914 - 1.139, 1000 items per 1000) and clindamycin prescribing (RR = 1.101, 95% CrI 0.926 - 1.305, items per 1000). For total CDI, however, the 95% CrI intervals spanned 1. The percentage of IZ population employment deprived was positively associated with an increased risk of total CDI, CA-CDI and HA-CDI.

Comparing the spatio-temporal model parameters for total CDI showed the fully adjusted model with cattle density, total antibiotics, employment deprivation (%) and year accounted for 19% of the spatio-temporal variability in comparison to the intercept only model. However, when assessed separately, antibiotic variables and cattle

density accounted for very little spatio-temporal variability with employment deprivation and year accounting for the majority of variation in total CDI. There was no temporal dependence estimated in the modelling of these data ($\tau = 0.042$), however, there were only three temporally varying points. The spatial dependence parameter reduced in the fully adjusted model for total CDI compared to the intercept model ($\tau_{intercept} = 0.543$ compare to $\tau_{adj} = 0.385$), however there was still some spatial association remaining. This implies that there was unaccounted for spatial association in the CDI data.

The impact of cattle density on CA-CDI was comparable with other studies that report environmental risk factors of CA-CDI. A study which presented a spatial analysis of CA-CDI showed increased proximity to livestock farming was associated with increased rates of CA-CDI after adjusting for population density and spatial clustering [53]. This study believed livestock farming to be a reasonable causal risk factor of CA-CDI when compared with other studies that report the presence of common strain of CDI in humans amongst farm animals such as pigs [170, 171], however, the presence of CDI has also been reported in cattle [177]. It has been seen that livestock animals such as piglets contain common human *c-difficile* isolates in their faeces [170] even prior to any administered antibiotics in the animals. The use of antibiotics in livestock animals is likely to have an impact the presence of *c-difficile* in animals [53]. From the available research, it is thought that this the first study to show cattle density as a risk factor of CA-CDI in Scotland with the novel use of spatio-temporal modelling of CDI.

Antibiotics have a well-defined relationship with the risk of CDI. Total antibiotics was seen to have a positive association with total CDI (RR = 1.020, 95% CrI 0.914 - 1.139, 100 items per 1000), however, CrI spanned 1. Clindamycin prescribing was also seen to have a strong positive association with HA-CDI (RR = 1.361, 95% CrI 1.128 - 1.639). These results are comparable to Chapter 6, which presents a study of CDI incidence

related to antibiotic prescribing by GP surgeries in Wales: total antibiotic prescribing (1000 items per 1000 STAR-PU) was associated with total CDI incidence (inpatient and non-inpatient CDI cases combined) (RR = 1.1413, 95% CI 0.9714–1.3404). Clindamycin prescribing was also shown to have a positive association with total CDI incidence in Chapter 6 [1]. The negative associations seen for CA-CDI related to total antibiotic prescribing and prescribing of high-risk antibiotic groups (clindamycin and coamoxiclav) were unexpected and difficult to explain. Prescribing of clindamycin and coamoxiclav have well-defined causal relationships with increased risk of CDI, particularly with community-acquired CDI [58].

To model these data, cattle density and GP antibiotic prescribing were transformed from different spatial scales to match the IZ level CDI data. Comparing the methods of interpolation, inverse-distance weighted (IDW) interpolation was found to perform better than ordinary kriging interpolation when compared using RMSE. This result is supported by a comparable study of spatial interpolation methods for rainfall predictions which showed IDW interpolation to perform better in terms of prediction accuracy compared to ordinary kriging, particularly in the absence of a strong spatial correlation structures. This study highlighted the flexibility of IDW interpolation as it is not restricted to strongly spatially correlation data, however, kriging interpolation performed well for strong spatial correlation structures [105]. Due to the lack of spatial correlation, IDW was the only method assessed for the GP antibiotic prescribing data, however, cattle density did show strong spatial correlation and was assessed using ordinary kriging and IDW interpolation. The results showed IDW RMSE to be smaller than the kriging RMSE although the results were very close ($\text{RMSE}_{Krig} = 0.286$ and $\text{RMSE}_{IDW} = 0.274$). Kriging interpolation is a more statistically sophisticated method as it provides measures of uncertainty and therefore may be augured as the more suitable method. It is a noted limitation that IDW does not provide measures of prediction uncertainty, although not performed in these analyses this could have been obtained by simulating

values based on the IDW spatial predictions multiple times and obtaining the average, similar to a bootstrapping approach.

A sensitivity analysis showed that the chosen level of smoothing largely affected the magnitude of model estimates and p-values, however, the direction of association remained the same. Both methods of interpolation have subjective elements to them and, therefore, it is difficult to illuminate bias when using these methods. This study has shown that these methods are useful to preliminary highlight trends in the context of transforming levels of spatial data for analysis, however, the confidence surrounding model estimates should be handled with caution. This is an important point for the interpretation of model estimates in the above paragraph, highlighting the sensitivity of results.

Multi-level spatial data is a burden for many fields of study [178]. Causal inference and modelling of these data can be extremely beneficial in highlighting important ecological risk factors, however, data are frequently unavailable on the desirable scale and, therefore, often leave relationships unmeasured or unaccounted for. This is a particular problem in observational research of routinely collected data [179]. Interpolation methods are a popular means of transforming data, however, there are multiple methods available [178]. For example, the cattle density data in these analyses could have been estimated at IZ level by applying a CAR Leroux model and predicting cattle density estimates at IZ centroids. This method was not adopted for these analyses as this study aimed to maintain as much variability as possible from one spatial scale to another and therefore did not want to account for spatial variability prior to making predictions. It was also of interest to compare between interpolation methods.

In these analyses, GP antibiotic prescribing data have been treated as point-location data. However, a study proposing a process-convolution model for quantifying health inequalities in respiratory prescriptions in Scotland disputes GP practice prescription data being considered as geostatistical or point-location data [156]. This study highlights that GP practice prescribing data relate to a single geographical coordinate, however, the patient populations are representative of the surrounding areas and, therefore, are not of the typical form of single measure point-location data. Furthermore, in urban areas where there are multiple GP practices, patient populations overlap and therefore cannot be represented by areal-level data either. Therefore, this study introduces a novel Bayesian random weighting spatio-temporal process-convolution model, which represents spatial correlation as an adaptive spatial smoother which allows for geographically close data to be modelled with autocorrelation or as marginally independent, and ensures rigid distance decay is not enforced by using a random weighting scheme. Treating GP prescribing data as point-location data has been previously assumed in other spatial studies as a standard approach, and have shown strong results [150, 157], however, the work in this chapter might benefit from implementing this process-convolution model in the context of CDI incidence and GP antibiotic prescribing as a mode of methods comparison and validation of results [156].

It is, therefore, noted as a limitation of this study that the point-location methods applied to these data may not be optimally appropriate for these spatial data, hence the necessity for the sensitivity analyses. Another limitation of this study is that intermediate zone centroids were used for the interpolation prediction grid and therefore spatial predictions were made based on the distance from IZ centroids to GP practice locations and centroids of agricultural parishes. The centroid of the IZs will not represent the most populated areas within each IZ. GP practices are also mostly based in populated locations, therefore the closest GP practice to an IZ centroid is most likely not representative of the whole IZ population. Similar to cattle density data, where

both agricultural parish and IZ centroids have been assumed, the distances between these are most likely not representative of the locations of true holdings with cattle.

However, this chapter has presented a novel spatio-temporal analysis of CDI in Scotland assessing antibiotic and environmental drivers of CDI infection. Associations between CDI and these risk factors have been presented but should be interpreted with caution due to the potential biases introduced when transforming these data to intermediate zone level. These analyses have also highlighted the issues in handling mutlilevel spatial data in observational research of routinely collected data. Future work of these data include the application the spatio-temporal process-convolution model [156] and assessing model estimates to compare between methods of handling multilevel and incompatible spatial data.

Chapter 6

Incidence of *Clostridiods Difficile* Infection in Welsh GP Practices Associated to Primary Care GP Antibiotic Prescribing

6.1 Introduction

Surveillance of *c-difficile* infection data in Wales indicates that CDI rates are high in comparison with England and Scotland: 36.7 CDI per 100,000 population in Wales compared with 23.9 CDI per 100,000 in England and 30.1 CDI per 100,000 in Scotland (2017) [48, 180, 181]. CDI in hospitalized patients results in poorer patient outcomes, increased length of hospital stays and increased treatment costs [182, 35]. CDI has a significant effect on patient morbidity and mortality with 30 day all-cause mortality rates suggested to be between 9% and 38% with attributable mortality varying between 5.7% and 6.9% [183].

Current Welsh Government policy encourages the prudent and appropriate use of antibiotics [184]. Welsh primary care antibiotic prescribing rates have fallen in recent years. From 2013/14 to 2017/18, Welsh primary care prescribing saw a 11.9% reduction in the total volume of items dispensed [134]. Stewardship of particular broad-spectrum antimicrobials associated with a high risk of CDI (e.g. in particular co-amoxiclav, cephalosporins, quinolones) is recommended to reduce the number of patients predisposed to CDI and lower transmission rates [63]. The four broad-spectrum antibiotics targeted by stewardship programmes under the group collectively called the '4C antimicrobials' are cephalosporins, clindamycin, ciprofloxacin and co-amoxiclav [64].

The epidemiology of *c-difficile* is clearly complex; observed patterns of disease may be due to the individual or combined effects of (i) outbreaks in healthcare settings, (ii) exposure to environmental sources of *c-difficile* and (iii) triggering of recently acquired or long-term colonisation by exposure to factors such as antibiotics that disrupt the gut microbiota.

The primary objective of this study was to understand the impact of total and high risk Welsh GP antibiotic prescribing on total CDI incidence, with a secondary aim of stratifying CDI incidence by inpatient and non-inpatient cases.

6.2 Data

This is a retrospective ecological study of the incidence of CDI across Wales between the financial years 2014–15 and 2017–18, including all cases of laboratory-confirmed CDI, from routine surveillance data collated by Public Health Wales every financial year [185] and linked to aggregated rates of antibiotic prescribing in the GP surgery at which the patient was registered.

6.2.1 Data sources and linkage

Clostridioides difficile Infection (CDI)

All glutamate dehydrogenase (GDH)-positive/toxin-positive CDI cases reported to the national surveillance system for *c-difficile* infection were provided by Public Health Wales including, when available, the GP surgery at which the case was registered. Following linkage of patients to practices and subsequent linkage of relevant practice-level data, the data were anonymised prior to analysis, including anonymising the practice and health board in which the practice was based. Classification of patients into 'inpatient', 'non-inpatient' or 'unknown' was part of routine Public Health Surveillance activity using the following definitions: an 'inpatient' was associated with a sample submitted from a hospital inpatient location, irrespective of specimen timing in relation to admission date. A 'non-inpatient' originated from a non-inpatient setting (GP, hospital, AE or admission units, with no assessment being made of time elapsed since admission). For patients with more than one positive sample in a financial year, the first positive result for any one patient was selected. CDI cases were excluded if the patients were registered with a practice outside Wales or the practice was unknown.

The number of CDI cases for patients registered at each of the Welsh GP surgeries was aggregated for each financial year (2014–15, 2015–16, 2016–17, 2017–18). The GP location for each patient was obtained from the Welsh demographic system using the patient's NHS number and GP surgery, allocated by Public Health Wales and the NHS Wales Informatics Service (NWIS).

Antibiotic Prescribing

Rates of antibiotic prescribing by practice were obtained from the Welsh pharmacy database Comparative Analysis System for GP Prescribing Audit (CASPA). Data were collated for all antibiotics and separately for those classes considered a high risk for CDI (cephalosporins, quinolones, co-amoxiclav and clindamycin). Total antibiotics were collated as items per 1000 specific therapeutic group age/sex-related GP prescribing units (STAR-PU); separate classes were collated as items per 1000 patients registered at the practice. The STAR-PU is a measure weighted to reflect the age and gender mix of the practice and specific drug type [134]. Antibiotic GP prescribing was collated by financial year. Rates of PPI prescription were obtained as items per 1000 registered patients.

GP Practice Demographics

Attributes relating to GP surgeries and rates of CDI were collated independently for each of the four financial years. Practice population size data were obtained from the CASPA GP prescribing database for each financial year. Mergers and changes in practice structure over the 4 years explained the practice population variation during the analysis period. Data regarding the percentage of patients aged $\geq$ 65 years, in each practice was obtained from the Public Health Wales Observatory. The social deprivation of each practice was estimated by the percentage of the practice population in the most deprived 40% of Wales's lower super output area using the 2014 version of the Welsh indices of multiple deprivation. The 2014 figures were also used for subsequent study years as these data were only available for 2014 and it is unlikely that there would be substantial changes over the subsequent 3 years.

The level of co-morbidities of the practice populations was estimated from disease and risk behaviour-specific prevalence rates reported as part of the general medical services contract Quality and Outcome Framework (QOF). These data were available separately for each of the four periods, where available. The following practice level prevalence indicators were included in initial analyses: patients with COPD; patients ever diagnosed with established hypertension; and patients at least 17 years old diagnosed with a specified diabetes and by type. QOF data were generally available for 2014–15 and 2015–2016 but less so for 2016–17, and no data available for 2017–18 due to relaxed NHS Wales data capture requirements [186]. Where data were missing for later years, previous practice measures were imputed using a last-one-carried-forward (LOCF) method [187]. This assumes that practice population indicators remain similar between years.

Data Linkage

The number of CDI cases associated with each of the GP practices across Wales were aggregated for each of the four financial years (2014-2015, 2015-2016, 2016-2017, 2017-2018, April 1st to March 31st) to obtain a count of CDI cases per GP practice. The total number of inpatient and non-inpatient CDI cases per GP practice was also available for each financial year.

The high-risk antibiotic data were already aggregated by GP practices and financial years, therefore could be directly linked to the restructured CDI data by the anonymised GP practice key. PPIs and total antibiotics (per STAR-PU) were provided in a separate data set which were then linked to the CDI and high-risk antibiotic data. GP demographic and co-morbidity data were provided for each financial year. These data were combined into a four-year data set then linked to the CDI data set by the anonymised GP practice key.

The data were then in the appropriate format for these analyses including aggregated counts of CDI cases (total, inpatient and non-inpatient) by financial year; antibiotic prescribing for total antibiotics (items per 1000 STAR-PU) and four high-risk antibiotics (items per 1000 registered patients); GP practice list size; percentage of practice population over 65 (%); residing in the most deprived lower super output area (%); with COPD (%), with hypertension (%) and diabetes (%).

6.2.2 Ethics and Data Storage

The study did not require ethical approval as it used data on CDI cases per GP surgery and publicly available GP practice level data, and did not use any identifiable individual patients. In addition, GP surgery identity and health board were anonymised within the analysis data set. Data are stored on a secure server at University of Bristol with access limited to a small number of study team members. This process was cleared by the Public Health Wales Information Governance team. Public Health Wales collates C. difficile data on an all-Wales basis for ongoing surveillance purposes.

6.3 Statistical Methods

6.3.1 Descriptive Statistics

Initially, a trend analysis was conducted to explore the relationship between Welsh GP antibiotic prescribing and CDI incidence. Scatter plots were assessed and generalised additive models (GAMs) were then examined to further understand the relationships between variables. (See Chapter 2, equation 2.16). GAMs are particularly suitable for assessing this type of noisy data, where the underlying relationships are often unclear [188]. The GAMs were modelled to assess if a linear assumption would be appropriate

to assume for the relationship between CDI incidence and each of the antibiotic variables.

CDI incidences (cases per 1000 registered patients) were presented over the four financial years with 95% confidence intervals using Byar's approximation for total, inpatient and non-inpatient CDI. Byar's approximation gives accurate estimations to exact Poisson probabilities, even for data with small counts. Therefore, this was suitable for these data [189].

Median (IRQ) GP antibiotic prescribing rates were presented for all antibiotic variables (total and individual high-risk) over the four financial years. The change in total antibiotic prescribing over time, for each GP practice, was then visualised using a Sankey plot [145]. To create the plot GP practice total antibiotic prescribing data were categorised for each year into "low, "medium" and "high" prescribers, defined by the tertiles of the earliest financial year (2014/15). This provided visual representation of how practice prescribing has evolved over time. Line plots were created to visualise Welsh health board CDI incidence over the four financial years for total, inpatient and non-inpatient CDI. This allowed individual health board trends to be assessed.

6.3.2 Generalised Linear Model (GLM)

The assumed distribution for the CDI data is Poisson, however, the data were found to be over-dispersed and the strict Poisson assumption of equal mean and variance did not hold. A negative-binomial model was used as an alternative method for over-dispersed data as it does not require the mean and variance to be equal with the addition of a dispersion parameter. See Chapter 3, section 3.2.2.

Negative-binomial regression models investigated the association between practice level GP antibiotic prescribing and CDI incidence in Welsh GP surgeries. Primary analysis assessed the association between total CDI incidence (inpatient/non-inpatient cases combined) and total and high-risk practice level antibiotic prescribing. Natural log transformations were applied to prescribing rates for co-amoxiclav, cephalosporins, clindamycin and quinolones, as the distribution of rates were heavily skewed, and therefore the units are expressed as log items per 1000 registered patients. The associations with potential confounding variables (financial year, health board, age, social deprivation, PPI use, COPD, diabetes and hypertension) were examined and adjusted for in regression models.

Here, y_{ijk} represents CDI counts (Total, Inpatient or Non-inpatient) with an *offset* for GP practice population (N_{ijk}) for each GP practice i, within health boards j and year k: $k = 1 - N_{years}$, $j = 1 - N_{HB}$ and $i = 1 - N_{GP(jk)}$ and N_{years}, N_{HB} and $N_{GP(jk)}$ equal the number of years, number of health boards and the number of GP practice within each year and health board. γ_j and α_k represent the contribution of fixed effects health-board and financial year, respectively.

$$E(\log(y_{ijk})) = \log(N_{ijk}) + \beta_0 + \gamma_j HB + \alpha_k Yr + \beta_1 Over65_{ijk} + \beta_2 Deprivation_{ijk} + \beta_3 PPI_{ijk} + \beta_4 COPD_{ijk} + \beta_5 Diabetes_{ijk} + \beta_6 Hypertension_{ijk} + \beta_7 Antibiotic_{ijk} \quad (6.1)$$

The variable $Antibiotic_{ijk}$ represents the addition of total antibiotics (items per 1000 STAR-PU) or one of the four high-risk antibiotic groups (co-amoxiclav, cephalosporin's, clindamycin, or quinolones, items per 1000 registered patients). The same model structure was applied for all analyses of total, inpatient and non-inpatient CDI. For the catergorised analyses, the same model structure was again used, replacing the $Antibiotic_{ijk}$

variable for a categorised version with four groups, with baseline group equal to quartile 1 (lowest prescribers).

A model including interaction terms between CDI case source (inpatient/ non-inpatient) and all other covariates was examined then backward selection was performed to identify any statistically significant terms at a 5% significance level. The results of these interaction tests motivated a secondary analysis of CDI incidence with stratification of inpatient and non-inpatient cases modelled against total and high-risk antibiotic GP prescribing, using the same model structure as 6.1. The interaction model is presented in equation 6.2 where I_{ijk} corresponds to the binary variable for CDI case source: inpatient ($I = 1$) or non-inpatient ($I = 0$).

$$E(\log(y_{ijk})) = \log(N_{ijk}) + \beta_0 + \beta_1 I_{ijk} + \beta_{2k} Year_{ijk} + \beta_{3j} HB_{ijk} + \beta_4 Over65_{ijk} + \beta_5 Deprivation_{ijk} + \beta_6 PPI_{ijk} + \beta_7 COPD_{ijk} + \beta_8 Diabetes_{ijk} + \beta_9 Hypertension_{ijk} + \beta_{10} Antibiotic_{ijk} + I_{ijk} : (\beta_{11k} Year_{ijk} + \beta_{12j} HB_{ijk} + \beta_{13} Over65_{ijk} + \beta_{14} Deprivation_{ijk} + \beta_{15} PPI_{ijk} + \beta_{16} COPD_{ijk} + \beta_{17} Diabetes_{ijk} + \beta_{18} Hypertension_{ijk} + \beta_{19} Antibiotic_{ijk}) \tag{6.2}$$

An additional analysis then assessed the association with a categorised form of total and high-risk, antibiotic GP prescribing. Each antibiotic group was categorised by quartiles of prescribing (quartile 1 to quartile 4) and then modelled, adjusting for GP practice demographics and co-morbidities for total, inpatient and non-inpatient CDI. Boxplots were used to visualize differences in both inpatient and non-inpatient CDI incidences by quartiles of GP prescribing level. This secondary analysis is a replication of the original analysis using the numerical values of the prescribing levels. It was carried out

to investigate if there was any indication that prescribing in the highest quartile was associated with significantly more risk that the lowest quartile and also to investigate any potential non-linearity.

6.4 Results

The following sections presents the exploratory trend analysis followed by descriptive analysis of the data set and finally the results from the negative-binomial GLMs. Scatterplots of the data compared to CDI incidence (cases per 1000 registered patients) highlighted the noise in the data.

There appeared to be no association between deprivation and CDI incidence, however, an increasing percentage of practice population with COPD, hypertension and diabetes showed some positive trends with CDI incidence. Similarly, increasing PPI prescriptions and percentage of practice over 65 showed a slight increasing association with CDI incidence (figure 6.1).

Comparing total CDI incidence with each of the antibiotic groups showed widespread data and noisy trends which were difficult to interpret. There were a few very high prescribing GP practices for the high-risk antibiotic groups (figure 6.2).

Log transformations of the high-risk antibiotics pulled high-prescribing practices closer to the main body of data, however, they were still highly scattered. There was some evidence of a positive trends between CDI incidence and cephalosporins, coamoxiclav and quiniones, but not clindamycin. Log transformation remained for the rest of the analyses (figure 6.3).

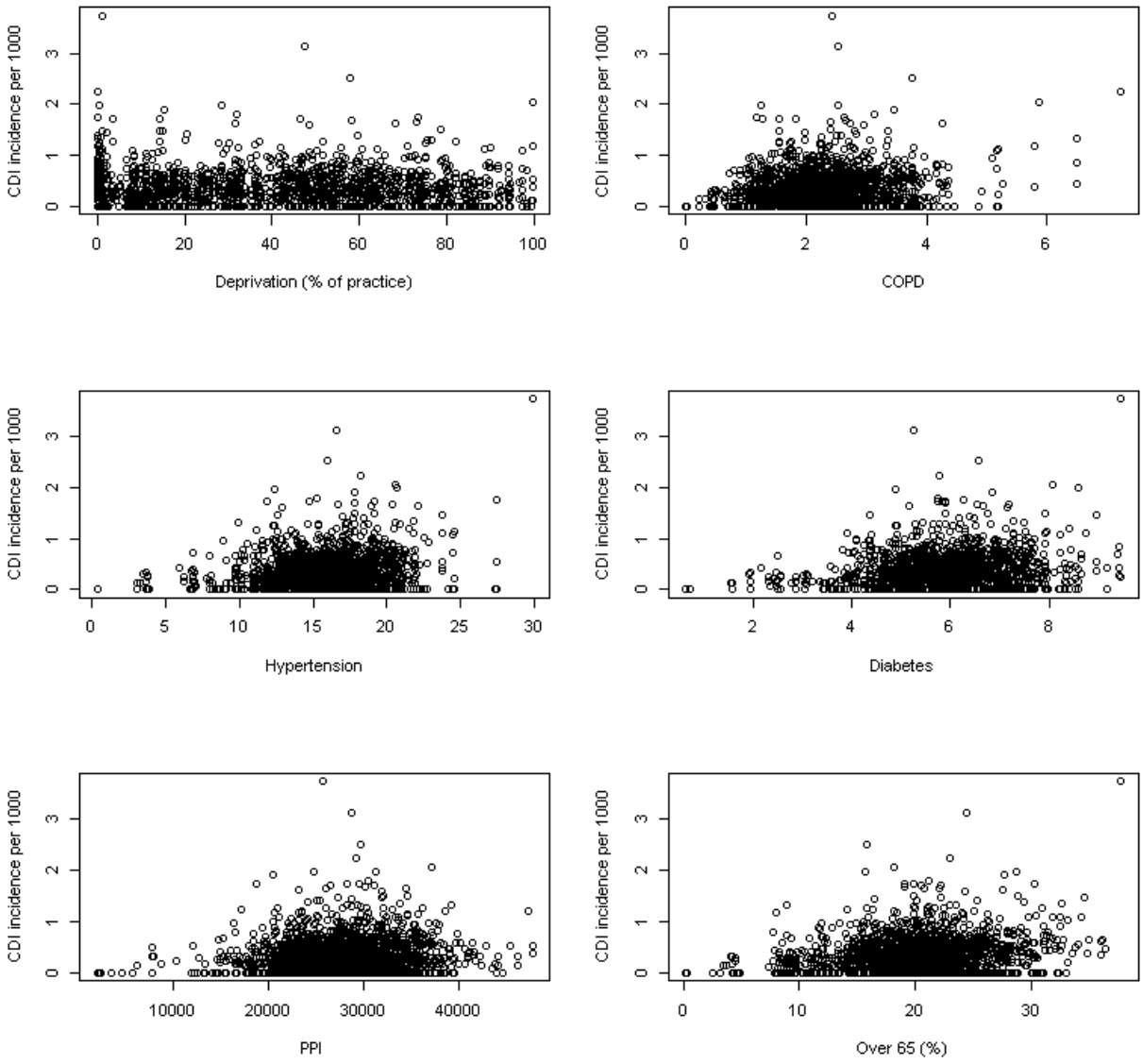


Figure 6.1: CDI incidence (per 1000 registered patients) against GP population deprivation (%) (top-left), COPD (%) (top-right), Hypertension (%) (middle-left), Diabetes (%) (middle-right), PPI (%) (bottom-left) and Over 65 (%) (bottom-right)

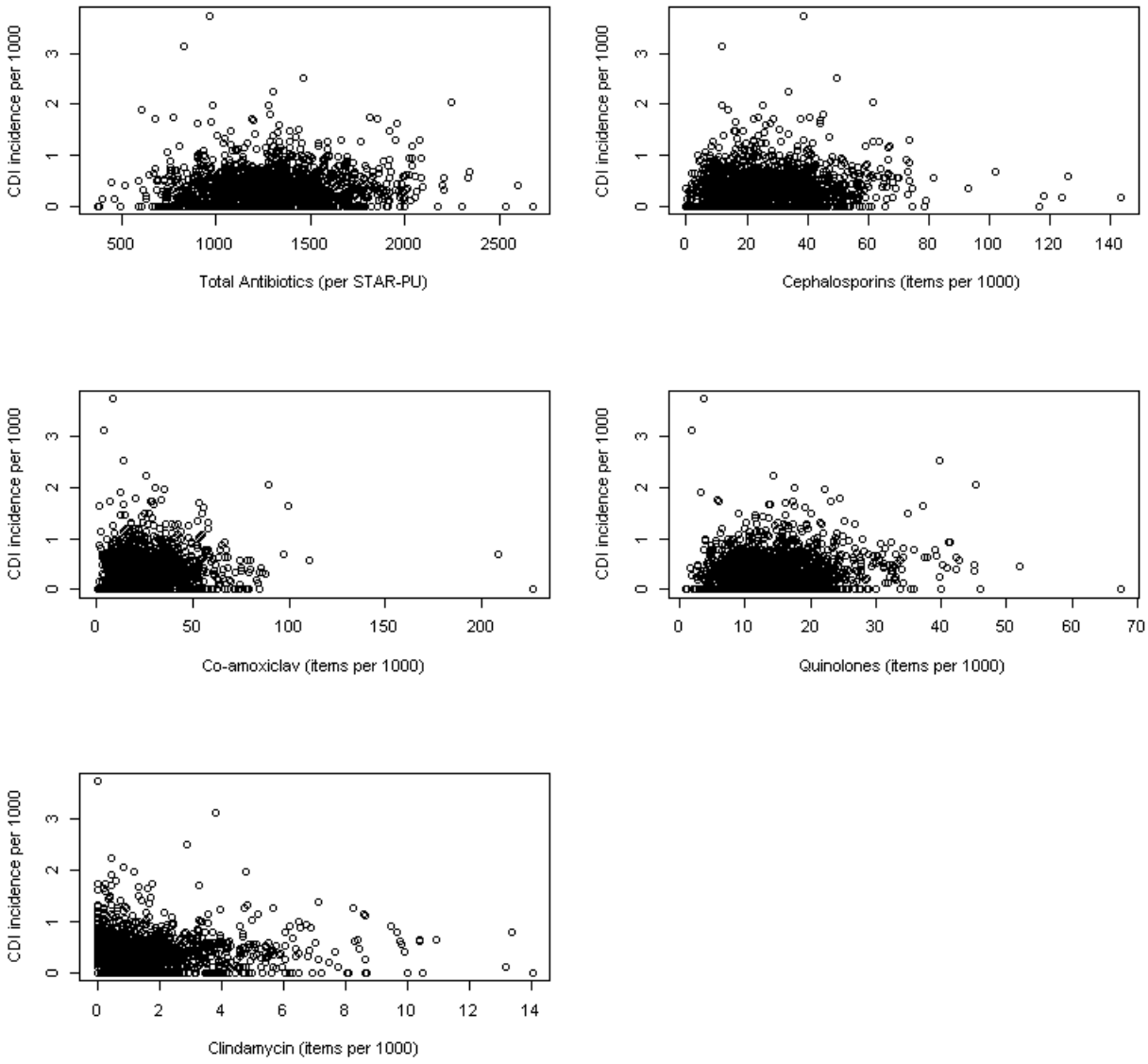


Figure 6.2: CDI incidence (per 1000) against Total (top-left) and High risk (Cephalosporins (top-right), Co-amoxiclav (middle-left), Quinolones (middle-right) and Clindamycin (bottom)) antibiotic prescribing

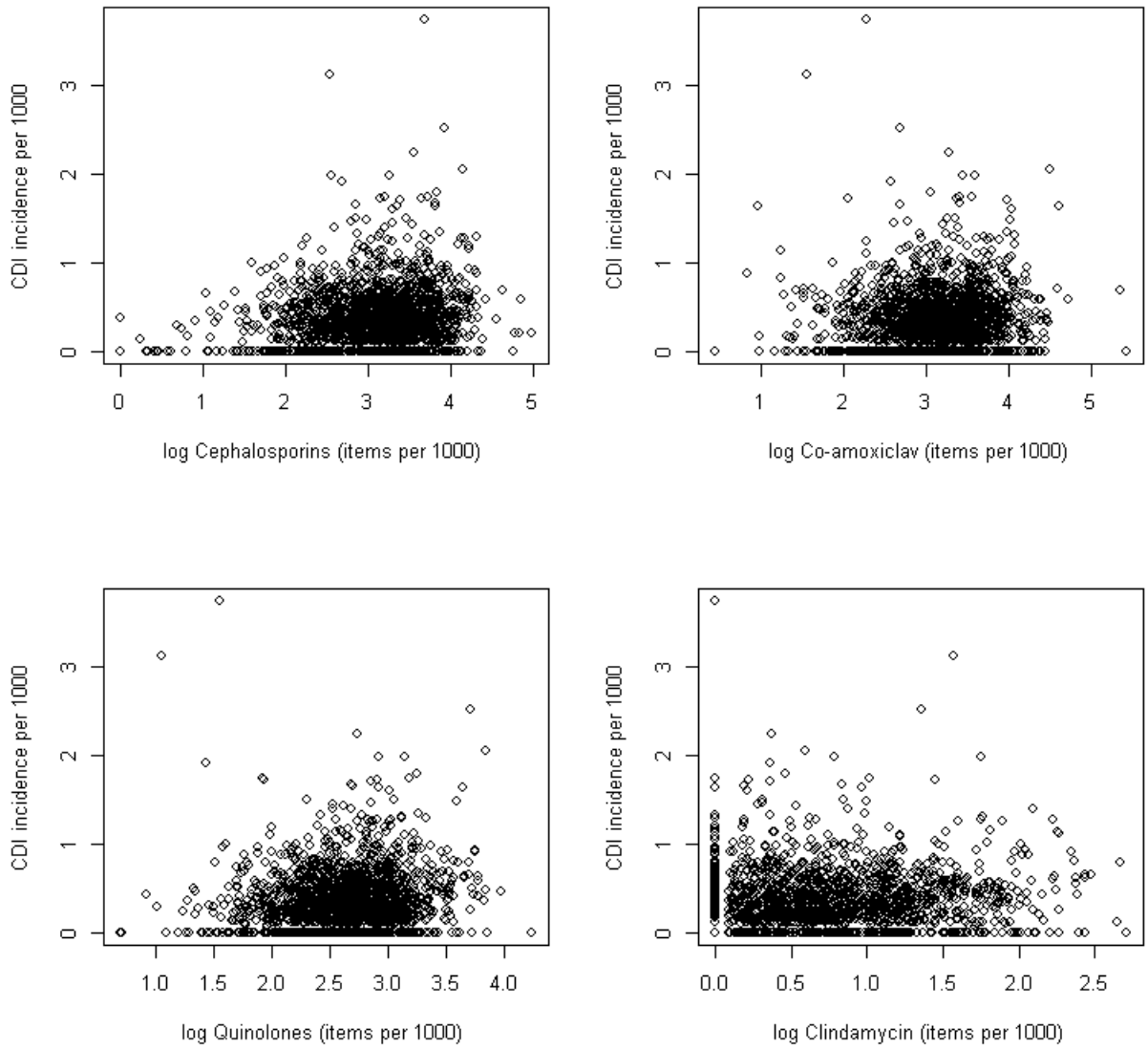


Figure 6.3: CDI incidence (per 1000) against log transformed high risk (Cephalosporins (top-left), Co-amoxiclav (top-right), Quinolones (bottom-right) and Clindamycin (bottom-left) antibiotic prescribing

Generalised Additive Models (GAMs)

The GAMs were created to aid the interpretation of trends existing within the data, however, were used only for exploratory purposes. The GAMs were set to be fully flexible to allow influential data points to surface. Plotting the GAMs allowed the visualisation of any non-linearity between each antibiotic variable and incidence of CDI.

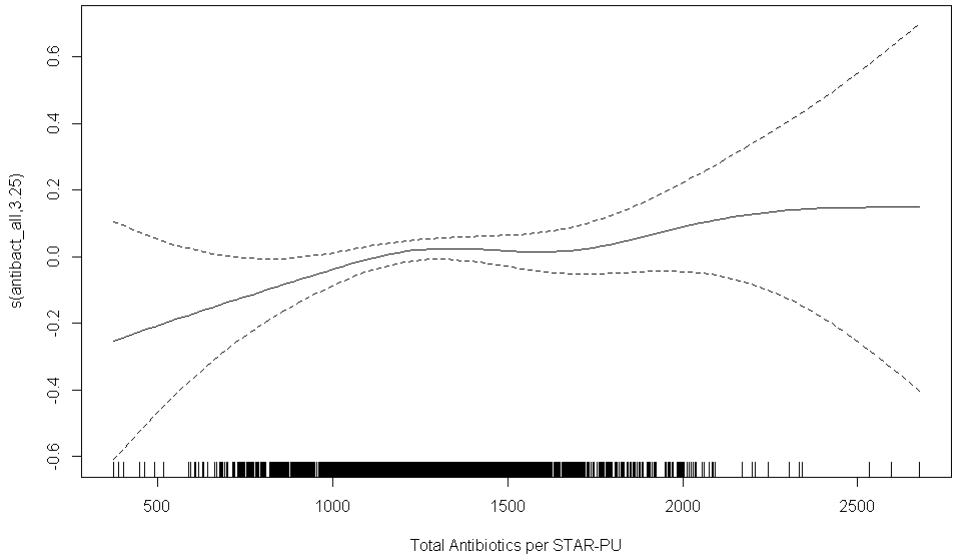


Figure 6.4: CDI incidence (per 1000 registered patients) against total antibiotic prescribing (items per 1000 STAR-PU).

Total antibiotics trend with CDI incidence showed increasing CDI in practices with high total antibiotic use. There did not appear to be an obvious non-linear trend and was suitable to be modelled with a linear assumption. The shaded black line along the x-axis indicated where the majority of data lay (figure 6.4).

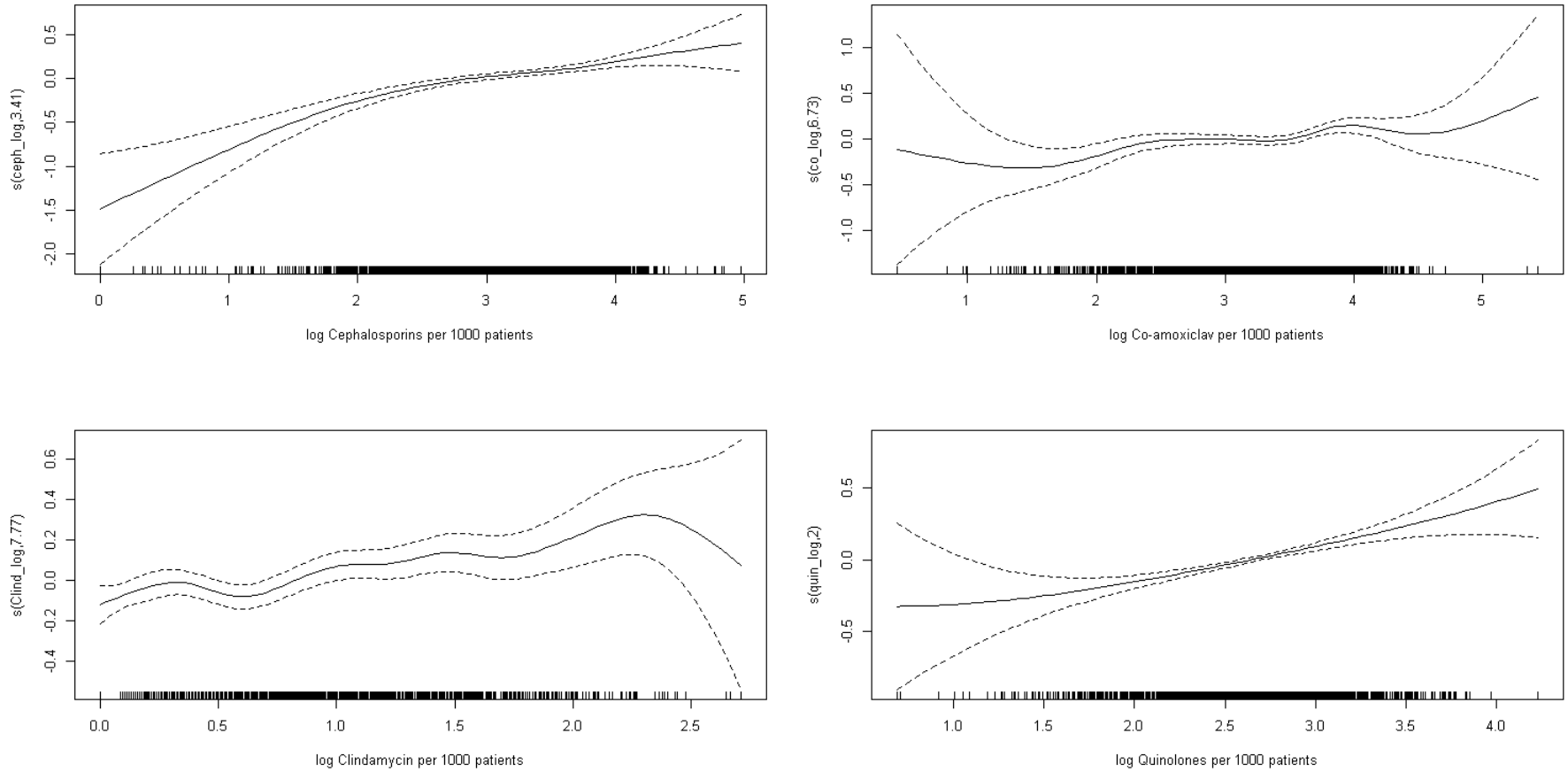


Figure 6.5: CDI incidence (per 1000 regsitered patients) against log transformed high risk (Cephalosporins (top-left), Co-amoxiclav (top-right), Quinolones (bottom-left) and Clindamycin (bottom-right) antibiotic prescribing.

The GAMs plots for log transformed high risk antibiotic prescribing compared to CDI incidence showed all four antibiotic groups with slight positive trends. None of the GAM plots highlight clear non-linear trends, therefore these data were suitable to be modelled under a linear assumption (figure 6.5).

6.4.1 Descriptive Analysis

There were 4613 total confirmed CDI cases from 2014–15 to 2017–18 linked to GP surgeries over seven Welsh health boards, serving a population of over 3 million: incidence was 1.44 (95% CI 1.40–1.48) cases per 1000 registered patients. The number of inpatient and non-inpatient cases were 2580 (61.8%) and 1763 (38.2%) with incidences (95% CI) per 1000 patients of 0.89 (0.86–0.92) and 0.55 (0.53–0.58). There was a 15.6% decrease in total CDI incidence from 2014–15 to 2017–18, from 0.40 (0.38–0.43) to 0.34 (0.32–0.36) cases per 1000 registered patients. However, 2017–18 showed a 7.6% increase in CDI incidence relative to 2016–17. The inpatient and non-inpatient split of CDI cases remained largely unchanged throughout the study period (table 6.1).

Table 6.1: CDI incidence (Inpatient, Non-inpatient and Total, cases per 1000 registered patients) by financial year with 95% confidence intervals.

CDI cases	Financial Year			
	2014-15	2015-16	2016-17	2017-18
Total	0.403 (0.381 - 0.425)	0.381 (0.360 - 0.403)	0.316 (0.297 - 0.336)	0.340 (0.320 - 0.361)
Inpatient	0.257 (0.240 - 0.275)	0.232 (0.215 - 0.249)	0.195 (0.180 - 0.211)	0.206 (0.191 - 0.222)
Non-inpatient	0.146 (0.133 - 0.160)	0.149 (0.136 - 0.163)	0.121 (0.110 - 0.134)	0.134 (0.122 - 0.147)

Median total antibiotic prescribing rates fell by 11.3% from 2014–15 to 2017–18 and consistently decreased each year (table 6.2). Prescribing rates for co-amoxiclav, cephalosporins and quinolones also decreased during this time with co-amoxiclav and cephalosporins displaying similar median prescribing rates throughout (29.22 and 27.15 items per 1000 in 2014–15 then 20.65 and 18.45 items per 1000 in 2017–18), whereas the rate of quinolone prescribing was lower (14.49 items per 1000 in 2014–15 then 12.39 items per 1000 in 2017–18). In comparison with all other high-risk antibiotic groups, clindamycin

was rarely prescribed, however, it was the only antibiotic group seen to increase each year (0.75 items per 1000 in 2014–15 then 1.04 items per 1000 in 2017–18) (table 6.2).

Table 6.2: Median antibiotic prescribing rates (Total, Co-amoxiclav, Cephalosporins, Clindamycin, Quinolones, items per 1000 STAR-PU) with IQR by financial year.

Antibiotics	Financial Year			
	2014-15	2015-16	2016-17	2017-18
Total	1353 (1157 - 1557)	1278.6 (1081.6 - 1453.6)	1230.4 (1049.3 - 1392.1)	1199.8 (1043.7 - 1363.0)
Co-amoxiclav	29.22 (19.20 - 41.26)	23.81 (15.76 - 33.97)	21.59 (14.73 - 30.14)	20.65 (14.04 - 27.95)
Cephalosporin	27.15 (18.27 - 40.27)	23.65 (14.74 - 34.59)	20.81 (12.12 - 29.70)	18.45 (11.04 - 27.86)
Clindamycin	0.75 (0.27 - 1.81)	0.88 (0.31 - 1.84)	0.97 (0.44 - 2.08)	1.04 (0.39 - 2.02)
Quinolones	14.49 (10.77 - 19.12)	13.13 (9.62 - 17.33)	12.97 (9.59 - 16.73)	12.39 (8.82 - 16.77)

The overall (2014/2015 to 2017/2018) total GP antibiotic prescribing rates varied greatly between surgeries and years; the median (IQR) prescribing rate was 1272 (1077 – 1450, items per 1000 STAR-PU) with minimum of 374 items per 1000 STAR-PU and maximum of 2677 items per 1000 STAR-PU. Overall high-risk antibiotic prescribing also varied between surgeries and years, with minimum GP prescribing rates <1 item per 1000 for all four high-risk antibiotics. Co-amoxiclav, cephalosporin, clindamycin and quinolone median (IQR and min–max) prescribing rates were 23.3 (15.6–33.7 and 0.6–226.8), 22.2 (13.4–33.8 and 0.00–143.7), 0.9 (0.4–2.0 and 0.00–14.1) and 13.3 (9.6–17.5 and 1.0–67.6) items per 1000 registered patients, respectively.

A Sankey plot was created to show the change in total antibiotic prescribing behaviours between years in Welsh GP practices the over four financial years. The proportion of GP practices prescribing 'High' total antibiotics have decreased over the four financial years, with the proportion of 'Low' prescribers also steadily increasing. The proportion of 'Medium' prescribers reduced between 2014/2015 and 2017/18, however, the difference was not as substantial as 'High' and 'Low' categories. There was also a proportion of practices moving from Low to Med and Med to High between financial years. However, this figure largely indicates a decrease in high prescribing GP practices (figure 6.6).

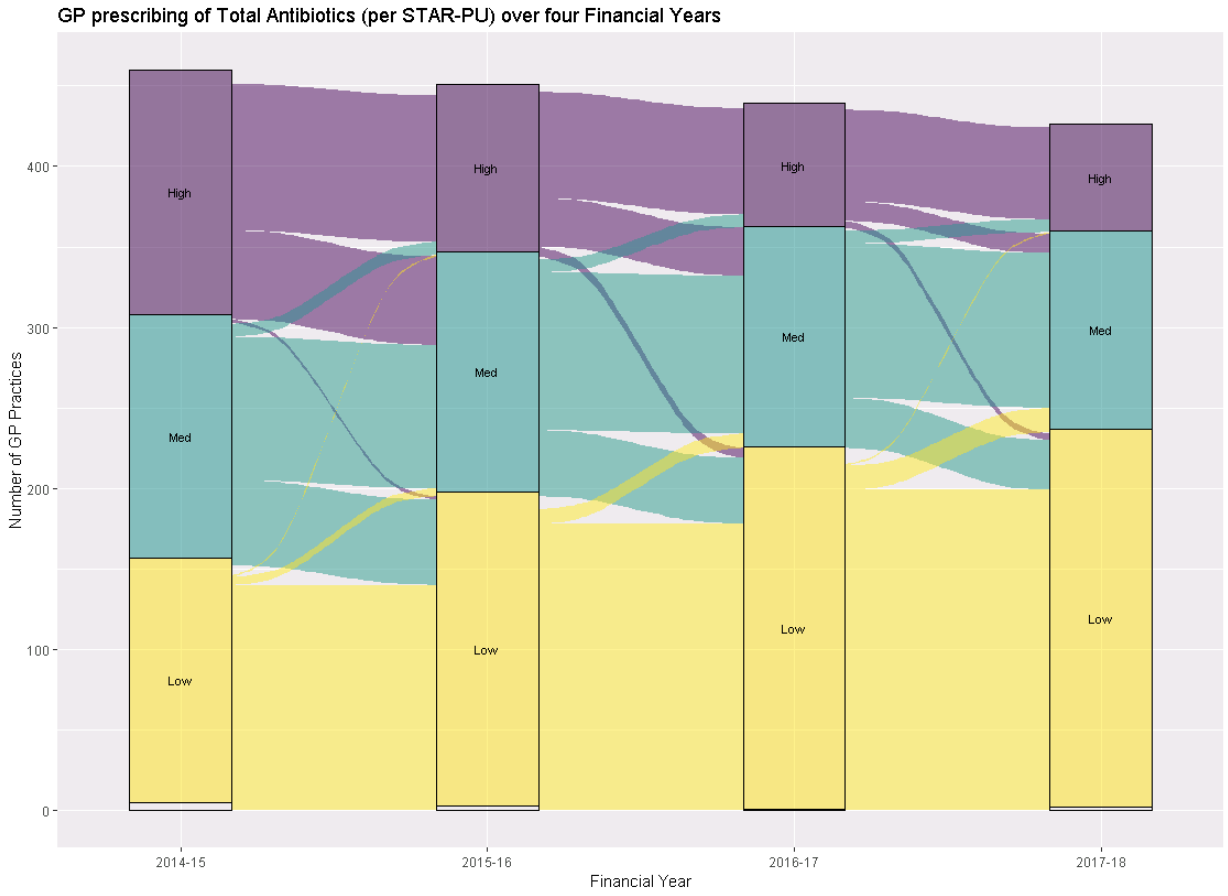


Figure 6.6: GP prescribing of Total Antibiotics (per 1000 STAR-PU) over four financial years: categorised into Low, Med or High prescribing determined by tertiles of 2014-15 rates.

Comparing total CDI incidence rates between Welsh health boards showed a decreasing trend from 2014-15 and 2016-2017 for the majority of health boards, however, all show an increase between 2016-17 and 2017-18 except for health board 4 (figure 6.7).

Separating by inpatient and non-inpatient CDI incidence showed that CDI incidence varied between health boards and over time. However, this also highlighted differences between inpatient and non-inpatient CDI trends. For inpatient CDI, the majority of health boards show a relatively constant trend over the financial years with two health boards (1 and 5) showing decreasing trends in inpatient CDI over time. Non-inpatient CDI incidence also showed varying trends between health boards: health

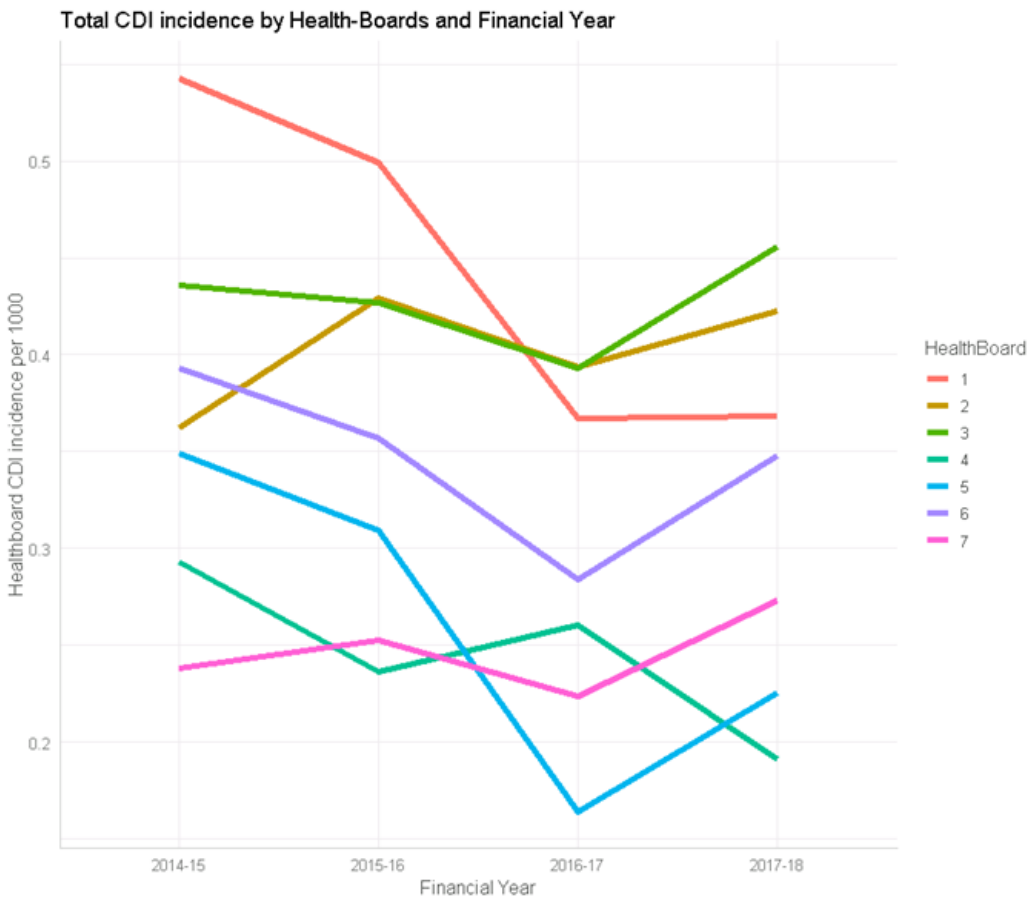
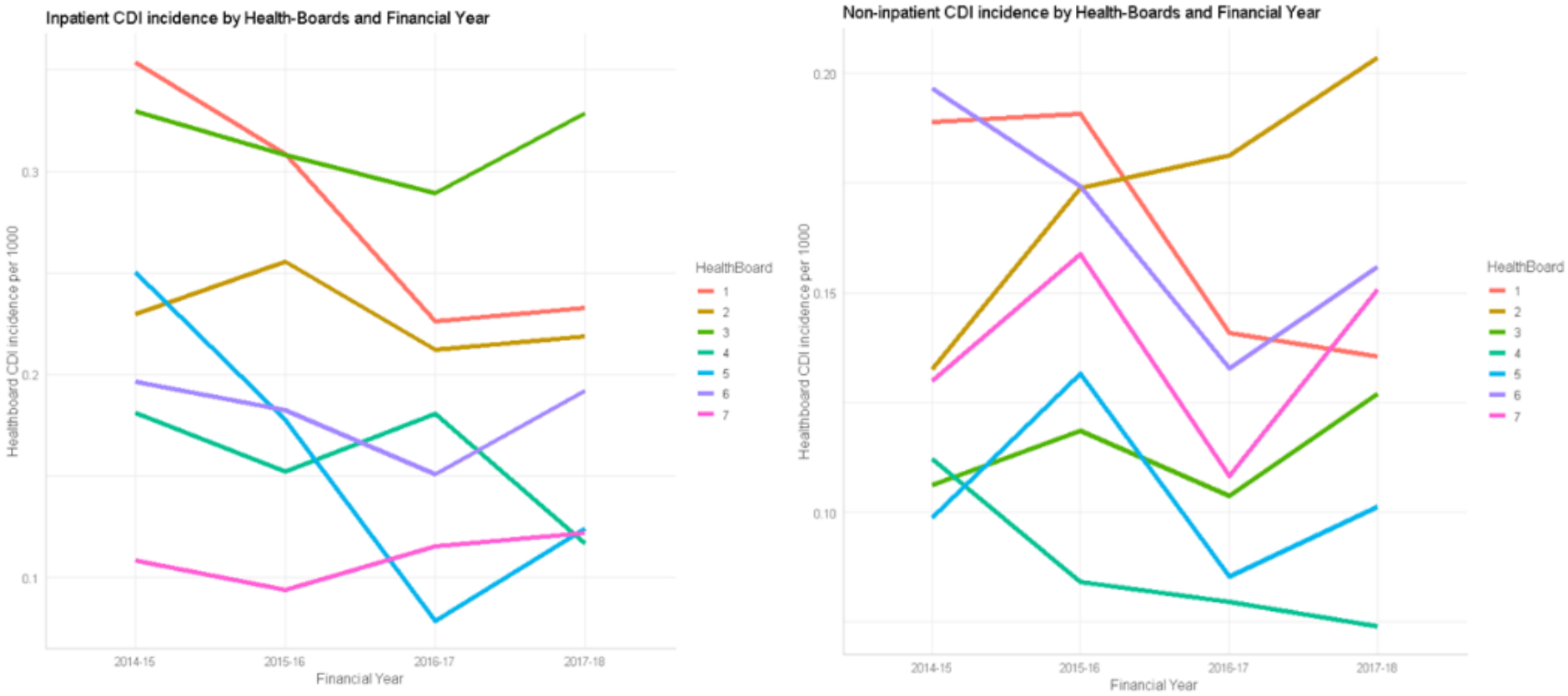


Figure 6.7: Mean health-board total CDI incidence per 1000 registered patients across four financial years.

board 2 showed a strictly increasing trend in CDI over time whereas health boards 3 and 7 varied between years, with an increase in non-inpatient CDI between 2016/17 and 2017/18 for all health boards except 1 and 4 (figure 6.8).



(a) Mean inpatient CDI incidence

(b) Mean non-inpatient CDI incidence

Figure 6.8: Health-board inpatient (a) and non-inpatient (b) CDI incidence per 1000 registered patients by financial year (2014/15 to 2017/18).

6.4.2 Total CDI Incidence

Unadjusted models for the primary analysis showed increased risk of total CDI incidence with increasing percentage of practice population comorbidities [for a 1% increase in the percentage of the practice population with COPD, the relative risk (RR = 1.173, 95% CI 1.120 – 1.230); with diabetes, RR = 1.152 ((95% CI, 1.114 – 1.191); and with hypertension, RR = 1.055 (95% CI, 1.043 – 1.068)]. An increased risk in total CDI was also associated with increasing percentage of patients aged ≥ 65 years [RR (95% CI) = 1.037 (1.029–1.045)] and PPI prescribing (per 1000 items per 1000 patients) [RR (95% CI) = 1.018 (1.011–1.025)] (table 6.3).

Table 6.3: Unadjusted and Adjusted RRs for Total CDI incidence compared to Total Antibiotic prescribing (items per 1000 STAR-PU), with 95% CI and p-values (from adjusted model).

	Unadjusted	Adjusted	
	RR (95% CI)	RR (95% CI)	P
Total Antibiotics (per 1000 items per 1000 STAR-PU)	1.3375 (1.1696–1.5293)	1.1413 (0.9714–1.3404)	0.108
Financial Year:			
2014-15	1	1	-
2015-16	0.9453 (0.8518 - 1.0490)	0.9397 (0.8518 - 1.0365)	0.214
2016-17	0.7901 (0.7095 - 0.8798)	0.7823 (0.7056 - 0.8672)	<0.001
2017-18	0.8388 (0.7541 - 0.9330)	0.8257 (0.7455 - 0.9144)	<0.001
Health boards (1-7)	-	-	<0.001
Patients 65 (%)	1.0370 (1.0293 - 1.0447)	1.0269 (1.0134 - 1.0407)	0.001
Social Deprivation Score: (%) most deprived	0.9995 (0.9980 - 1.0009)	1.0002 (0.9980 - 1.0025)	0.856
Proton Pump Inhibitor (per 1000 items per 1000 patients)	1.0175 (1.0105–1.0246)	0.9967 (0.9877–1.3404)	0.481
COPD (%)	1.1732 (1.1196 - 1.2292)	1.0594 (0.9906 - 1.1322)	0.093
Diabetes (%)	1.1518 (1.1138 - 1.1913)	1.0609 (1.0023 - 1.1227)	0.039
Hypertension (%)	1.0550 (1.0426 - 1.0676)	1.0051 (0.9841 - 1.0264)	0.639

*Unadjusted models include only one predictor variable (univariate analysis). Adjusted Models are adjusted for Financial Year, Health-board, (%) Patients *geq* 65, Social Deprivation Score - (%) most deprived, Proton Pump Inhibitor (per 1000 STAR-PU), (%) COPD, (%) Diabetes and (%) Hypertension. The global p-value for health boards is presented as the health bards are not identified.

Total CDI incidence was associated with higher total antibiotic prescribing [RR (95% CI) = 1.338 (1.170–1.529) per 1000 items per 1000 STAR-PU] (table 6.3). High-risk antibiotic classes were also seen to be positively associated with total CDI incidence; co-amoxiclav, clindamycin, cephalosporins and quinolones presented a 12.0%, 14.9%, 24.6% and 28.0% increase in risk of CDI per unit increase in log items per 1000 registered patients (table 6.4).

Table 6.4: Unadjusted and Adjusted RR of Total CDI incidence compared to rates of predefined high-risk antibiotic groups with 95% CI and P-values (from adjusted models)

	Unadjusted	Adjusted	
High risk antibiotic GP prescribing	RR (95% CI)	RR (95% CI)	P
Co-amoxiclav (log items per 1000)	1.1200 (1.0451 - 1.2005)	1.0803 (0.9993 - 1.1682)	0.054
Cephalosporin (log items per 1000)	1.2456 (1.1715 - 1.3250)	1.0638 (0.9858 - 1.1485)	0.111
Clindamycin (log items per 1000)	1.1485 (1.0707 - 1.2319)	1.0787 (1.0014 - 1.1618)	0.046
Quinolones (log items per 1000)	1.2798 (1.1690 - 1.4015)	1.0125 (0.9180 - 1.1168)	0.805

*Adjusted Models are adjusted for Financial Year, Health-board, (%) Patients *geq* 65, Social Deprivation Score - (%) most deprived, Proton Pump Inhibitor (per 1000 STAR-PU), (%) COPD, (%) Diabetes and (%) Hypertension. The global p-value for health boards is presented as the health bards are not identified.

The fully adjusted model for the primary analysis showed diminished effects and wider 95% CIs for total CDI incidence with total antibiotic prescribing [RR (95% CI) = 1.141 (0.971–1.340) per 1000 items per 1000 STAR-PU]. A higher percentage of the practice populations aged $\geq$ 65 years and with diabetes were both associated with increased total CDI incidence. Incidence also varied among the health boards and financial years (table 6.3). Increased prescribing of clindamycin was associated with higher total CDI incidence [RR (95% CI) = 1.079 (1.001–1.162) per log items per 1000 registered patients] in the fully adjusted model. The other high-risk antibiotics (co-amoxiclav, cephalosporins and quinolones) showed positive associations with total CDI incidence, however, results were weaker with wider 95% CIs [RR (95% CI) = 1.080 (0.999–1.168), 1.063 (0.986–1.149) and 1.013 (0.918–1.117), respectively] (table 6.4).

Interaction Model

An interaction model for CDI incidence showed the relationship with inpatient or non-inpatient cases to vary for deprivation and health board (tests for interaction P = 0.034 and P $<$ 0.001, respectively). Figure 6.9 presents the interaction between CDI case source and health boards for total CDI incidence, showing that inpatient and non-inpatient CDI incidence varied between health boards. In health boards 6 and 7 the predicted rates of CDI are similar for inpatient and non inpatient CDI whereas for the other 5 health boards the predicted rate was higher for inpatient CDI. Figure 6.10

shows the interaction between CDI case source and deprivation, showing inpatient CDI to have a positive linear relationship with increasing deprivation of the GP practice, whereas non-inpatient shows no strong directional association.

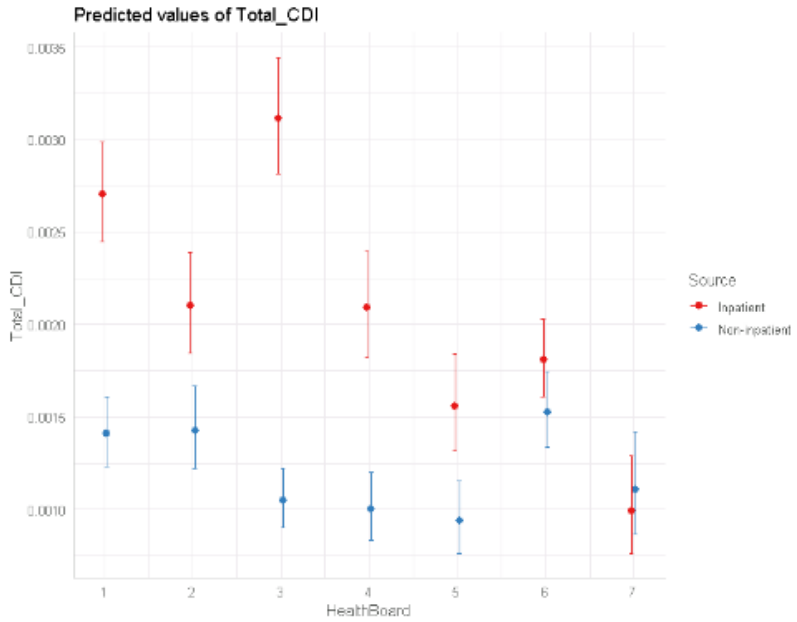


Figure 6.9: Interaction plot of total CDI incidence vs health board for inpatient and non-inpatient cases.

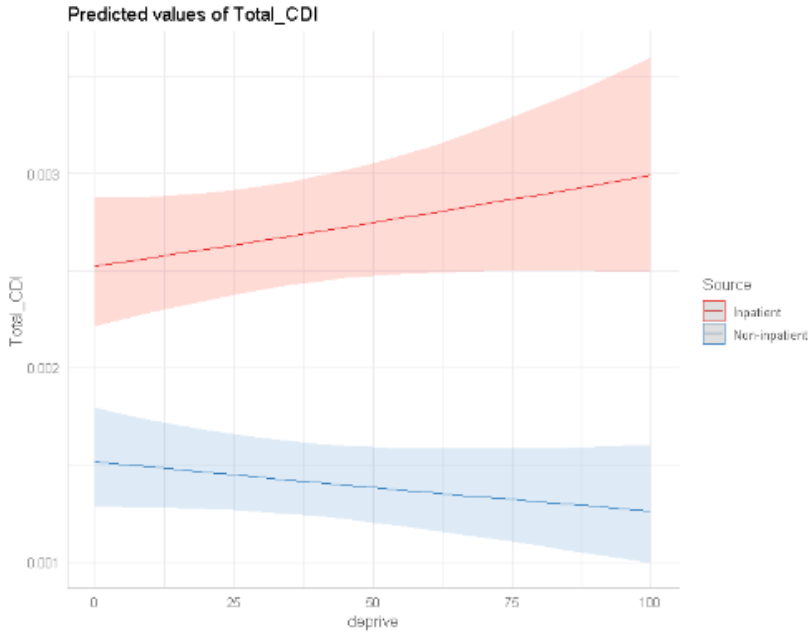


Figure 6.10: Interaction plot of total CDI incidence vs deprivation (%) by inpatient/non-inpatient.

6.4.3 Stratified CDI Incidence

The secondary analysis, where inpatients and non-inpatient CDI are modelled separately, showed inpatient CDI to be weakly associated with total antibiotic prescribing (per 1000 items per 1000 STAR-PU) [RR (95% CI) = 1.101 (0.907–1.335)], with non-inpatient CDI incidence showing similar, but slightly larger, results [RR (95% CI) = 1.213 (0.942–1.560)]. Inpatient CDI incidence showed an association with financial years. Each year was lower in comparison with 2014–15, however, this was not seen for non-inpatient cases. Inpatient CDI incidence was also seen to be associated with the percentage of practice population with diabetes and aged $\geq$ 65 years [RR (95% CI) = 1.079 (1.008–1.155) and 1.034 (1.018–1.050), respectively] (table 6.5).

Table 6.5: Adjusted RRs of Inpatient and Non-inpatients CDI incidence, with 95% CIs and P-values (from adjusted model).

	Inpatients		Non-Inpatients	
	RR (95% CI)	P	RR (95% CI)	P
Total Antibiotics (per 1000 items per 1000 STAR-PU)	1.1010 (0.9070–1.3353)	0.329	1.2128 (0.9418–1.5596)	0.13
Financial Year:				
2014-15	1	-	1	-
2015-16	0.8884 (0.7902 - 0.9987)	0.0479	1.0340 (0.8872 - 1.2051)	0.668
2016-17	0.7482 (0.6613 - 0.8462)	<0.001	0.8430 (0.7171 - 0.9906)	0.037
2017-18	0.7785 (0.6889 - 0.8795)	<0.001	0.9173 (0.7824 - 1.0752)	0.287
Health boards (1-7)	-	<0.001	-	0.012
Patients >65 (%)	1.0339 (1.0176 - 1.0504)	<0.001	1.0152 (0.9941 - 1.0367)	0.161
Social Deprivation Score: (%) most deprived	1.0014 (0.9987 - 1.0041)	0.297	0.9985 (0.9950 - 1.0020)	0.39
Proton Pump Inhibitor (per 1000 items per 1000 patients)	0.9981 (0.9871–1.0092)	0.734	0.9946 (0.9804 –1.0090)	0.453
COPD (%)	1.0532 (0.9712 - 1.1411)	0.21	1.0608 (0.9536 - 1.1782)	0.274
Diabetes (%)	1.0794 (1.0082 - 1.1552)	0.027	1.0317 (0.9437 - 1.1272)	0.488
Hypertension (%)	0.9949 (0.9700 - 1.0203)	0.329	1.0240 (0.9906 - 1.0582)	0.158

*Adjusted Models are adjusted for Financial Year, Health-board, (%) Patients $\geq$ 65, Social Deprivation Score - (%) most deprived, Proton Pump Inhibitor (per 1000 STAR-PU), (%) COPD, (%) Diabetes and (%) Hypertension. The global p-value for health boards is presented as the health bards are not identified.

The confidence intervals for the risk ratios for all the high risk antibiotics for both inpatient and non-inpatient CDI spanned 1. Clindamycin prescribing was the only high-risk antibiotic to show a positive estimate for both inpatient and non-inpatient CDI (table 6.5).

Table 6.6: Adjusted RR (95% CI) of Total CDI incidence associated with rates of high-risk antibiotic groups.

	Inpatient CDI incidence		Non-inpatient CDI incidence	
	RR (95% CI)	P	RR (95% CI)	P
Co-amoxiclav (log items per 1000)	1.0019 (0.9990- 1.0047)	0.193	0.9994 (0.9955- 1.0032)	0.131
Cephalosporins (log items per 1000)	0.9996 (0.9963 - 1.0028	0.798	1.0015 (0.9973 - 1.0056)	0.489
Clindamycin (log items per 1000)	1.0039 (0.9769 – 1.0114)	0.776	1.0124 (0.9769 - 1.0486)	0.493
Quinolones (log items per 1000)	1.0050 (0.9978 - 1.0122)	0.175	0.9935 (0.9839 - 1.0031)	0.188

Categorised Antibiotics

Inpatient and non-inpatient CDI incidences were compared to categorised antibiotic variables, initially visualised using boxplots. A gradual increase in median inpatient CDI incidence was observed in the boxplot in figure 6.11 across increasing quartiles of high-risk antibiotic (co-amoxiclav, cephalosporin, clindamycin and quinolone) prescribing (quartile 1–quartile 4), with quartile 4 showing the widest variability. For non-inpatient CDI there was an increase in median incidence from quartile 1 prescribing compared with all other prescribing categories (quartile 2, quartile 3 and quartile 4) for all high-risk antibiotic classes (figure 6.12).

Fully adjusted negative-binomial GLMs then assessed the categorised antibiotic variables for inpatient and non-inpatient CDI incidence, comparing this to crude CDI incidences by categories. Again, the confidence intervals for the risk ratios for all the high risk antibiotics for both inpatient and non-inpatient CDI spanned 1. Inpatient CDI incidence steadily increased from quartile 1 to quartile 3 total antibiotic GP prescribing, then decreased for quartile 4 GP prescribing. This was reflected in the RR in the adjusted model with quartile 2, quartile 3 and quartile 4 prescribers at increased risk of inpatient CDI compared to quartile 1 prescribers. Similar results were seen for non-inpatient CDI cases which showed increased risk estimates for quartile 2, quartile 3 and quartile 4 total antibiotics prescribers compared to quartile 1. All confidence intervals contained 1 with high p-values (table 6.7).

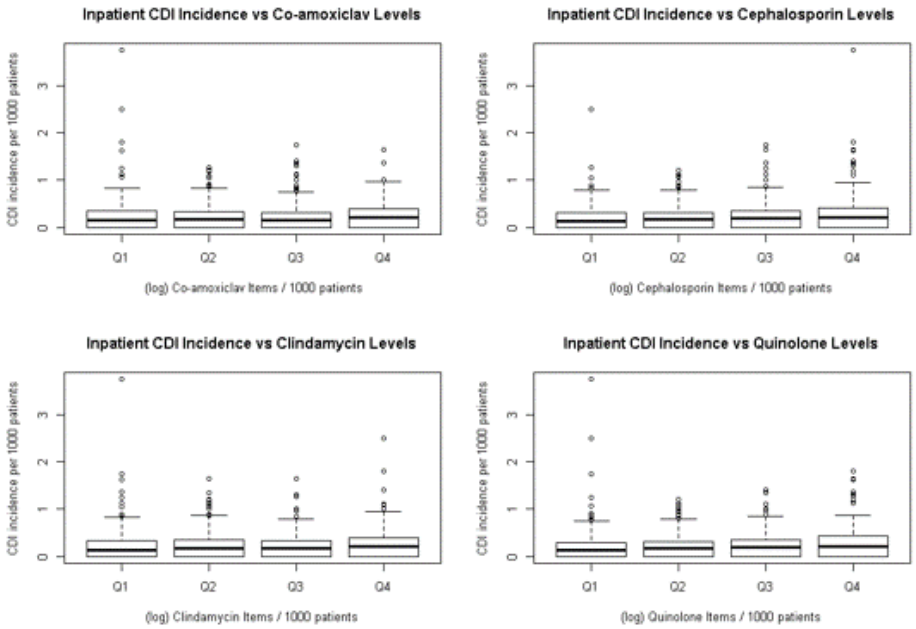


Figure 6.11: Inpatient CDI incidence rate (cases per 1000 registered patients) compared to categorised high-risk (co-amoxiclav (top-left), cephalosporins (top-right), clindamycin (bottom-left) and quinolones (bottom-right)) antibiotic prescribing rates (quartile 1 – quartile 4) by Welsh GP practices (log items per 1000 registered patients).

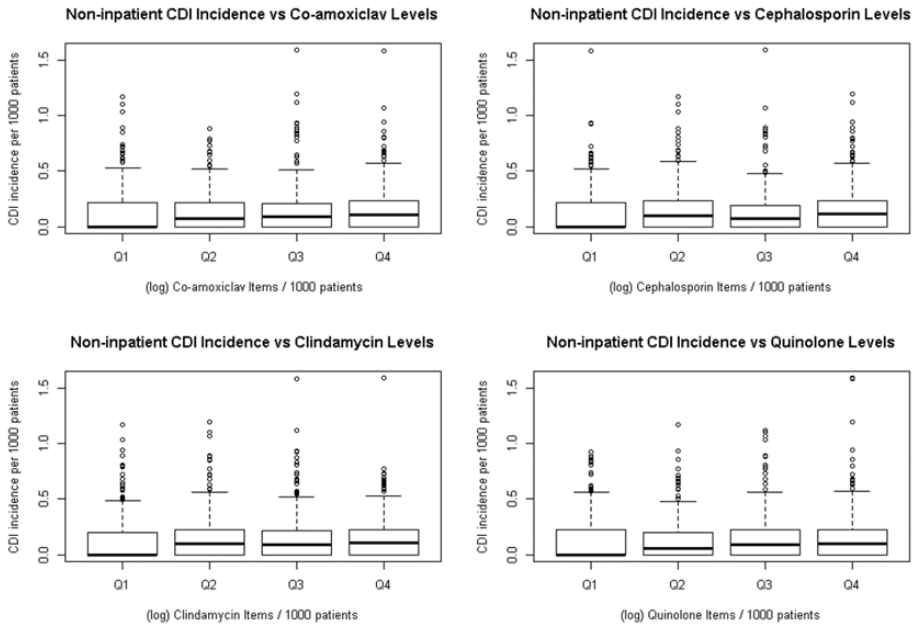


Figure 6.12: Non-inpatient CDI incidence rate (cases per 1000 registered patients) compared to categorised high-risk (co-amoxiclav (top-left), cephalosporins (top-right), clindamycin (bottom-left) and quinolones (bottom-right)) antibiotic prescribing rates (quartile 1 – quartile 4) by Welsh GP practices (log items per 1000 registered patients).

Table 6.7: Stratified CDI incidence associated with categorised practice GP prescribing rates of total antibiotics: adjusted RR, with 95% CI and p-values. Mean CDI incidence for each category of GP prescribing.

	CDI Incidence	Adjusted	
Total Antibiotics (items per 1000 STAR-PU)	Cases per 1000	RR (95% CI)	P
Inpatient			
quartile 1	0.183	1	-
quartile 2	0.220	1.023 (0.898 - 1.167)	0.730
quartile 3	0.246	1.080 (0.942 - 1.238)	0.269
quartile 4	0.240	1.044 (0.897 - 1.216)	0.578
Non-inpatient			
quartile 1	0.117	1	-
quartile 2	0.148	1.152 (0.977 – 1.359)	0.095
quartile 3	0.147	1.149 (0.964 - 1.369)	0.124
quartile 4	0.137	1.100 (0.903 - 1.339)	0.342

The risk of inpatient CDI increased for quartile 4 clindamycin GP prescribing compared to quartile 1 (RR= 1.1485, 95% CI 1.0008 - 1.3189), other quartiles all showed positive estimates but with 95% CI's spanning 1. Cephalosporin and quinolones both show increasing CDI incidence from quartile 1 prescribers to quartile 4 prescribers, however, estimates were low. RR estimates for all other high-risk antibiotics showed positive association with increased inpatient CDI for quartile 1 compared to quartile 4, but the 95% CIs spanned 1 (table 6.8).

Non-inpatient CDI was shown to have an increase in risk for quartile 1 GP prescribing of clindamycin compared with quartile 2, quartile 3 and quartile 4 GP prescribing (RR = 1.1362, 95% CI 0.9641 - 1.3403; RR = 1.1336, 95% CI 0.9574 - 1.3437 and RR = 1.1111, 95% CI 0.9263 - 1.3340), however, 95% CIs contained 1. Cephalosporins, coamoxiclav and quinolones showed varying RR associated non-inpatient CDI with positive and negative estimates between quartiles of high-risk antibiotic classes, however, all 95% contained 1 (table 6.9).

Table 6.8: Inpatient CDI incidence associated with categorical GP prescribing rates of high-risk antibiotics: adjusted RR, with 95% CI and p-values. Mean CDI incidence are shown for each category of GP prescriber.

	CDI Incidence	Adjusted	
Co-amoxiclav (log items per 1000)	Cases per 1000	RR (95% CI)	P
quartile 1	0.202	1	-
quartile 2	0.229	1.0788 (0.9487 - 1.2272)	0.249
quartile 3	0.207	1.0109 (0.8843 - 1.1559)	0.875
quartile 4	0.248	1.1527 (0.9973 - 1.3327)	0.055
Cephalosporins (log items per 1000)			
quartile 1	0.176	1	-
quartile 2	0.216	1.0433 (0.9461 - 1.1913)	0.531
quartile 3	0.234	1.0105 (0.8797 - 1.1616)	0.883
	0.257	1.0652 (0.9152 - 1.2422)	0.413
Clindamycin (log items per 1000)			
quartile 1	0.183	1	-
quartile 2	0.220	1.0296 (0.9055 - 1.1713)	0.656
quartile 3	0.246	1.0150 (0.8894 - 1.1591)	0.825
quartile 4	0.238	1.1485 (1.0008 - 1.3189)	0.049
Quinolones (log items per 1000)			
quartile 1	0.200	1	-
quartile 2	0.210	1.0514 (0.9240 - 1.1969)	0.448
quartile 3	0.214	1.0031 (0.8783 - 1.1461)	0.964
quartile 4	0.261	1.0278 (0.8939 - 1.1825)	0.700

Table 6.9: Non-inpatient CDI incidence associated with categorical GP prescribing rates of High-Risk Antibiotics: adjusted RR, with 95% CI and P-Values. Average CDI incidence are shown for each category of GP prescriber.

	CDI Incidence	Adjusted	
Co-amoxiclav (log items per 1000)	Cases per 1000	RR (95% CI)	P
quartile 1	0.132	1	-
quartile 2	0.129	0.9093 (0.7695 - 1.0747)	0.266
quartile 3	0.140	1.0520 (0.8893 - 1.2451)	0.556
quartile 4	0.148	1.0972 (0.9119 – 1.3207)	0.975
Cephalosporins (log items per 1000)			
quartile 1	0.117	1	-
quartile 2	0.143	1.0848 (0.9184 - 1.2826)	0.340
quartile 3	0.132	0.9466 (0.7915 - 1.1331)	0.551
quartile 4	0.156	1.1396 (0.9370 - 1.3874)	0.194
Clindamycin (log items per 1000)			
quartile 1	0.117	1	-
quartile 2	0.148	1.1362 (0.9641 - 1.3403)	0.128
quartile 3	0.147	1.1336 (0.9574 - 1.3437)	0.147
quartile 4	0.137	1.1111 (0.9263 - 1.3340)	0.257
Quinolones (log items per 1000)			
quartile 1	0.127	1	-
quartile 2	0.136	0.8856 (0.7535 - 1.0410)	0.143
quartile 3	0.139	1.0083 (0.8556 - 1.1886)	0.922
quartile 4	0.145	0.9145 (0.7661 - 1.0918)	0.324

6.5 Alternative Analysis

Health board Effect

Throughout this study, a strong health board effect has been seen to be associated with risk of CDI. All health boards and GP locations were anoynimised for this study, therefore it was not possible to explore these data spatially. However, the health board effect may be masking the effects of some other variable.

6.5.1 Methods

It was hypothesised that the health board differences might be described by hospital level prescribing in Wales. Data were obtained from the national point prevalence survey on hospital prescribing in Wales and were added to the data as an aggregated continuous covariate of hospital prescribing by health board for each financial year. Model comparisons were made between a **Null Model** and a **Health board Model** then the **Null Model** and a **Hospital Prescribing Model** (defined below in equations: 6.3, 6.4 and 6.5). Models were compared by using the deviance, AIC and likelihood ratio tests. Models assessed total CDI (y_{ijk}) with *offset* for GP practice population (N_{ijk}) for each GP practice i, within health boards j and year k: $k = 1 - N_{years}$, $j = 1 - N_{HB}$ and $i = 1 - N_{GP(jk)}$

Null Model

The null model did not include health boards or health board prescribing:

$$E(\log(y_{ijk})) = \log(N_{ijk}) + \alpha_k Yr + \beta_1 Over65_{ijk} + \beta_2 Deprivation_{ijk} + \beta_3 PPI_{ijk} + \beta_4 COPD_{ijk} + \beta_5 Diabetes_{ijk} + \beta_6 Hypertension_{ijk} + \beta_7 Antibiotic_{ijk} \quad (6.3)$$

Health board Model

The health boards j were then included back into the model as a fixed effect:

$$E(\log(y_{ijk})) = \log(N_{ijk}) + \alpha_k Yr + \gamma_j HB + \beta_1 Over65_{ijk} + \beta_2 Deprivation_{ijk} + \beta_3 PPI_{ijk} + \beta_4 COPD_{ijk} + \beta_5 Diabetes_{ijk} + \beta_6 Hypertension_{ijk} + \beta_7 Antibiotic_{ijk} \quad (6.4)$$

Hospital Prescribing Model

The hospital prescribing model included aggregated hospital prescribing for each health board j, therefore was the same for each GP practice i within health boards:

$$E(\log(y_{ijk}) = \log(N_{ijk}) + \alpha_k Yr + \beta_1 Over65_{ijk} + \beta_2 Deprivation_{ijk} + \beta_3 PPI_{ijk} + \beta_4 COPD_{ijk} + \beta_5 Diabetes_{ijk} + \beta_6 Hypertension_{ijk} + \beta_7 Antibiotic_{ijk} + \beta_8 HospitalPrescribing_{jk} \quad (6.5)$$

6.5.2 Results

There were no hospital prescribing data available for health board 7, therefore it was removed from these analyses. There was a large spread of hospital antibiotic prescribing rates for health boards 3, 5 and 6. Health board 1 shows the largest median prescribing rate (figure 6.13).

The health board model showed the lowest AIC and the highest percentage of deviance explained whereas there was little difference between the null model and the hospital prescribing model (table 6.10).

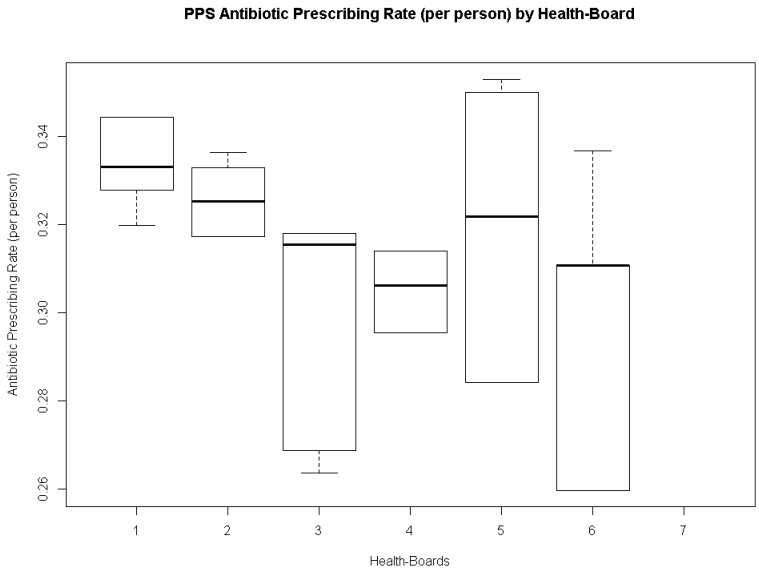


Figure 6.13: Boxplots of point prevalence survey hospital prescribing data (no data available for health-board 7).

Table 6.10: Model comparisons: AIC and deviance. Comparing the inclusion of point prevalence hospital prescribing and health boards

Model	AIC	d.o.f	Residual Deviance	Deviance Explained (%)
(1) Null Model	6481.5	1698	1943.611	8.86
(2) Health-Board Model	6428.8	1693	1937.555	11.88
(3) Hospital Prescribing Model	6482.7	1697	1943.899	8.92

Results from the likelihood ratio test confirmed that the addition of hospital prescribing did not result in an improved model fit. There was not enough evidence to suggest that the health board effect can be explained by hospital antibiotic prescribing and therefore health board remained in the model for the main analyses (table 6.11).

Table 6.11: Model comparison: likelihood ratio tests compared model the addition of health board and health board prescribing to the null model.

Model	theta	Resid. df	2 x log-lik.	Test	df	LR stat.	Pr(Chi)
(1) Null	6.221601	1698	-6457.548				
(2) Health-board	7.106374	1693	-6394.772	1 vs 2	5	62.77548	3.24E-12
(1) Null	6.221601	1698	-6457.548				
(3) Hospital Prescribing	6.236964	1697	-6456.737	1 vs 3	1	0.810269	0.368041

6.6 Discussion

Key Findings

This retrospective ecological study in Wales confirmed the hypothesis that overall GP surgery-level antibiotic prescribing rates were associated with an elevated risk of CDI. Unadjusted analysis showed a higher risk of total CDI incidence with total practice antibiotic prescribing [RR (95% CI) = 1.337 (1.170–1.529) per 1000 items per 1000 STAR-PU]. However, this effect was lower after adjusting for practice demographic covariates [RR (95% CI) = 1.141 (0.971–1.340)]. For the unadjusted estimate, this approximates to a 10% increase in risk of CDI between first- and third-quartile (1043.7 - 1363.0, items per 10000 registered patients) prescribers in 2017–18, which reduces to a 5% increase in risk after accounting for practice demographics. Higher total CDI incidence was associated with high percentages of practice population aged $\geq$ 65 years and with diabetes. Incidence decreased between 2014–15 and 2017–18 and also varied between health boards. However, the health board differences could not be explained by hospital prescribing at health board level. An increased risk of total CDI incidence was associated with antibiotic groups known to be high risk for CDI (co-amoxiclav, clindamycin, quinolones and cephalosporins) in the unadjusted analysis. An elevated risk of CDI was associated with clindamycin after adjusting for covariates [RR (95% CI) = 1.079 (1.001–1.162) per log items per 1000 registered patients], relating to a 4% increased risk between first- and third-quartile prescribers in 2017/18. Effects weakened for all other high-risk groups after adjusting for confounders. The secondary analyses indicated an increased risk of both inpatient and non-inpatient CDI associated with higher total antibiotic prescribing. High-risk antibiotic groups also suggested increased risk for both inpatient and non-inpatient CDI. Evidence was weaker in the secondary analyses as statistical power was lost by stratification.

Comparative Analysis

There was an overall incidence rate of 1.44 (95% CI 1.40–1.48) per 1000 patients for total CDI across four financial years. Yearly total CDI incidence decreased by 15.6% from 2014–15 to 2017–18 but increased between 2016–17 and 2017–18. Overall, antibiotic prescribing rates fell by 11.3% from 2014–15 to 2017–18, comparable with Public Health Wales reports, presumably in response to antibiotic stewardship efforts [134]. Prescribing rates of high-risk antibiotics also decreased during this time, by 29.3%, 32.0% and 14.5% for co-amoxiclav, cephalosporins and quinolones respectively. Clindamycin prescribing rates were seen to be low (≤ 1.04 item per 1000 registered patients) for each financial year, reflecting restricted indications in primary care guidelines; however, they increased by 38% from 2014–15 to 2017–18 (0.75 items per 1000 to 1.04 items per 1000). The reasons for this isolated increase are not clear. Common reasons for prescribing clindamycin include skin and soft tissue infection (including MRSA) and diabetic foot infection, particularly in the context of penicillin allergy, for which alternative appropriate antibiotics could be considered [63, 190]. Penicillin allergies may provide the reasoning behind high-risk antimicrobial prescribing such as clindamycin, however, inaccurate records of penicillin allergies can lead to unnecessary prescribing of such antibiotics [191]. Patients with a noted penicillin allergy are more likely to be prescribed clindamycin and experience worse health outcomes [192]. Penicillin allergies show an increased risk of MRSA and C. difficile [HR (95% CI) = 1.69 (1.51–1.90) and 1.26 (1.12–1.40, respectively] alongside increased use of macrolides, clindamycin and fluoroquinolones [193]. Improving the accuracy of recording penicillin allergy labels may be a good target for improving antibiotic stewardship and, in turn, affecting incidence of C. difficile.

The risks of CDI associated with the 4C antibiotic group are widely recognized and considered in antibiotic stewardship frameworks [194]. A meta-analysis of case-control studies investigating the association between CA-CDI and antibiotics identified clindamycin to have the strongest association with CA-CDI [OR (95% CI) = 20.43 (8.50–49.09), followed by fluoroquinolones and cephalosporins [5.65 (4.38–7.28) and 4.47 (1.60–12.50), respectively] [57]. The impact of the lasting effects of 4C prescribing can be seen in the risk of CA-CDI. A population-based case–control study on the cumulative and temporal effects of antimicrobial prescribing on CA-CDI showed that individuals exposed to $\geq$ 29 DDDs of any high-risk antimicrobial (cephalosporins, clindamycin, co-amoxiclav and quinolones) had an OR (95% CI) of 17.9 (7.6–42.2) [58]. Hence, these studies reiterate the importance of monitoring primary care antibiotic prescribing as small changes, such as a rise in clindamycin prescribing, could present serious problems.

This study reports 38% of all cases as non-inpatient and 62% as inpatient. Inpatient CDI incidence decreased over the study period, with slight increases in 2017–18, while non-inpatient CDI incidence fluctuated throughout this time. A study showing that a reduction of 10% in outpatient antibiotic prescribing could lead to a 17% (95% CI 6%–29.3%) decrease in CA-CDI highlighted a gap in the literature describing population-level impact of antibiotic use on CA-CDI [195]. Other work modelling inpatient and outpatient antibiotic stewardship interventions in a regional healthcare networks, suggested that a 30% reduction in inpatient and outpatient antibiotic prescribing could lead to a 17% decrease in healthcare-onset (HO) CDI and a 7% reduction in CA-CDI [36].

Limitations

A limitation of this study was that the inpatient/non-inpatient definition of CDI cases may not robustly measure the actual exposures to the healthcare system prior to disease presentation. For example, a patient who had recently been hospitalized, discharged and then presented at an emergency department would be classified as non-inpatient.

Strengths

Individual-level studies have shown the risks associated with antibiotics and 4C prescribing in the community and hospitals [58, 57, 196]. However, we believe this to be one of the few to report this association at a population-based ecological level. Although the associations shown at this level of analysis are less striking, the evidence of any relationship between primary care antibiotic prescribing and risk of CDI, particularly after accounting for differing patient demographics, is important.

Conclusions

In conclusion, this study shows that, even with high variability GP-level prescribing data, an increased risk of CDI can be seen to reflect antibiotic prescribing rates, particularly clindamycin, and demonstrates the continuing importance of antibiotic stewardship by prescribers.

Chapter 7

Scottish COVID-19 Testing Rates Compared with COVID-19 Symptom Reporting Platforms

7.1 Introduction

During the first wave of the pandemic, those who experienced COVID-19 symptoms were discouraged visiting general practitioner (GP) practices, pharmacies or hospitals. In Scotland, the recommended route to care was through NHS 24, as long as symptoms were not severe enough to call 999. NHS 24 is a telephone (111) and online service providing people with health information and advice 24 hours a day and is historically utilised as a route to care out of hours. Advice given during the first wave of the pandemic was that if a person was experiencing even mild COVID -19 symptoms, they should self-isolate for 14 days from the start of the symptoms and arrange a COVID-19 test. If symptoms worsened during this time, particularly if they had additional risk factors, developed breathlessness or symptoms lasted longer than 10 days then they were advised to call NHS 24 for support. Calls with symptoms related to COVID were flagged for surveillance purposes. NHS helplines such as NHS 24 are an established

resource for disease surveillance which can be used to provide early warning detection systems and impact policy decisions. For example, monitoring respiratory and gastrointestinal infections during a winter outbreak, health impacts of severe flooding in southern England and poor air quality episodes [197]. Data sources such as NHS 24 and NHS 111 are very important in disease surveillance and may provide vital information leading to the detection of events prior to the event itself occurring [198].

At the end of March 2020, the health science company ZOE launched The COVID Symptom Study (CSS) app in collaboration with King's College London. This app was a not-for-profit initiative to support COVID-19 research worldwide [199]. There are over 4 million contributors and it is, at the time of study, the world's largest ongoing study of COVID-19. The app promotes daily self-reporting of a person's health, regardless of whether they feel unwell or not, to understand COVID-19 spread. The app includes a user profile (age, sex, location, health conditions, medications, etc.) and then a separate platform for daily reporting on wellness (temperature, presence of cough, breathlessness, fatigue, etc.), however, features of the app grew with expanding knowledge of the disease. This app was not designed as a diagnostic tool, however, early results showed promising predictability capacities. Between March 28th to April 28th, the app recorded 2,450,569 UK and 168,293 US participants who self-reported symptoms. Modelling found age, sex, loss of taste or smell, persistent cough, serve fatigue and skipped meals to be the strongest predictors of a positive COVID-19 test, with a sensitivity of 0.65 (0.62 - 0.67) and specificity of 0.78 (0.74 - 0.80). Loss of taste and smell was seen to be the strongest predictor of COVID-19 [200].

The analysis in this chapter presents a spatial and spatio-temporal analysis of COVID-19 positive testing data at postcode districts (PCD) level in Scotland, aiming to assess the correlation of test positivity with COVID flagged NHS 24 Calls and predicted COVID cases from the CSS app users to determine the strength of these data as surveillance tools in the initial months of a pandemic.

7.2 Methods

7.2.1 Data sources and linkage

COVID-19 Data Sources

Three data sets were used for these analyses. COVID-19 testing positivity data and NHS 24 calls data flagged as COVID-related were both provided by Public Health Scotland (PHS). The CSS app users were obtained through the Secure Anonymised Information Linkage (SAIL) Databank as part of the BREATHE consortium [197]. The data sets contained non-identifiable aggregated weekly counts of activity by postcode districts (PCD) for NHS 24 and COVID-19 testing, however the CSS was converted from Lower Super Output Area (LSOA) to PCDs. These data sets cover the first wave of COVID-19 in Scotland - from March 2020 to June 2020. COVID-19 testing data and NHS 24 were both available for this entire time period, however, the CSS app data was only available from 30th March onwards. These three data sets were initially presented separately including all available weeks, however, modelling of these data was performed on a combined and subset version, covering the same 12-week period.

The COVID-19 testing data included weekly counts of the total number of COVID-19 tests carried out by NHS Scotland and the number of positive results in each PCD. There was a total of 230,759 tests carried out during this time period with 17,941 positive cases.

The NHS 24 data contained weekly total numbers of phone calls to NHS 24 and the number of these calls flagged as reporting COVID-19-related symptoms by PCD. A phone call was flagged as a 'COVID-19'-related call if any of the classic COVID-19 symptoms were mentioned. These symptoms include: cold and flu like symptoms, fever, continuous cough, difficulty breathing and, sickness and diarrhoea. During this time there were 393,233 calls made to NHS 24 and 118,993 of these were flagged with COVID-19-related symptoms. However, NHS 24 did not introduced a COVID-19 classification system until the 14th of April, therefore all classifications prior this time were back predicted (5-weeks) using a prediction model developed by Public Health Scotland for NHS 24 calls from mid April to the end of May relating to respiratory and gastrointestinal syndromes plus the patients age. These analyses had no access to the prediction model itself and only obtained model output data.

The third set of data was from the CSS app users which were obtained from the SAIL Databank. Throughout the pandemic SAIL have stored daily updates from the CSS app. These data were aggregated by week and Lower Super Output Area (LSOA) then converted to PCD within the secure remote desktop SAIL databank. These data were then in an unidentifiable form and were extracted for analysis. The data included the number of active app users per week per PCD and the number of users within each PCD *predicted* to have COVID-19 based on their symptom reporting, in the same week. The total number of users per PCD and the total number of *predicted* positive COVID-19 users from 30th March to 16th June were also obtained. There was a total of 209,975 users with 10,605 positive *predicted* cases in Scotland during this time.

Covariate Data

There were a number of potential covariates identified to help describe the spatial variability of COVID-19 risk across Scotland. Deprivation, gender, population density and age are some of the key risk factors shown to be associated with severe COVID-

19 illness [73, 201, 198] and, therefore, a measure of these covariates were obtained. This included measures of the percentage of population employment deprived, income deprived and living in overcrowded spaces, which were all sourced from Scottish Index of Multiple Deprivation (SIMD) [202]. Age population distribution for the young and the elderly, percentage of male population, urban/rural classification and population density were sourced from 2011 census data and information service division (ISD) [140, 93, 149]. No temporally varying covariates were collected over this short period of time (see table 7.1 for full descriptions).

Table 7.1: Covariate definitions for income deprived, employment deprived, overcrowded living, standardised mortality ratio, urban/rural, male population, population density, population ages under 5, 12, and 17 years old and population ages over 64, 74, and 84 years old.

Covariate	Definition
Income Deprived (% PCD population)	Income deprivation, as defined by the Scottish Index of Multiple Deprivation (SIMD), is a measure of the percentage of the population (adults and their dependents) in receipt of Income Support, Employment and Support Allowance, Job Seekers Allowance, Guaranteed Pension Credits, Child and Working Tax Credits, or Universal Credit (excluding those in the category 'working with no requirements'), or in Tax Credit families on low income.

Employment Deprived (% PCD population)	Employment deprivation, as defined by the Scottish Index of Multiple Deprivation (SIMD), is a measure of the percentage of the working-age population (men aged 16-64 and women aged 16-60) who is on the claimant count, those who receive Incapacity Benefit, Employment and Support Allowance or Severe Disablement Allowance, and Universal Credit claimants who are not in employment.
Overcrowded Households (% PCD population)	The proportion of household population that live in overcrowded housing based on the occupancy rating. This compares the actual number of rooms in the house to the number of rooms which are required by the household, based on the relationships between them and their ages. Overcrowding is defined to mean households with an occupancy rating of -1 or -2 i.e. that there is either 1 or 2 rooms too few in the household.

Weight average Standardised Mortality Ratio (SMR) per PCD	Standardized Mortality Ratio (SMR) is a ratio between the observed number of deaths in an study population and the number of deaths expected, based on the age- and sex-specific rates in a standard population and the size of the study population by the same age/sex groups.
Urban Rural Classification (2-fold)	Areas with a population of less than 3,000 people and more than a 30-minute drive time of a settlement of 10,000 or more as classified as rural; otherwise urban.
Male Population (% PCD population)	Percentage of PCD population recorded as Male.
Population Density per PCD	Population density is defined by the number of people per km-squared per PCD.
Aged under 5, under 12, under 17 (% PCD Population.)	Percentage of PCD population aged under 5, 12 and 17
Aged over 64, over 74, over 84 (% PCD Population.)	Percentage of PCD population aged over 64, 74 and 84.

Spatial Scale

The COVID-19 testing data and NHS 24 COVID-19 calls were available by PCD level ,however, all other data were converted to PCDs. The covariate data were only available by IZs and the CSS data were available by LSOA. There are 429 PCDs in Scotland, defined by the first four characters of the postcode, e.g. AB10 and there are 1279 IZs,

however, these do not match up completely. To convert IZ to PCD, the total number of PCD postcode and IZ unique pairings were calculated. The total number of postcodes within each IZ was also counted to then allow for a proportion of postcodes per IZ within each PCD to be calculated. This proportion could then be used to distribute IZ data such as population, assigning a proportion of the information to the correct PCD. The same technique was applied when converting the ZOE app data from LSOA to PCD. This could not be done by population as there were no population data available for PCDs within IZs (figure 7.1).

	PC1	IntZone	N	Total	prop
	<chr>	<chr>	<int>	<int>	<dbl>
1	AB13	S02001236	3	155	0.0194
2	AB14	S02001236	146	155	0.942
3	AB31	S02001236	5	155	0.0323
4	AB32	S02001236	1	155	0.00645
5	AB13	S02001237	75	125	0.6
6	AB14	S02001237	5	125	0.04
7	AB15	S02001237	44	125	0.352
8	AB21	S02001237	1	125	0.008
9	AB10	S02001238	1	201	0.00498
10	AB15	S02001238	200	201	0.995

Figure 7.1: Data structure screenshot of transforming IZ to PCD by proportion of postcode.

7.2.2 Statistical Methods

A primary analysis initially assessed spatial covariates compared to COVID-19 positive testing, conducting backward selection and testing residuals for spatial autocorrelation. These data were then modelled spatially, adjusting for spatial covariates and comparing the effects on the spatial model parameters.

Exploratory Spatial Analysis

Each data set was initially explored individually, assessing maps of each aggregated data set by PCD from March to June, including insets of Scotland's most densely populated health boards: NHS Greater Glasgow and Clyde and NHS Lothian. Moran's

I test for spatial association was applied to quantify the spatial correlation between PCDs. Summaries (median and IQR) of the raw counts and proportions per PCD were then produced. The distributions of covariates were assessed individually and log transformations were applied for positively skewed variables. The relationship between positive testing rates, COVID-19 NHS 24 calls and CSS app predicted cases were compared with each of the covariates and, scatter-plots with a fitted generalised additive model (GAMs) were produced to explore the suitability of a linear assumption.

Binomial Generalised Linear Models (GLMs)

The COVID-19 testing data assume a binomial distribution for the total number of COVID-19 tests N_i and the number of positive COVID-19 tests Y_i for each PCD i:

$$Y_i \sim Binomial(N_i, p_i)$$

therefore, the logit link function, $g(.)$, was applied for binomial GLMs (Chapter 2, section 2.15).

Univariable binomial GLMs were applied to assess the relationship between sociodemographic covariates and the proportion of COVID-19 positive tests per PCD. The correlation matrix of covariates assessed as a potential collinearity problem was expected due to multiple measures of deprivation. Multivariable GLMs were then assessed and backward selection was applied to finalise model covariates.

Three multivariable binomial GLMs were presented, each including different age covariate pairs (under 5 and over 84, under 12 and over 74, under 17 and over 64) then models were compared using AIC.

Covariate Models

$$E(logit(p_i)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{3i} + \beta_5 AGE_{1i} + \beta_6 AGE_{2i} \quad (7.1)$$

such that p_i represents the proportion of positive COVID-19 tests in PCD i. X_1 represents log population density; X_2 represents PCD population male (%) and X_3 represents Urban (=0) /Rural (=1) classification. Three binomial models (BM) were then compared for different age pairs (figure 7.2 for a diagram of the analyses plan):

- **BM1**: AGE_1 = % population under 5 and AGE_2 = % population over 84
- **BM2**: AGE_1 = % population under 12 and AGE_2 = % population over 74
- **BM3**: AGE_1 = % population under 17 and AGE_2 = % population over 64

The percentage of COVID-19 related NHS 24 calls and the percentage of CSS app users with *predicted* COVID-19 were then introduced to the multivariable binomial models as key predictors, adjusting for covariates. The residuals from each model were extracted and tested for residual spatial association using Moran's I test for spatial association (figure 7.2 for a diagram of the analyses plan).

Fully Adjusted Models

The variable COV_i represents either NHS 24 calls (%) or the CSS app (%) for PCD i.

$$E(logit(p_i)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 AGE_{1i} + \beta_6 AGE_{2i} + \beta_7 COV_i \quad (7.2)$$

- **BM4**: $COVID_i$ = NHS 24 COVID-19 calls (%)
- **BM5** $COVID_i$ = COVID positive CSS app users (%)
- **BM6** $COVID_{1i}$ = NHS 24 COVID-19 calls (%) and $COVID_{2i}$ = COVID-19 positive CSS app users (%)

Spatial CAR Leroux Models

These data were then modelled spatially with a CAR Leroux prior to account for spatial autocorrelation (Chapter 2, equation 2.21). Models were assessed with proportion of positive COVID-19 tests as a binomial response. Covariates were successively introduced into the models to allow for comparisons between the spatial variability (τ) and spatial dependence (ρ) parameters. This would indicate the impact of key predictors on COVID-19 variability. Initially, an intercept model was assessed and then univariable spatial models for NHS 24 COVID-19 calls (%) and COVID-19 positive CSS app users (%). A covariate's only spatial model was then assessed, and finally covariates were combined with COVID-19 key predictors following the same structure as equation 7.2:

Spatial Models

- **S1**: Positive Tests $\sim$ 1 (intercept model)
- **S2**: Positive Tests $\sim$ NHS 24 COVID-19 calls (%)
- **S3**: Positive Tests $\sim$ COVID CSS app users (%)
- **S4**: Positive Tests $\sim$ Covariates
- **S5**: Positive Tests $\sim$ NHS 24 COVID-19 calls (%) + Covariates.
- **S6**: Positive Tests $\sim$ COVID CSS app users (%) + Covariates.
- **S7**: Positive Tests $\sim$ NHS 24 COVID-19 calls (%) + COVID CSS app users (%) + Covariates.

Model estimates, with credible intervals (Cr.I's) were compared, and model fit assessed from the Deviance Information Criterion (DIC), the corresponding estimated effective number of parameters (p.d), and the Log Marginal Predictive Likelihood (LMPL). The best fitting model minimises the DIC while maximising the LMPL. Figure 7.2 presents a diagram to visually explain the GLM models and spatial analyses in this section.

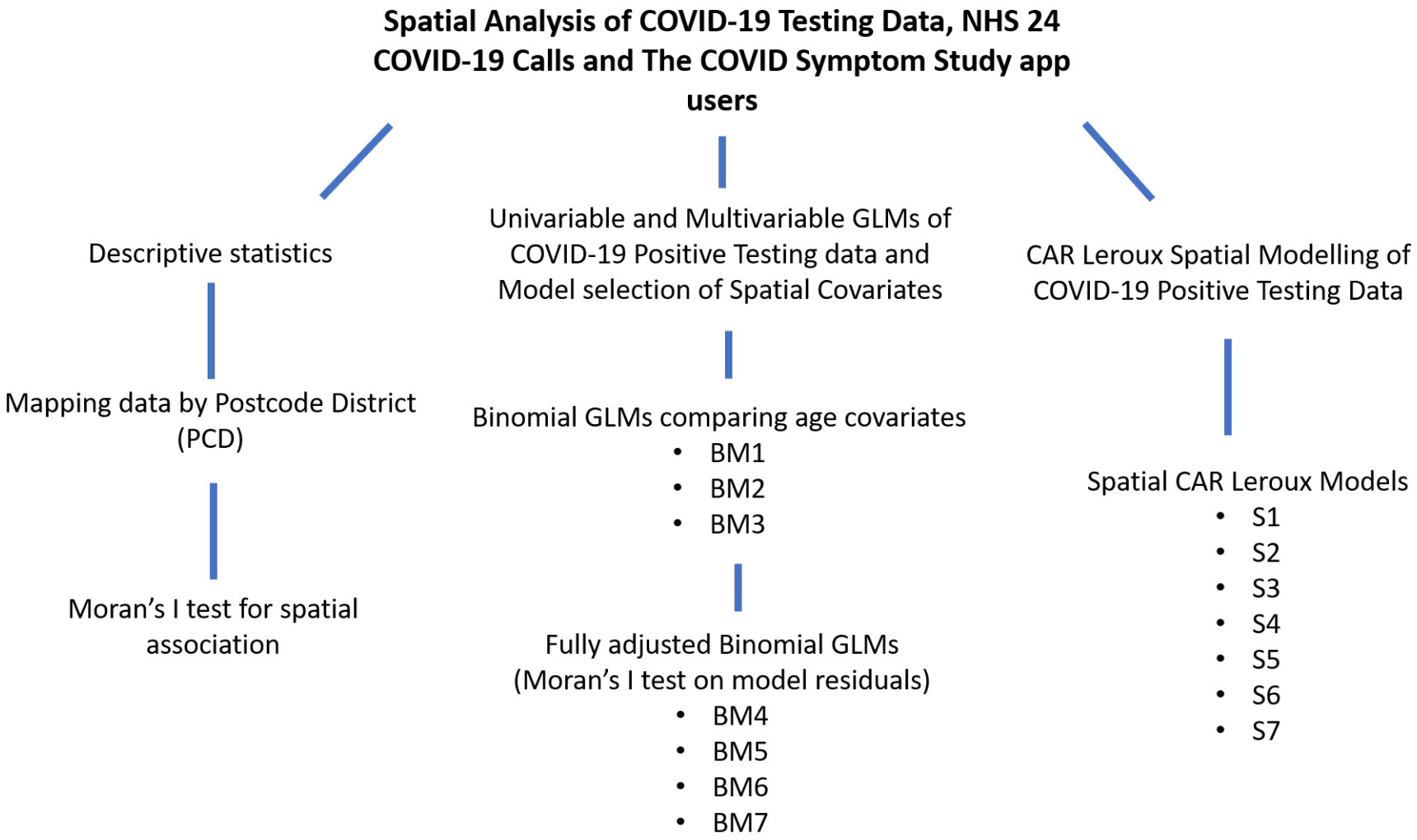


Figure 7.2: Diagram of Spatial Analysis plan for COVID-19 Positive Testing Data.

Spatio-Temporal Analysis

A secondary analysis introduced the temporal variation of the COVID-19 testing, NHS 24 calls and CSS app data, presenting weekly counts per PCD.

Temporal Exploratory Analysis

Each variable was visualised over time using boxplots of weekly proportions per PCD of COVID-19 positive testing, NHS 24 COVID calls and CSS app predicted COVID-19 per weekly active users. Summary tables were produced for raw counts and proportions.

The Auto-Correlation Function (ACF) was calculated for of COVID positive tests, NHS 24 COVID calls and CSS app predicted COVID cases. The spatio-temporal data were aggregated by week to produce a data set of total counts and proportions for each of the 12 weeks. The ACF was the calculated for lags of 1 and 2 weeks as there were only a small number of data points.

Spatio-temporal AR(1) Models

This analysis applied multiple spatio-temporal AR(1) model (Chapter 2, section 2.3.3). Multiple spatio-temporal models were assessed, successively introducing variables to compare by the amount of spatio-temporal variability (τ) accounted for in each model and how spatial (ρ_S) and temporal (ρ_T) dependence parameters changed. Model estimates with 95% CrI's, were assessed and model comparisons were made using DIC, p.d. and LMPL information. This analysis followed the same process as the spatial analyses previously described.

- **ST1**: Positive Tests $\sim$ 1 (intercept model)
- **ST2**: Positive Tests $\sim$ NHS 24 COVID-19 calls (%)
- **ST3**: Positive Tests $\sim$ COVID positive CSS app users (%)

- **ST4**: Positive Tests $\sim$ Covariates
- **ST5**: Positive Tests $\sim$ NHS 24 COVID-19 calls (%) + Covariates.
- **ST6**: Positive Tests $\sim$ COVID positive CSS app users (%) + Covariates.
- **ST7**: Positive Tests $\sim$ NHS 24 COVID-19 calls (%) + COVID positive CSS app users (%) + Covariates.

Spatio-temporal models (1-week lag)

It was of interest to explore lagged versions of the NHS 24 and CSS apps data. COVID-19 test results should be returned within 24 hours, however, can take up to 3 days [203]. This combined with the expected time taken between reporting symptoms and taking a test, it was hypothesised that at the time of study there may be delays between onset of COVID-19 symptoms and testing positive. The relationship between positive COVID testing data and 1-week lagged symptom variables were assessed using spatio-temporal AR(1) models. These models were computed on 11 weeks of data and, therefore, the DIC of these models cannot be used to compare to the previous spatio-temporal models which were modelled on 12-weeks of data.

- **ST5**: Positive Tests $\sim$ lagged NHS 24 COVID-19 calls (%) + Covariates.
- **ST6**: Positive Tests $\sim$ lagged COVID positive CSS app users (%) + Covariates.
- **ST7**: Positive Tests $\sim$ lagged NHS 24 COVID-19 calls (%) + lagged COVID positive CSS app users (%) + Covariates.

Sensitivity Analysis

Spatio-temporal ANOVA model

The spatio-temporal AR(1) model has the assumption that each time period of data has the same spatial structure as all other time periods. This assumption was tested by calculating Moran's I for each week of COVID-19 testing data.

A spatio-temporal ANOVA model was then assessed to determine whether a varying spatial structure over time, would affect model estimates. The spatio-temporal ANOVA model (section 2.24) splits the spatio-temporal variation into three components: the overall spatial effect common to all time periods; the overall temporal effect common to all spatial units; and a set of independent space-time interactions. This model was applied for COVID-19 testing data, with binomial response and fully adjusting for NHS 24 COVID calls (%); CSS app positive COVID users (%); age (%); deprivation (%) and population density. See figure 7.3 for a diagram of spatio-temporal analyses structure.

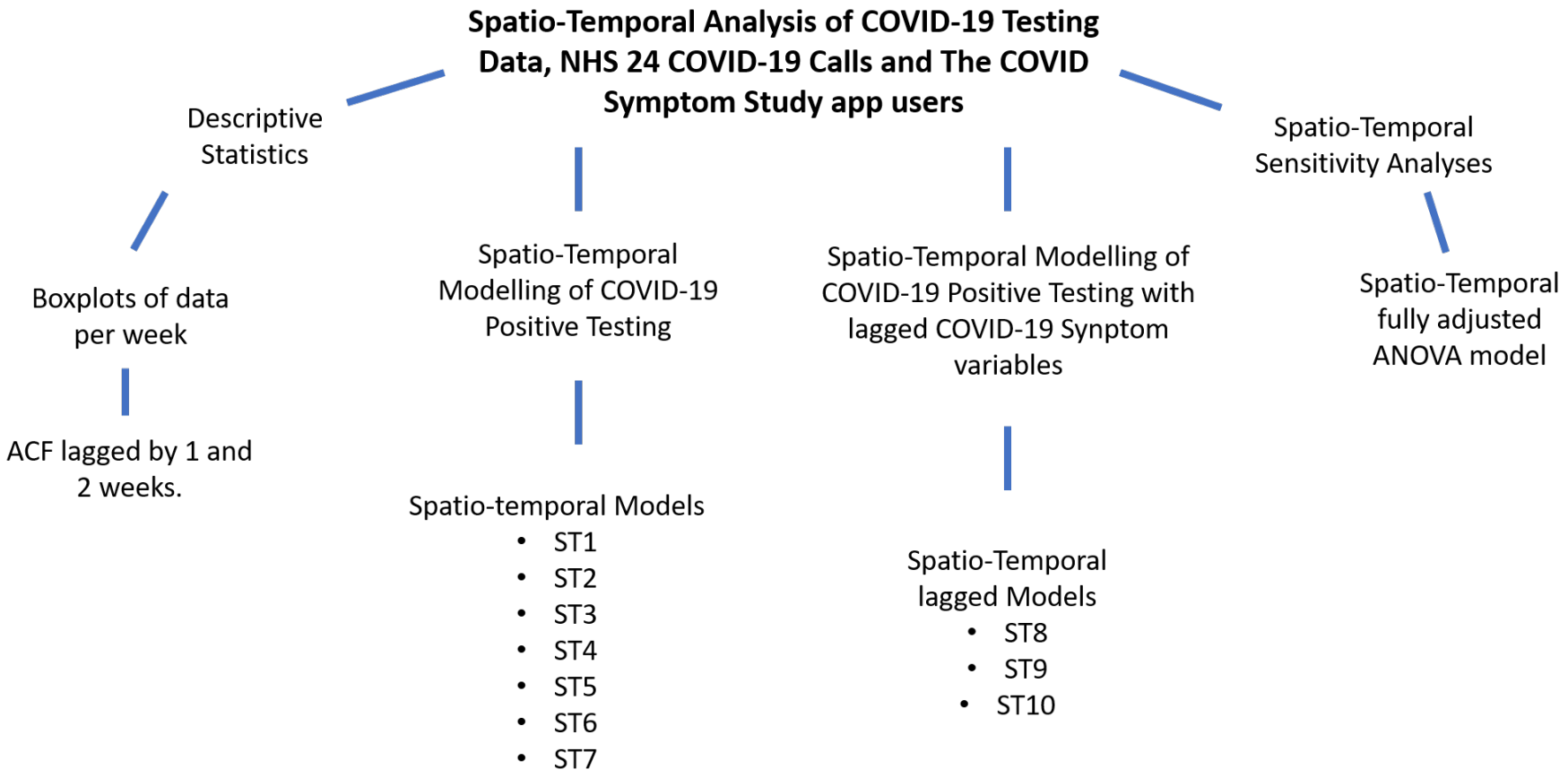


Figure 7.3: Diagram of Spatio-Temporal Analysis plan for COVID-19 Positive Testing Data.

7.3 Spatial Analyses Results

7.3.1 Exploratory Spatial Analysis

There was a total of 230,759 COVID-19 tests carried out in Scotland from the 2nd March 2020 to 15th June 2020, with 17,941 positive cases. Median (IQR) number of tests was 280 (53 - 868) per PCDs with median of 13 (1 - 63) positive tests per PCD. The median (IQR) percentage of positive tests by PCD was 5.3% (1.9% - 8%) with a maximum of 34% positive tests on outskirts of the City of Edinburgh (table 7.2)).

Table 7.2: Median (IQR) with maximum and minimum values for aggregated positive testing data, NHS 24 calls data and CSS app users from March to June, by PCD. Median, interquartile ranges, maximum and minimum values with proportions for each variable.

	Minimum	1st Q	Median	3rd Q.	Maximum
COVID-19 Testing					
Number of tests	16	53	280	868	2610
Number of positive COVDI-19 tests	0	1	13	66	306
% positive COVID-19 tests	0	1.9	5.3	8.1	3.4
NHS 24 COVID-19 related calls					
Number of NHS 24 calls	16	87	484	1514.5	5019
Number of COVID-19 related calls	0	23.5	141	451	1658
% NHS 24 COVID-19 calls	0	25.3	29.4	32.1	53.6
COVID Symptom Study app users					
Number of users	2	75	289	767	3719
Number of positive COVID-19 predictions	0	3	13	38	164
% CCS users positive COVID-19	0	3.1	4.5	5.8	11.9

Mapping the proportion of positive tests (figure 7.4) shows higher proportions in more densely populated areas of Scotland with the Central Belt of Scotland showing higher positive testing proportions in comparison to the North of Scotland. Proportion of positive tests appears low in the majority of the Scottish islands except for a notable outbreak in the Isle of Skye where in May 2020 approximately 60 residents and staff from a care home in Portree contracted COVID-19 which resulted in an outbreak across the island [204]. The Isle of Skye shows high positive testing rates, with more than 20% of all tests carried out returning positive results.

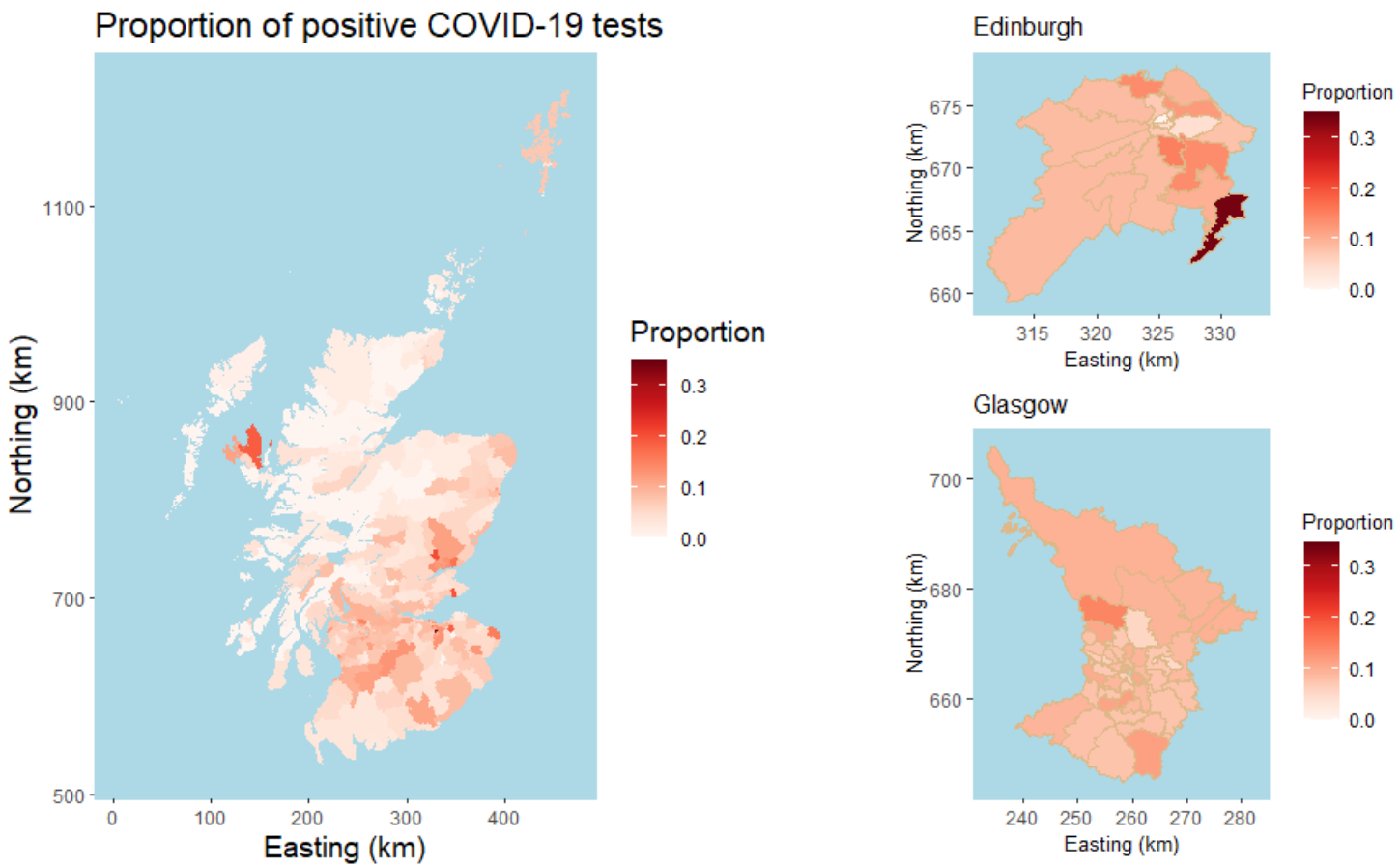


Figure 7.4: Proportion of positive COVID-19 tests per PCD in Scotland, aggregated from March to June with insets for Scotland's two most populated cities and surround areas: Edinburgh (postcode districts beginning EH - top-right) and Glasgow (post code districts beginning G - bottom-right).

Edinburgh (EH postcodes) and Glasgow (G postcodes) show similar patterns to each other with the majority of PCDs ranging between 10-20% positive testing percentage, however, one PCD in the Lothian area showed a high proportion of positive tests: $>$ 30% positive tests. Moran's I test for spatial association presented an I statistic of 0.537, with associated p-value < 0.001 indicating a strong positive spatial correlation in the COVID-19 positive testing proportions in Scotland (figure 7.4).

NHS 24 experienced a total of 393,233 calls from 2nd March 2020 to 15th June 2020 with 118,993 flagged as reporting COVID-19-related symptoms. The median (IQR) percentage of NHS 24 calls flagged as COVID-19-related by PCD was 29% (25% - 32%), with a maximum of 54% NHS 24 calls highlighting COVID-19 symptoms in the South of the Isle of Skye. The majority of PCDs have values between 25% and 32% for the percentages all NHS 24 calls which were COVID-19 related during this time period (table 7.2).

A map of the proportion of NHS 24 calls shows the majority of Scotland reporting $\geq$ 20% of NHS 24 calls with COVID-like symptoms, with many places reporting more than 40% of calls as COVID related. The spatial pattern appears slightly more sporadic compared to the testing data, however, areas such as Skye show high proportions of NHS 24 COVID-19 calls: this is comparable with the testing data. Conversely, many of the other Scottish islands show high proportions of COVID related NHS 24 calls which is not reflected in the testing data. Edinburgh and Glasgow show similar rates of COVID-related calls with the majority of PCDs reporting between 20% and 35% COVID flagged calls (figure 7.5).

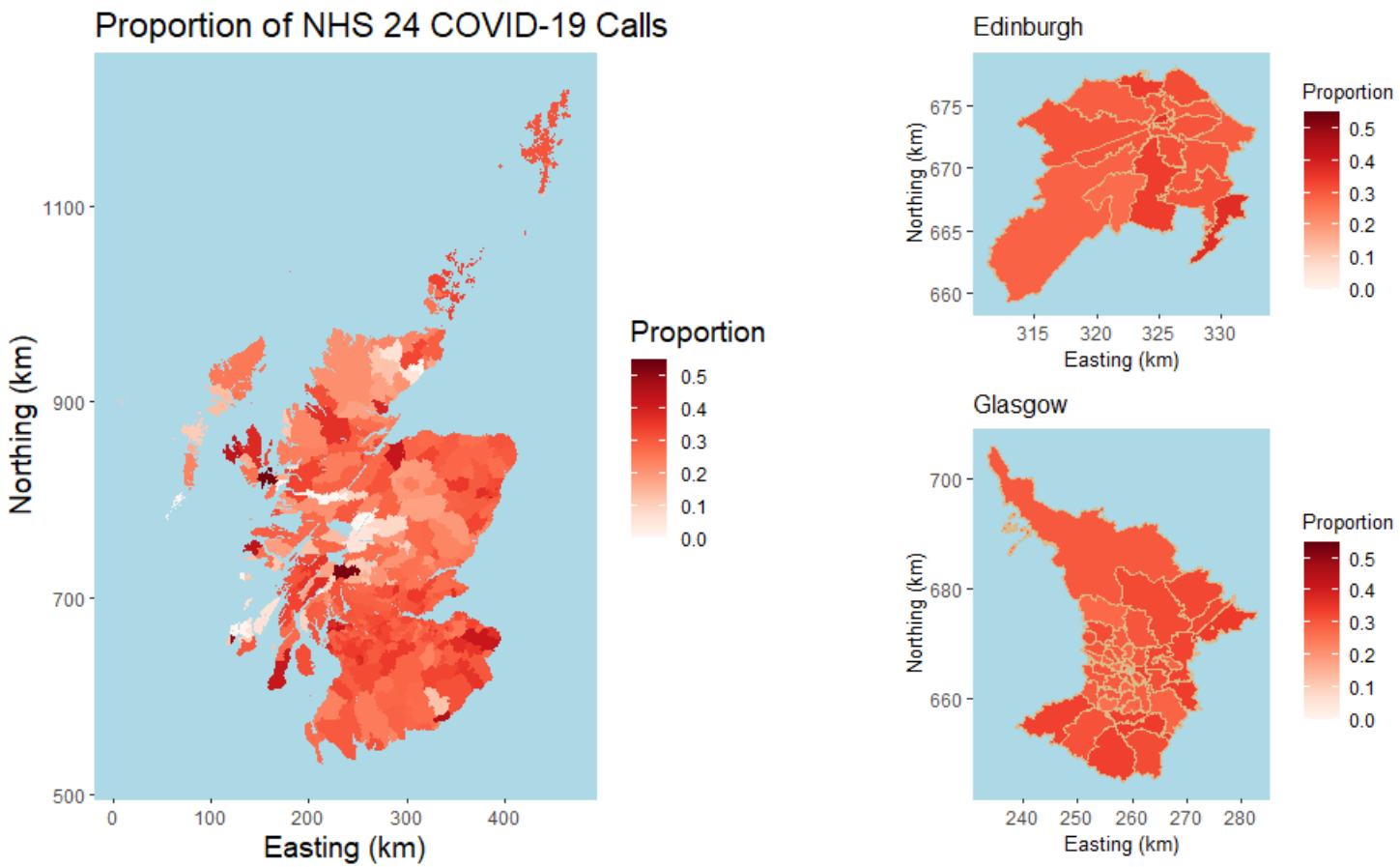


Figure 7.5: Proportion of COVID-19 related NHS 24 calls per PCD in Scotland, aggregated from March to June with insets for Scotland's two most populated cities and surrounding areas: Edinburgh (postcode districts beginning EH - top-right) and Glasgow (post code districts beginning G - bottom-right)

Moran's I test for spatial association returned I = 0.300 with an associated p-value < 0.001. This indicates a positive spatial association among PCDs in the proportion of COVID related NHS 24 calls. Spatial correlation was not as strong as the positive testing data.

In Scotland, there were 209,975 participants of the CSS (app users) from 30th March to the 15th June. The total number of *predicted* positive COVID-19 cases was 10,605. The median (IQR) percentage of users with *predicted* COVID-19 per PCD was 5% (3% - 6%) and a maximum of 12% of total users *predicted* to have COVID-19, by PCD. *Predicted* COVID-19 cases and positive tests, show comparable figures. However, the testing data are more widely spread (table 7.2).

The Central Belt of Scotland appeared to be consistently higher for the proportion of CSS app users with *predicted* COVID-19 compared to the north of Scotland where a few PCDs show 0 proportion of *predicted* cases. Edinburgh showed consistent proportions across PCD, however, Glasgow showed a few PCDs with a high *predicted* proportion, with some PCDs showing between 9%-12% of all user *predicted* to have had COVID-19 (figure 7.6). Note that these data contain one month's less data than for testing and NHS 24.

Moran's I test gave an I statistic = 0.314 and associated p-value < 0.001, indicating positive spatial correlation in CSS app proportion of *predicted* COVID-19 cases of the same order of magnitude as for NHS24.

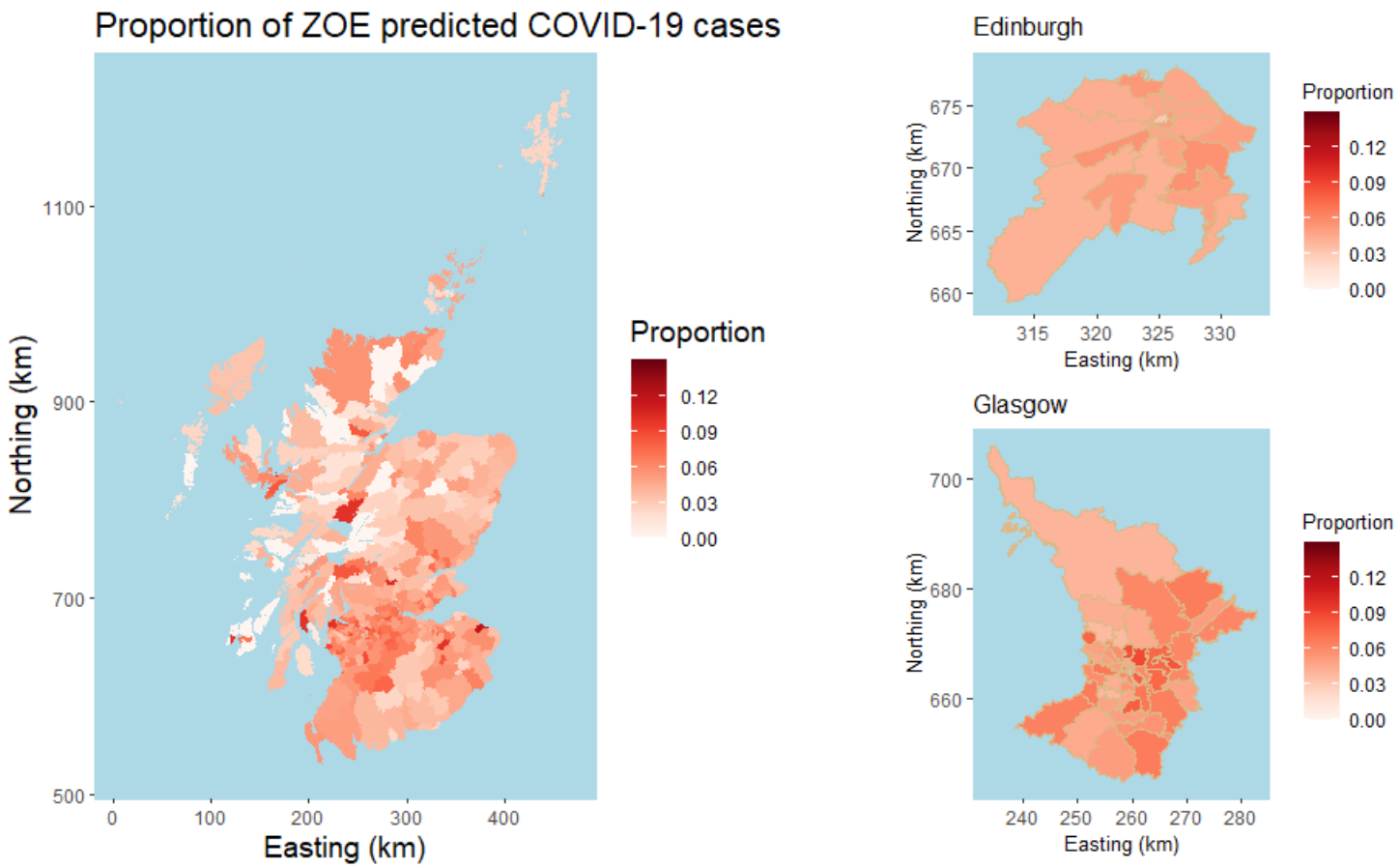


Figure 7.6: Proportion of COVID-19 positive CSS app users per PCD in Scotland, aggregated from March to June with insets for Scotland's two most populated cities and surrounding areas: Edinburgh (postcode districts beginning EH - top-right) and Glasgow (post code districts beginning G - bottom-right).

Spatial Covariates

The distributions of PCD population density and, percentages of PCD population income deprived, employment deprived and overcrowded living were all right skewed. Natural log transformations were taken to reduce spread of data (figure 7.7).

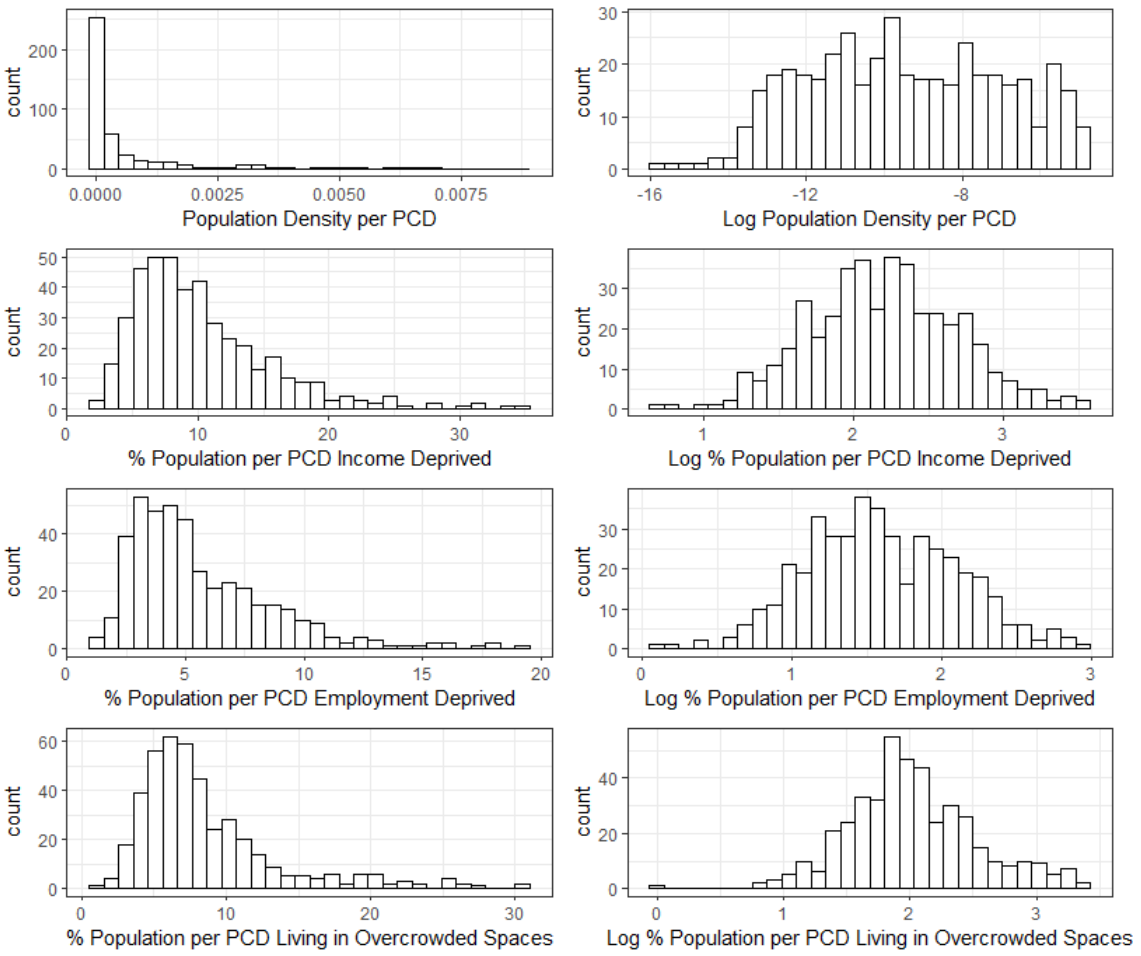


Figure 7.7: Log transformations of spatial covariate. Left hand side plots show population density (top), % population income deprived (second row), % population employment deprived (third row), % population overcrowded living (bottom). Right hand side plots show log transformed spatial covariates.

Multiple scatter plots were created to compare the proportion of positive COVID-19 tests against spatial covariates with fitted GAMs. The percentage of population aged under 5 and over 84 were plotted to represent PCD age distributions.

The percentage of PCD population aged under 5 showed a positive association with the proportion of positive COVID-19 tests, whereas the percentage of population aged over 84 was less clear and appeared relatively flat. The proportion of positive tests showed a positive association with log population density. The percentage of PCD population who are male showed an increasing trend at the beginning where there are few PCDs but appears to show a negative trend over the majority of data points. Deprivation and health factors (weighted average SMR, percentage of population in overcrowed living, income and employment deprived) all showed similar shallow increasing trends with the proportion of positive tests. The median proportion of positive tests for urban PCDs was much higher than rural PCD's (figure 7.7).

These plots gave reasonable evidence to assume a linear assumption when modelling these data. There are no strong non-linear trends seen from the fitted GAMs, with most curvature seen at the extremes of the variable ranges.

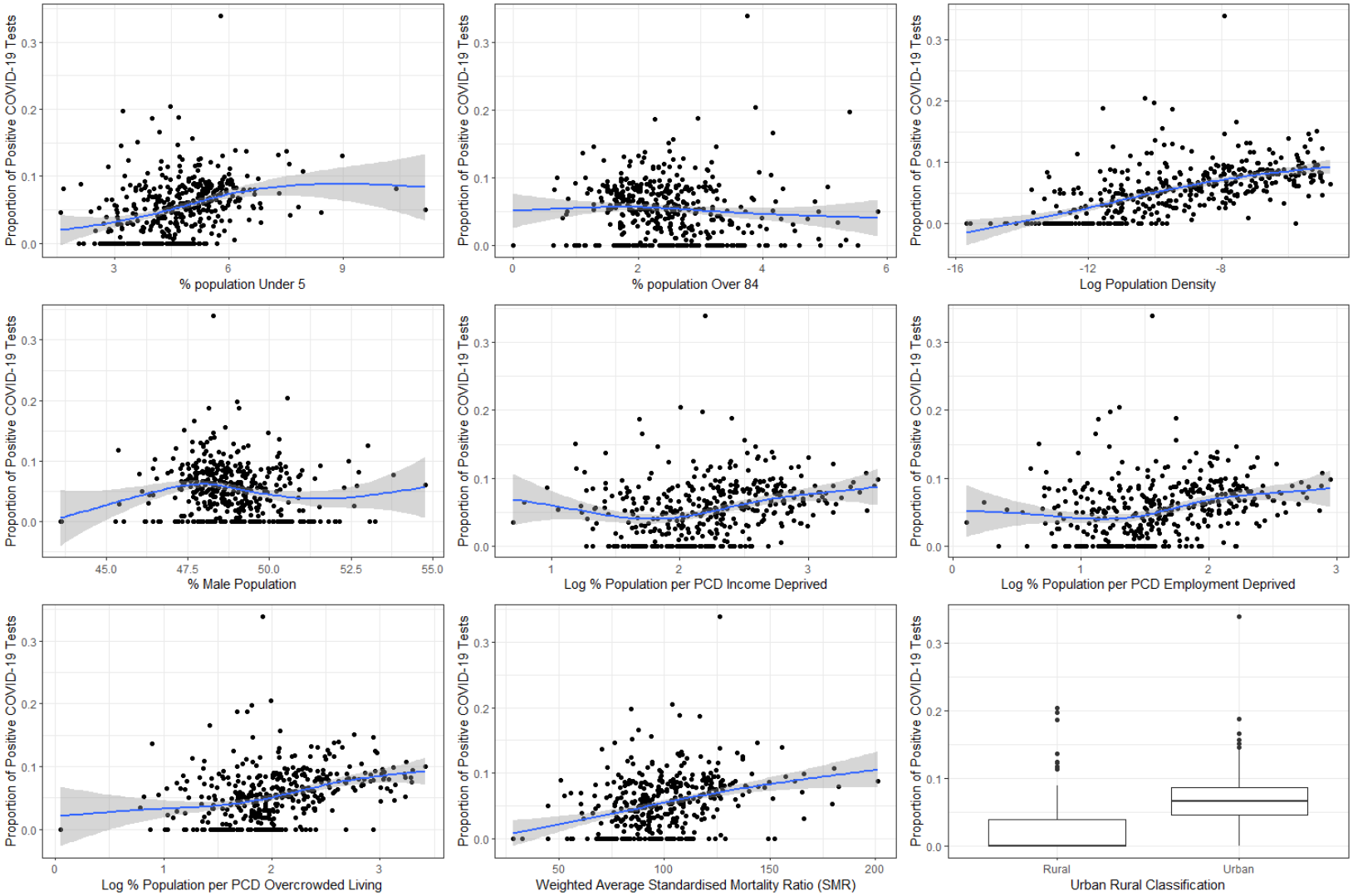


Figure 7.8: Proportion of positive COVID-19 tests compared to spatial covariates with fitted GAM's: percentage of PCD population aged under 5 (top-left), aged over 84 year (top-middle), log population density (top-right), percentage of PCD population male (left second row), income deprived (log) (middle second row), employment deprived (log) (right second row), overcrowded living (log) (bottom-left), weight average standardised mortality ratio (SMR) (bottom-middle) and urban rural (bottom-right).

7.3.2 Binomial GLM

This regression estimates in this section refer to the odds of testing positive for COVID-19 per 1 unit increase in model covariates: age covariate estimates relate to a 1% increase in percentage of PCD population, similarly for percentage of male population; deprivation covariates (income, employment and overcrowded living) are associated with a 1 unit increase in log percentage of population; population density refers to a 1 unit increase in the log number of people per PCD area (m^2); estimates of weighted average SMR (observed number of deaths divided by the number of expected deaths) also refers to a 1 unit increase per PCD and estimates for Urban/ Rural compare the odds of testing positive for Urban PCDs compared to Rural PCDs.

Univariate analyses of the spatial covariates showed a positive association with positive COVID-19 tests and the percentage of population under 5, 12 and 17 (table 7.3). The decreasing trend for percentage of population aged over 64 and 74 was seen (OR = 0.977, 95% CI 0.973 - 0.981 and OR = 0.964, 95% CI 0.956 - 0.971), whereas aged over 84 showed a 95% CI to contain 1 but also suggested a negative trend (OR = 0.978, 95% CI 0.956 - 1.001). COVID-19 positive tests showed an increasing association with log population density and weighted average standard mortality ratio (SMR) (OR = 1.112, 95% CI 1.102 - 1.123) and OR = 1.004, 95% CI 1.003 - 1.005). The percentage of the PCD population who are male showed a negative association (OR = 0.944, 95% CI 0.931 - 0.958), whereas deprivation factors (percentage of population overcrowded living, income and employment deprived) all presented an increasing trend with proportion of positive tests (OR = 1.265, 95% CI 1.224 - 1.308; OR = 1.134, 95% CI 1.100 - 1.169 and OR = 1.160, 95% CI 1.125 -1.197). Urban PCDs had twice the odds of increased COVID-19 positive testing in comparison to rural PCD's (OR = 2.171, 95% CI 1.901 - 2.494) (table 7.3).

Table 7.3: Univariable Binomial GLMs compared to spatial covariates with unadjusted OR with 95% CIs

	Unadjusted OR (95% CI)
% Population aged under 5	1.093 (1.074, 1.112)
% Population aged under 12	1.026 (1.018, 1.035)
% Population aged under 17	1.010 (1.004, 1.016)
% Population aged over 64	0.977 (0.973, 0.981)
% Population aged over 74	0.964 (0.956, 0.971)
% Population aged over 84	0.978 (0.956, 1.001)
log population density	1.112 (1.102, 1.123)
Male population (%)	0.944 (0.931, 0.958)
Log (%) population income deprived	1.134 (1.100, 1.169)
Log (%) population employment deprived	1.160 (1.125, 1.197)
Log (%) population overcrowded deprived	1.265 (1.224, 1.308)
Urban: Urban vs Rural	2.171 (1.901, 2.494)
Weight average SMR	1.004 (1.003, 1.005)

A number of the spatial covariates showed strong positive correlations with one another which presented a multicollinearity problem when constructing the multivariable models (figure 7.9). Unstable estimates were seen in the GLM models with the inclusion of employment deprivation, overcrowded living, income deprivation and average SMR mortality ratio while performing model selection, therefore, these were not included in the final model.

Model selection was originally performed including all deprivation factors. Initially, all covariates showed to be significantly associated with positive COVID testing except overcrowded (%) (p=0.25). Therefore, this was removed from the model. All covariates in the model then presented as statistically significant covariates at a 5% significance level. However, income deprivation (p=0.003) had a negative association with COVID positive testing and employment deprivation showed a positive association (p=0.019) which was queried given the similar directional association seen from the univariable analyses (table 7.3).

One deprivation was removed to see the effect on other model estimates: removing income deprivation caused the direction of association for employment deprivation to reverse and reduced the p-value (p=0.068). The direction of association did not change for SMR, however, the p-value of weight average SMR also reduced from (p = 0.0039 to p = 0.011). Removing employment deprivation then reduced the p-value of weighted average SMR further (p = 0.070) and was therefore removed. Other model estimates and p-values remained fairly constant during this process, therefore, it was decided to remove the deprivation covariates completely, other than population density, which was correlated with deprivation. The AIC including all covariates was $\text{AIC}_{FULL} = 1432.0$ and increased to $\text{AIC}_{REDUCED} = 4141.6$ with the removal of deprivation factors. A likelihood ratio test showed no significant difference between models.

Each of the deprivation covariates were strongly correlated with one another ($\rho > 0.6$), however, each of these covariates also had a relatively strong positive correlation with population density which remained in the final model.

Three multivariable binomial models compared the inclusion of the different age brackets. Effect sizes and direction of association did not change dramatically between models, therefore, the chosen model included the percentage of over 84 and percentage of under 5. This decision was informed by the lowest AIC, indicating a better fitting model. The demographic variables describing a PCD included in the final model were aged under 5 (%); aged over 84 (%); log population density; male population (%) and urban/rural classification (table 7.4).

Moran's I test for spatial association was carried out on the residuals from the final covariate multivariable model (**BM1** in table 7.4), and returned an I statistic = 0.299, with $p < 0.001$, indicating residual spatial correlation.

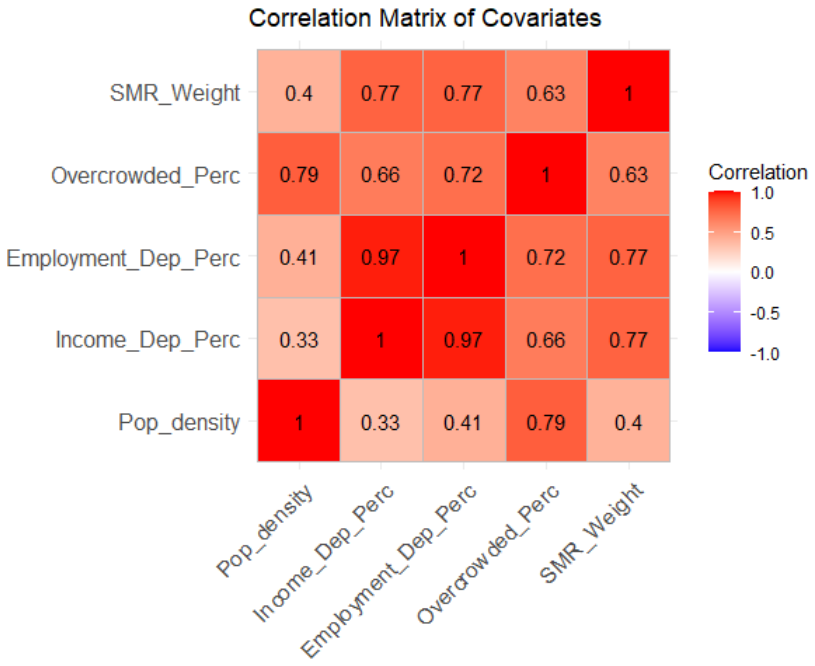


Figure 7.9: Pearson correlation matrix comparing spatial covariates.

The COVID-19 positive testing rate was more strongly correlated with the proportion NHS 24 COVID-19-related calls ($\rho = 0.439$) compared to the proportion of CSS app users with predicted COVID ($\rho = 0.377$), however, both showed fairly strong positive correlation. NHS 24 COVID calls also showed positive correlation with CSS app predictions ($\rho = 0.419$) (table 7.5).

The estimates for models including NHS 24 and the CSS refer to an increase in the odds of testing positive for COVID-19 per PCD for a 1% increase in percentage of COVID-19 related NHS 24 calls per PCD and a 1% increase in the percentage of The CSS app users predicted to have COVID-19 per PCD.

Table 7.4: Multivariable binomial GLM models with spatial covariates with odd ratios and 95% confidence intervals

	Adjusted OR (95% CI)		
	BM1	BM2	BM3
% Population aged under 5	1.06 (1.04,1.08)	-	-
% Population aged over 84	1.11 (1.08,1.15)	-	-
% Population aged under 12	-	1.02 (1.01,1.03)	-
% Population aged over 74	-	1.02 (1.01,1.03)	-
% Population aged under 17	-	-	1.0094 (1.0031,1.0157)
% Population aged over 64	-	-	1.01 (1,1.02)
log Population Density	1.11 (1.1,1.12)	1.12 (1.1,1.13)	1.12 (1.11,1.14)
Male population (%)	0.98 (0.97,0.99)	0.98 (0.96,0.99)	0.98 (0.96,0.99)
Urban: Urban vs Rural	1.48 (1.29,1.71)	1.49 (1.29,1.71)	1.5 (1.31,1.73)
AIC	4141.6191	4186.4355	4190.3959
Log-Likelihood	-2064.8096	-2087.2178	-2089.1979

Table 7.5: Correlation matrix for positive testing, NHS 24 calls and CSS predicted cases aggregated over time.

	Positive Testing	NHS 24 Flagged Calls	CSS Predicted COVID Cases
Positive Testing	1.000	0.439	0.377
NHS 24 Flagged Calls	0.439	1.000	0.419
CSS Predicted Cases	0.377	0.419	1.000

The percentage of NHS 24 calls flagged COVID (**BM4**) and CSS app predicted COVID (**BM5**) were both associated with an increase in the odds of testing positive for COVID-19 (OR=1.013, 95% CI 1.008 - 1.018 and OR = 1.041, 95% CI 1.029 - 1.054, respectively). Combining both key predictors into the same model showed both NHS 24 calls and CSS app to have a positive association with COVID positive testing (**BM6**) with similar estimates to the reduced models (**BM4** and **BM5**) (table 7.6). Covariate effects remain fairly consistent across all three models and are comparable with effects seen in the covariate only model (table 7.5).

Moran's I test on residuals from each model suggested evidence of positive spatial association after adjusted for spatial covariates and key predictor variables. The Moran's I statistics was comparable between all three models (I = 0.296, I = 0.279 and I =

Table 7.6: Multivariable binomial GLMs with OR (95% CI) including NHS 24 COVID flagged calls and CSS app user predicted COVID with Moran's I test for spatial association on model residuals.

	Adjusted OR (95% CI)		
	BM4	BM5	BM6
NHS 24 COVID flagged calls (%)	1.013 (1.008 - 1.018)	-	1.011 (1.006 - 1.016)
CSS app predicted COVID user (%)	-	1.041 (1.029 - 1.054)	1.037 (1.025 - 1.050)
% Population aged under 5	1.060 (1.039 - 1.081)	1.053 (1.033 - 1.074)	1.055 (1.035 - 1.077)
% Population aged over 84	1.12- (1.089 - 1.152)	1.144 (1.111 - 1.178)	1.146 (1.113 - 1.180)
log population density	1.109 (1.097 - 1.120)	1.102 (1.090 - 1.114)	1.100 (1.088 - 1.112)
Male population (%)	0.982 (0.967 - 0997)	0.995 (0.979 - 1.010)	0.995 (0.980 - 1.011)
Urban: Urban vs Rural	1.416 (1.233 - 1.635)	1.439 (1.253 - 1.661)	1.405 (1.223 - 1.622)
Log-likelihood	-2045.0776	-2037.1649	-2027.4613
AIC	4104.1552	4088.3298	4070.9225
Moran's I	I =0.296 (p <0.001)	I =0.279 (p <0.001)	I =0.284 (p <0.001)

0.284 for models **BM4**, **BM5**, and **BM6** respectively). This implied that the spatial correlation could not be explained by these spatial covariates, therefore invalidating the independence assumptions and supporting the modelling of these data spatially.

7.3.3 Spatial CAR Leroux Model

The following results present separate spatial models of COVID-19 positive testing data and then compare the effect of successively introducing spatial covariates on spatial model parameters. Comparisons were made between spatial variance (τ) and dependence (ρ) parameters. Models were adjusted for spatial covariates and presented in table 7.8. The percentage of male population (%) and urban/rural classification were not associated with the COVID-19 positive testing proportion when included into the spatial models and therefore were removed from this analysis.

Firstly, the impact of introducing spatial covariates was compared between the spatial model parameters in table 7.7: **S1** showed the intercept only model spatial variability estimated as $\tau = 0.812$ for the COVID-19 testing data. Introducing NHS 24 COVID-19 calls reduced the spatial variability (**S2**, $\tau = 0.718$), however, this was relatively unchanged with the inclusion of CSS app COVID users (**S3**, $\tau = 0.803$). Assessing an adjusting model including spatial covariates reduced τ to a similar level as seen

for the NHS 24 COVID calls model (**S4**, $\tau = 0.716$) compared to the intercept model (**S1**). Combining NHS 24 and spatial covariates into one model reduced τ more (**S5**, $\tau = 0.637$) whereas **S6** with CSS app and spatial covariates, showed no difference in τ from the covariates only model (**BM4**). These results suggested that the CSS app had little effect on spatial variability. **S7** combined all covariates into one model and τ reduced the most, compared to the intercept estimate. However, this was comparable to **S5** which did not include CSS app as a predictor. Hence, **S5** and **S7** were seen to account for the most spatial variability for COVID-19 positive testing. It was noted that all model τ estimates remained within all model 95% credible intervals (table 7.7).

The inclusion of spatial covariates showed a reduction in the spatial dependence ρ compared to the intercept and univariable models, however, all estimates again remained within the 95% credible intervals (table 7.7).

Table 7.7: Comparison of spatial models spatial variability and spatial dependence estimates for data by PCD with 95% credible intervals.

CAR Leroux model	τ	ρ (95% Credible Interval)
S1: Intercept Model	0.812 (0.654 - 1.013)	0.978 (0.922, 0.998)
S2: NHS 24 Calls Flagged Covid (%)	0.718 (0.576 - 0.900)	0.976 (0.916 - 0.998)
S3: CSS App Predicted Positive Cases (%)	0.803 (0.645 - 1.002)	0.975 (0.911, 0.998)
S4: Covariates	0.716 (0.573 - 0.896)	0.945 (0.812, 0.994)
S5: Adj. NHS 24 Calls Flagged Covid (%)	0.637 (0.510 - 0.802)	0.946 (0.818, 0.994)
S6: Adj. CSS App Predicted Positive Cases (%)	0.717 (0.572 - 0.899)	0.943 (0.801 - 0.994))
S7: Adj. Both Key Variables	0.636 (0.507 - 0.801)	0.945 (0.814 - 0.994))

S1 - S3 are univariable and do not include any other covariates. S4 - S7 are adjusted and include covariates: % population under 5, % population over 84 and log population density.

Three adjusted spatial models are presented in table 7.8. The percentage of NHS 24 COVID-19 calls were positively associated with positive COVID-19 testing (OR = 1.038, 95% CrIs 1.024 - 1.052 per 1% increase) (**S5**). The percentage of CSS app COVID-19 predictions was also positively associated with positive COVID-19 testing (OR = 1.014, 95% Cred. I 0.974 - 1.056 per 1% increase), however, credible interval

spanned 1 (**S6**). Combing both NHS 24 and CSS app into the same model gave similar effect sizes and 95% CrIs as seen in separate models.

Table 7.8: Spatial CAR Leroux Models: Adjusted NHS 24 COVID calls, Adjusted CSS app COVID users and Fully Adjusted Model with both key variables with OR (95% CI). DIC, p.d and LMPL for model comparison.

	S5	S6	S7
NHS 24 COVID flagged calls (%)	1.038 (1.024, 1.052)	-	1.038 (1.025, 1.051)
CSS app predicted COVID user (%)	-	1.014 (0.974, 1.056)	1.012 (0.972, 1.053)
% Population aged under 5	1.073 (1.011, 1.139)	1.066 (0.999, 1.135)	1.070 (1.008, 1.135)
% Population aged over 84	1.123 (1.039, 1.214)	1.136 (1.044, 1.238)	1.126 (1.039, 1.219)
log population density	1.111 (1.053, 1.170)	1.120 (1.053, 1.190)	1.106 (1.047, 1.165)
DIC	2309.437	2329.222	2311.634
p.d	233.9769	241.4274	234.7981
LMPL	-1251.57	-1258.04	-1249.3

Deviance Information Criterion (DIC), the log Marginal Predictive Likelihood (LMPL) and corresponding estimated effective number of parameters (p.d).

Comparing model fit: NHS 24 model (**S5**) was the best fitting model according to DIC and LMPL, however, combined model (**S7**) showed similar results (table 7.8). The NHS 24 model (**S5**) and the combined model (**S7**) accounted for the most spatial variability ($\tau = 0.637$ and $\tau = 0.636$) compared to the intercept model ($\tau = 0.812$). The spatial covariate estimates were similar across all three models: population density, percentage of population under 5 and over 84 positively associated with COVID-19 positive testing, although CrIs spanned 1 for percentage of population aged under 5 in **S6**.

Spatial Model Predictions

A map of the final spatial model predictions was visualised to assess model smoothing and the effect on COVID-19 testing proportions. The final spatial model included COVID-19 NHS 24 calls and spatial covariates: log population density, age under 5 (%) and aged over 84 (%), however, did not include CSS app predictions.

Comparing crude COVID-19 testing to COVID-19 model predictions in figure 7.10, the distribution of COVID-19 positive testing was very similar between the crude and predicted maps with similar variability between areas. There was some smoothing, particularly over PCDs in the Highlands and North of Scotland, however, areas of high positive COVID-19 rates were detected in the predictions. The median (IQR) COVID-19 proportion was 5.4% (1.9% - 8.1%) and the COVID-19 predictions were 5.5% (2.4% - 7.4%). The spread was slightly reduced and maximum and minimum PCDs were slightly different: $\text{Min}_{PRED} = 0.5\%$ and $\text{Max}_{PRED} = 29.3\%$, whereas $\text{Min}_{CRUDE} = 0\%$ and $\text{Max}_{CRUDE} = 34.0\%$.

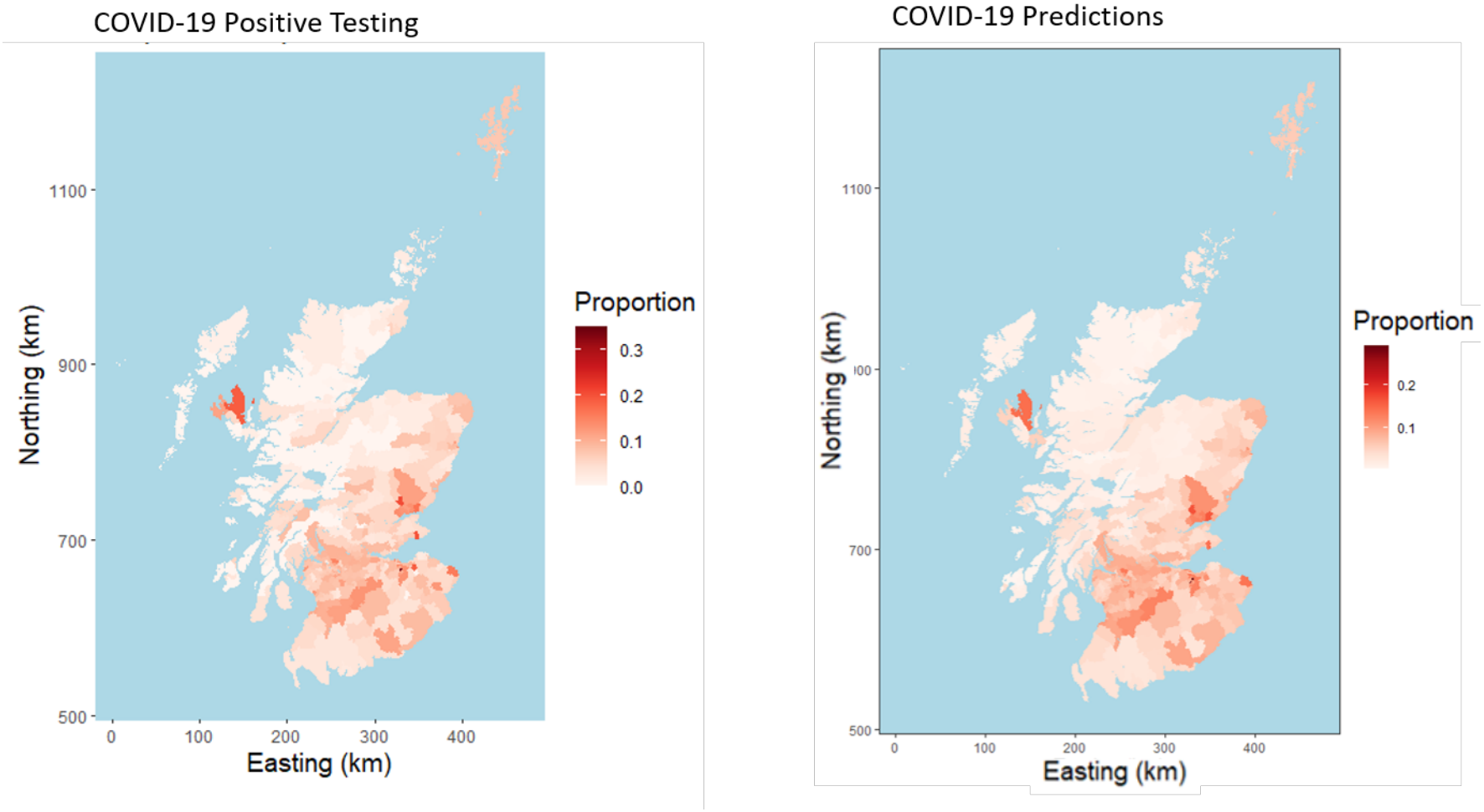


Figure 7.10: Comparison of crude COVID-19 positive testing proportions, from 30th March to June 15th, to COVID-19 positive testing predictions from final spatial model.

7.4 Spatio-Temporal Results

7.4.1 Exploratory Analysis

The median (IQR) number of tests per PCD per week was 14 (3 - 47), with a maximum of 427 tests. The median (IQR) percentage of positive tests per PCD per week was 0 (0 - 8.5%). Several PCDs show a 100% positive testing as some PCDs had very small numbers of tests which all returned positive cases. Many PCDs throughout the time period show a low, or zero, proportion of positive tests (table 7.9).

Table 7.9: Median (IQR) with maximum and minimum values for positive testing data by PCD and weeks.

	Number of Tests	Number of Positive Tests	Proportion of Positive Tests
Minimum	1	0	0.000
1st Qu	3	0	0.000
Median	14	0	0.000
3rd QU	47	3	0.085
Maximum	427	64	1.000

The COVID-19 positive testing data ranges from 2nd March to 15th June. The proportion of positive tests increases consistently from early March before reaching a peak between 23rd March and 6th April, at which point proportion of positive tests begins to decrease (figure 7.11).

The median (IQR) percentage of COVID-19 related NHS 24 calls per PCD per week was 28% (17% - 38%) with a maximum of 100% of calls. Similar to the testing data, there were a number of PCDs with small numbers of NHS 24 calls that were all identified as COVID-19-related calls giving a maximum proportion of 1 (table 7.10).

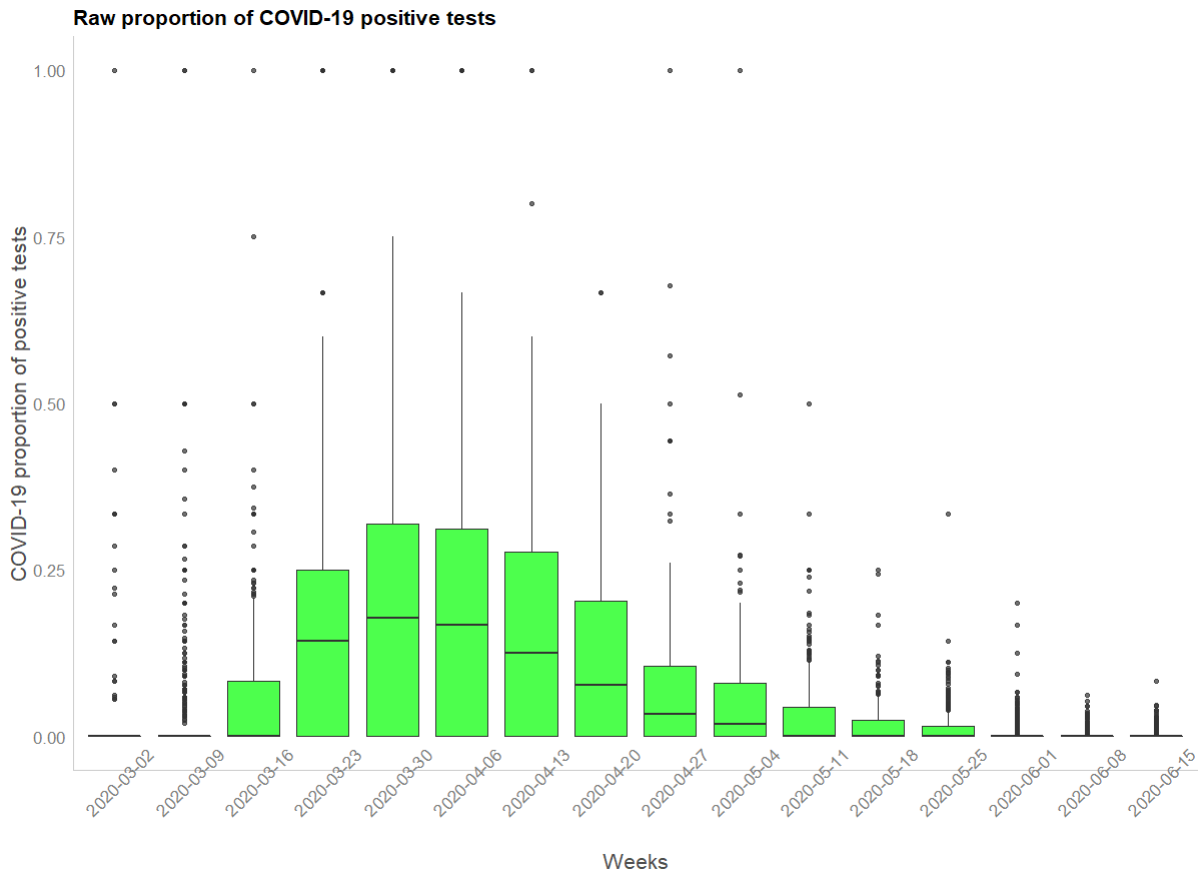


Figure 7.11: Proportion of positive COVID-19 tests by week from 2020-03-02 to 2020-06-15

Table 7.10: Median (IQR) with maximum and minimum values for NHS 24 calls data by PCD and Week.

	Total Number of NHS 24 Calls	COVID Flagged Calls	Proportion of COVID Flagged Calls
Minimum	1	0	0
1st Qu	8	2	0.17
Median	34	9	0.28
3rd Qu	100	28	0.38
Max	653	180	1

The proportions of NHS 24 COVID calls increased gradually from 3rd March onwards. The temporal curve for NHS 24 calls is flatter than corresponding plot for the testing data, reaching a maximum proportion of calls between 16th March and 13th April, with fluctuations between these dates. NHS 24 calls then slowly decrease from 20th April into the summer months. There is a wide spread of COVID related NHS calls between PCDs throughout all weeks (figure 7.12).

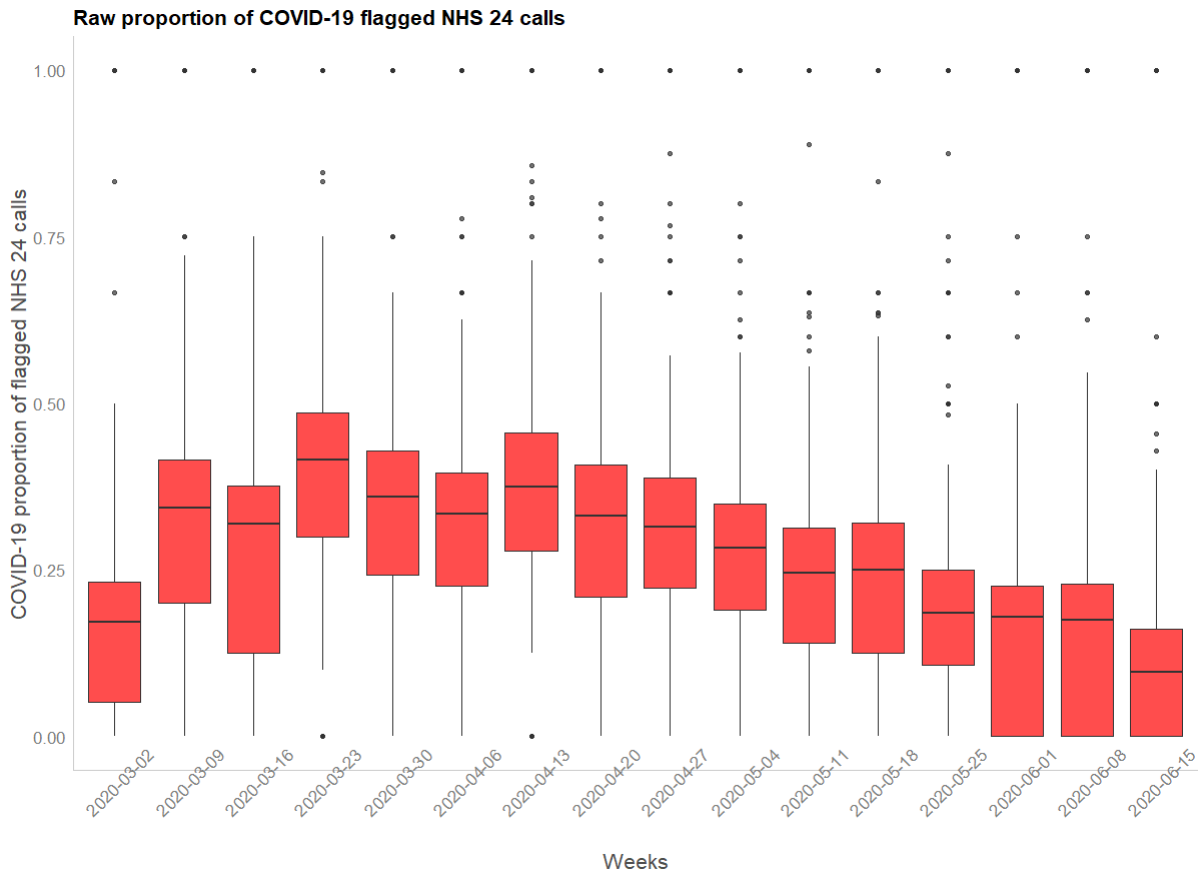


Figure 7.12: Proportion of COVID Flagged NHS 24 Calls by Week from 2020-03-02 to 2020-06-15

The median (IQR) percentage of predicted COVID-19 cases was 0.9% (0% - 1%) of active users per week, with a maximum of 25% of users predicted with COVID-19 in a week. The CSS app data does not include the same number of weeks in comparison to testing and NHS 24 data (table 7.11).

Table 7.11: Median (IQR) with maximum and minimum values for COVID-19 symptom study data by PCD and Week.

	Total Number of Active Users	Predicted COVID Cases	Proportion of Predicted COVID Cases
Minimum	1	0	0.000
1st Qu	26	0	0.000
Median	114	0	0.009
3rd Qu	315	2	0.010
Maximum	2218	54	0.2500

The CSS app data ranged from 30th March to 15th June. The median proportion of users with predicted COVID-19 decreases each week with the spread of data narrowing each week, however there are a number of outlying PCDs with high proportions of predicted cases throughout. This curve captures the decline from the peak of the first wave of the pandemic when comparing to the testing data (figure 7.13).

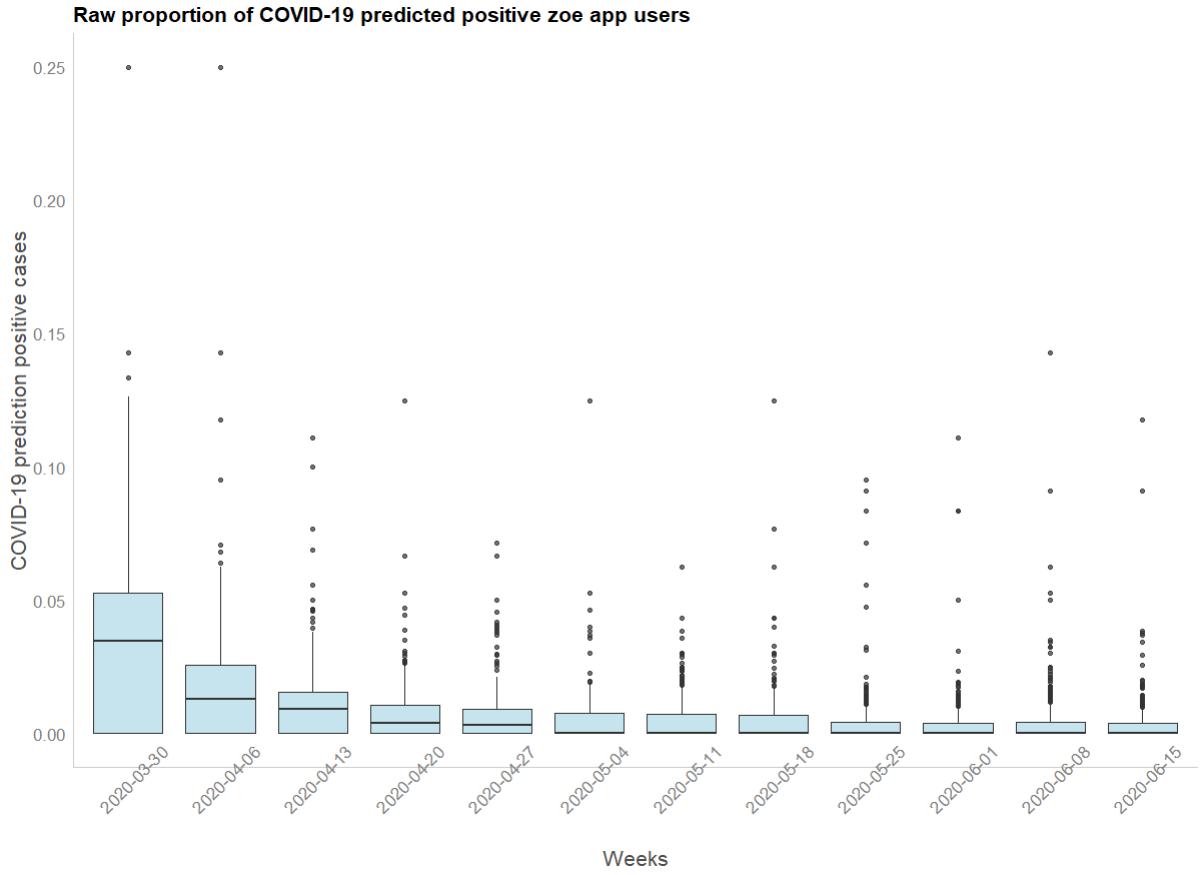


Figure 7.13: Proportion of Predicted COVID Cases by week from 2020-03-30 to 2020-06-15

CSS app COVID predicted users was seen to be more strongly correlated with COVID positive testing ($\rho = 0.371$) compared to NHS 24 COVID calls ($\rho = 0.281$). NHS 24 COVID calls and CSS app predicted COVID cases showed a positive correlation with each other, however, it is relatively weak ($\rho = 0.199$). Correlations between key variables were lower for the temporally varying data than observed in the spatially

aggregated data in table 7.5, therefore, it was expected that results would have less predictive power (table 7.13).

Table 7.12: Correlation matrix for raw positive testing, NHS 24 calls and CSS predicted cases data that vary by week and PCD.

	Positive Testing	NHS 24 Flagged Calls	CSS Predicted Cases
Positive Testing	1.000	0.281	0.371
NHS 24 Flagged Calls	0.281	1.000	0.199
CSS Predicted Cases	0.371	0.199	1.000

Temporal Autocorrelation

The results from the ACF showed COVID-19 positive testing to be correlated at lag 1. The NHS COVID-19 calls was also correlated at 1-lag, however, the CCS predicted COVID cases did not show any evidence of autocorrelation at a 5% significance level (figure 7.14).

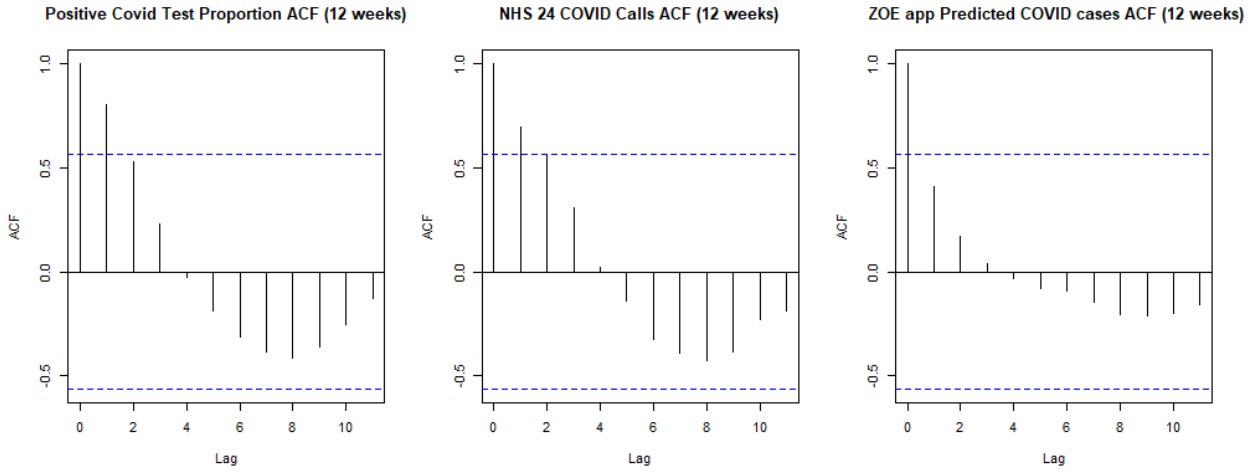


Figure 7.14: Autocorrelation function (AFC) for COVID-19 testing (left), NHS 24 calls (middle) and CSS predicted COVID cases (right).

The correlation statistics, lagging by 1 and 2 weeks, showed that at lag 1 there was strong positive correlation for positive COVID testing = 0.80 and NHS 24 COVID calls = 0.70, however, was much lower for ZOE app = 0.41 (as seen in figure 7.14). The strength of the correlation diminished at a 2-week lag: testing and NHS 24 were 0.53 and 0.56 respectively, CSS app also reduced further = 0.173.

7.4.2 Spatio-Temporal Model

Multiple spatio-temporal models with a multivariate auto-regressive process of order 1 were run to assess covariate effects on the spatio-temporal variance parameter (τ) and dependence parameters: ρS represents the spatial dependency parameter and ρT represents the temporal dependency parameter.

The spatio-temporal variability was estimated as $\tau = 0.53$ in the intercept only model (**ST1**). Introducing COVID-19 key predictors, NHS 24 (**ST2**) and CSS app COVID predictions (**ST3**) reduced the spatio-temporal variability comparably ($\tau = 0.507$ and $\tau = 0.510$, respectively), but were not very different from the intercept only model. The spatial covariates model accounted for very little variability and, remained similar to intercept (**ST1**). The spatio-temporal variability reduced the most with the full model which included all covariates, $\tau = 0.489$ (**ST7**). The spatial dependency parameter was very high for each model and very close to 1 ($\rho_S = 0.997$ full adjusted model, **ST7**). The temporal dependency parameter was also estimated high ($\rho_S = 0.858$ full adjusted model, **ST7**) and remained similar across all models (table 7.13).

The results for the three adjusted models are presented in table 7.14. Percentage of NHS 24 COVID calls was still seen to be positively associated with COVID-19 positive testing, however the strength of association was reduced in comparison to the spatial model (OR = 1.008, 95% CrIs 1.005 - 1.012) (**ST5**). The percentage of CSS app users predicted with COVID-19 per week per PCD was also seen to be positively associated

Table 7.13: Comparison of spatio-temporal model variability, spatial dependence and temporal dependence parameters for data by PCD and week with 95% credible intervals.

ST.CARar model	τ	ρS 95% Cred. I	ρT 95% Cred. I
1. Intercept Model	0.529 (0.446 - 0.623)	0.998 (0.996 - 0.999)	0.871 (0.825 - 0.913)
2. NHS 24 COVID flagged calls (%)	0.507 (0.425 - 0.599)	0.998 (0.996 - 0.999)	0.873 (0.827 - 0.917)
3. CSS app predicted COVID user (%)	0.510 (0.4285 - 0.602)	0.998 (0.996 - 0.999)	0.875 (0.829 - 0.918)
4. Spatial Covariates Only	0.525 (0.443 - 0.618)	0.998 (0.996 - 0.999)	0.854 (0.805 - 0.900)
5. Adj. NHS 24 COVID flagged calls (%)	0.503 (0.422 - 0.598)	0.998 (0.996 - 0.999)	0.855 (0.804 - 0.900)
6. Adj. CSS app predicted COVID user (%)	0.511 (0.427 - 0.609)	0.998 (0.996 - 0.999)	0.859 (0.810 - 0.904)
7. Adj. Both Key Variables	0.489 (0.407 - 0.581)	0.998 (0.995 - 0.999)	0.859 (0.808 - 0.905)

with positive COVID-19 testing (OR = 1.045, 95% Cred. I 1.019 - 1.073) (**Model ST6**). Combining both predictors into one model showed positive association for percentage of NHS 24 COVID-19 calls and CSS app *predicted* COVID-19 cases with positive COVID testing. Spatial covariate effects remained similar across all three models and are consistent with effect sizes seen in the spatial model (table 7.8). **ST5** showed the lowest DIC, however, the combined model (**ST7**) showed the best fitting model with lowest p.d and LMPL (table 7.14). The combined model (**ST7**) also accounted for the most spatio-temporal variability ($\tau = 0.4893$) compared to NHS 24 and The CSS models ($\tau = 0.503$ and $\tau = 0.\ 512$).

Table 7.14: Spatio-temporal AR(1) models: adjusted NHS 24 COVID calls, adjusted CSS app COVID users and fully adjusted model with both key variables with ORs and 95% credible intervals. DIC, p.d and LMPL are presented for model comparison.

	ST 5	ST 6	ST 7
NHS 24 COVID flagged calls (%)	1.008 (1.005, 1.012)	-	1.008 (1.005, 1.012)
CSS app predicted COVID user (%)	-	1.045 (1.019, 1.073)	1.043 (1.017, 1.071)
% Population aged under 5	1.065 (1.010, 1.121)	1.058 (1.005, 1.118)	1.054 (1.000, 1.111)
% Population aged over 84	1.080 (1.004, 1.162)	1.091 (1.013, 1.173)	1.088 (1.012, 1.170)
log population density	1.142 (1.097, 1.191)	1.135 (1.088, 1.185)	1.132 (1.084, 1.178)
DIC	11450.7	11454.35	11452.45
p.d	803.6914	810.1581	795.2783
LMPL	-5855.687	-5856.727	-5855.502

Lagged NHS 24 COVID calls and CSS app COVID predictions (1 - week)

Lagging key variables accounted for less spatio-temporal variability compared to results seen in table 7.15, with τ estimated higher than the intercept model (**ST1**) for all three lagged models (**ST8**, **ST9** and **ST10**), however, it should be noted that these results were modelled on a reduced data set and therefore are not necessarily comparable.

Table 7.15: Comparison of spatio-temporal model variability, spatial dependence and temporal dependence parameters with 95% credible intervals - 1-week lag for NHS 24 and CSS app outcomes.

ST.CARar model	τ	ρS 95% Cred. I	ρT 95% Cred. I
8. Adj. NHS 24 Calls Flagged Covid (%)	0.539 (0.448 - 0.644)	0.998 (0.996 - 0.999)	0.863 (0.811 - 0.912)
9. Adj. CSS App Predicted Positive Cases (%)	0.548 (0.457 - 0.655)	0.998 (0.996 - 0.999)	0.864 (0.813 - 0.912)
10. Adj. Both Key Variables	0.539 (0.449 - 0.645)	0.998 (0.996 - 0.999)	0.862 (0.810 - 0.911)

Model estimates were slightly weaker with the lagged covariates when compared with estimates seen in table 7.14, particularly for CSS app users. **ST8** was seen to fit these data best with the lowest DIC and p.d, however, **ST9** shows the lowest LMPL although all three models show comparable results for AIC, p.d. and LMPL (table 7.16). It is noted that DIC cannot be compared between table 7.14 and table 7.16 as the lagged analysis was based on less data (table 7.16).

Table 7.16: Spatio-Temporal AR(1) models with lagged covariates: adjusted NHS 24 COVID calls (lag 1), adjusted CSS app COVID users (lag 1) and fully adjusted model with both key variables (lag 1) with ORs and 95% credible intervals. DIC, p.d and LMPL presented for model comparison.

	ST 8	ST 9	ST 10
NHS 24 COVID flagged calls (%)	1.007 (1.004, 1.011)	-	1.007 (1.003, 1.011)
CSS app predicted COVID user (%)	-	1.012 (0.9850, 1.040)	1.011 (0.983, 1.039)
% Population aged under 5	1.070 (1.013, 1.132)	1.069 (1.008, 1.129)	1.067 (1.009, 1.127)
% Population aged over 84	1.106 (1.023, 1.196)	1.112 (1.028, 1.202)	1.120 (1.072, 1.172)
log population density	1.121 (1.070, 1.174)	1.124 (1.071, 1.176)	1.106 (1.047, 1.165)
DIC	10122.86	10129.36	10125.17
p.d	730.3278	735.9233	731.2189
LMPL	-5181.837	-5181.276	-5182.734

7.4.3 Sensitivity Analysis

Spatio-temporal ANOVA Model

The ST.CAR AR model assumes the same spatial structure across all time periods. This sensitivity analysis was conducted to assess whether the spatial pattern of COVID-19 positive testing changed over time. Moran's I test for spatial association was conducted for positive COVID testing on each week. Moran's I decreased throughout the weeks, implying that the strength of positive spatial association weakened over time, however the direction of association did not change (figure 7.15).

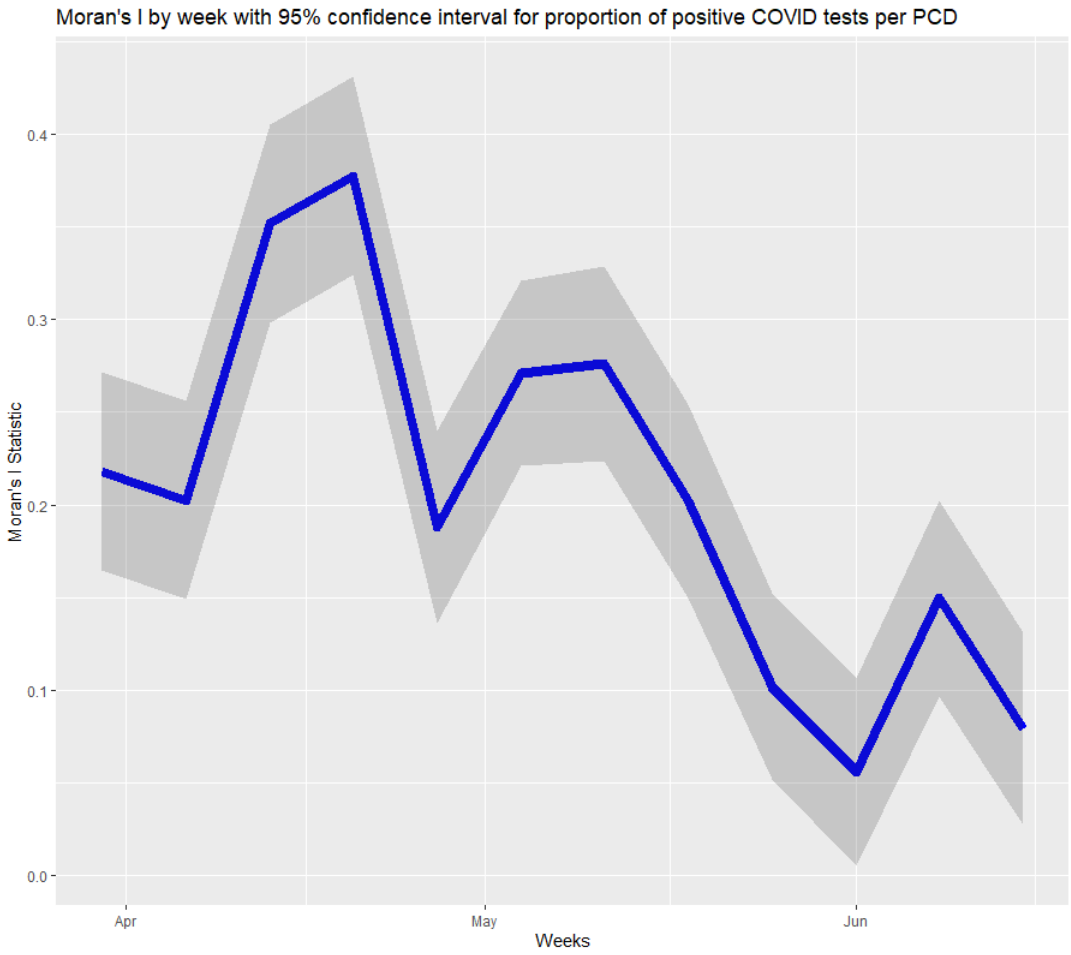


Figure 7.15: Moran's I by Week from April to June with 95% confidence interval for proportion of positive COVID-19 tests per PCD

The spatial distribution of the positive COVID testing data appeared similar for the first 2 weeks then decreased across the whole of Scotland in weeks 3 and 4. From then on, weeks 5 to 12, there were occasional instances of PCDs with high positive testing outbreaks, however, the majority of Scotland remained low. Moran's I ranged from 0.38 (4th April 2020) to 0.06 (1st June 2020) (figures 7.16 and 7.16).

The CAR ANOVA model was run for the fully adjusted model including NHS 24 COVID calls, CSS app predicted COVID cases and spatial covariates. The spatio-temporal ANOVA model allows for a space-time interaction, therefore allowing the spatial dependency to vary with time. The estimates of the fixed effects from the ST.CAR ANOVA model showed similar results to that seen in table 7.14 (**ST7**). This suggested that modelling these data allowing for a space-time interaction did not affect the estimates or interpretation of the model output. Model fit was not improved comparing DIC and LMPL, however, there were more effective samples (p.d.).

Table 7.17: Spatio-Temporal ANOVA model with OR and 95% credible intervals. DIC, p.d and LMPL presented.

	OR 95% Cred. I
NHS 24 COVID flagged calls (%)	1.010 (1.006, 1.014)
CSS app predicted COVID user (%)	1.046 (1.013, 1.080)
% Population aged under 5	1.059 (0.993, 1.127)
% Population aged over 84	1.122 (1.031, 1.222)
log population density	1.152 (1.090, 1.224)
τS	0.5640 (0.4251, 0.7431)
τT	0.2036 (0.0946, 0.5529)
τI	0.1781 (0.1530, 0.2057)
ρS	0.9615 (0.8476, 0.9961)
ρT	0.9770 (0.8737, 0.9981)
DIC	11753.88
p.d	1078.143
LMPL	-6088.466

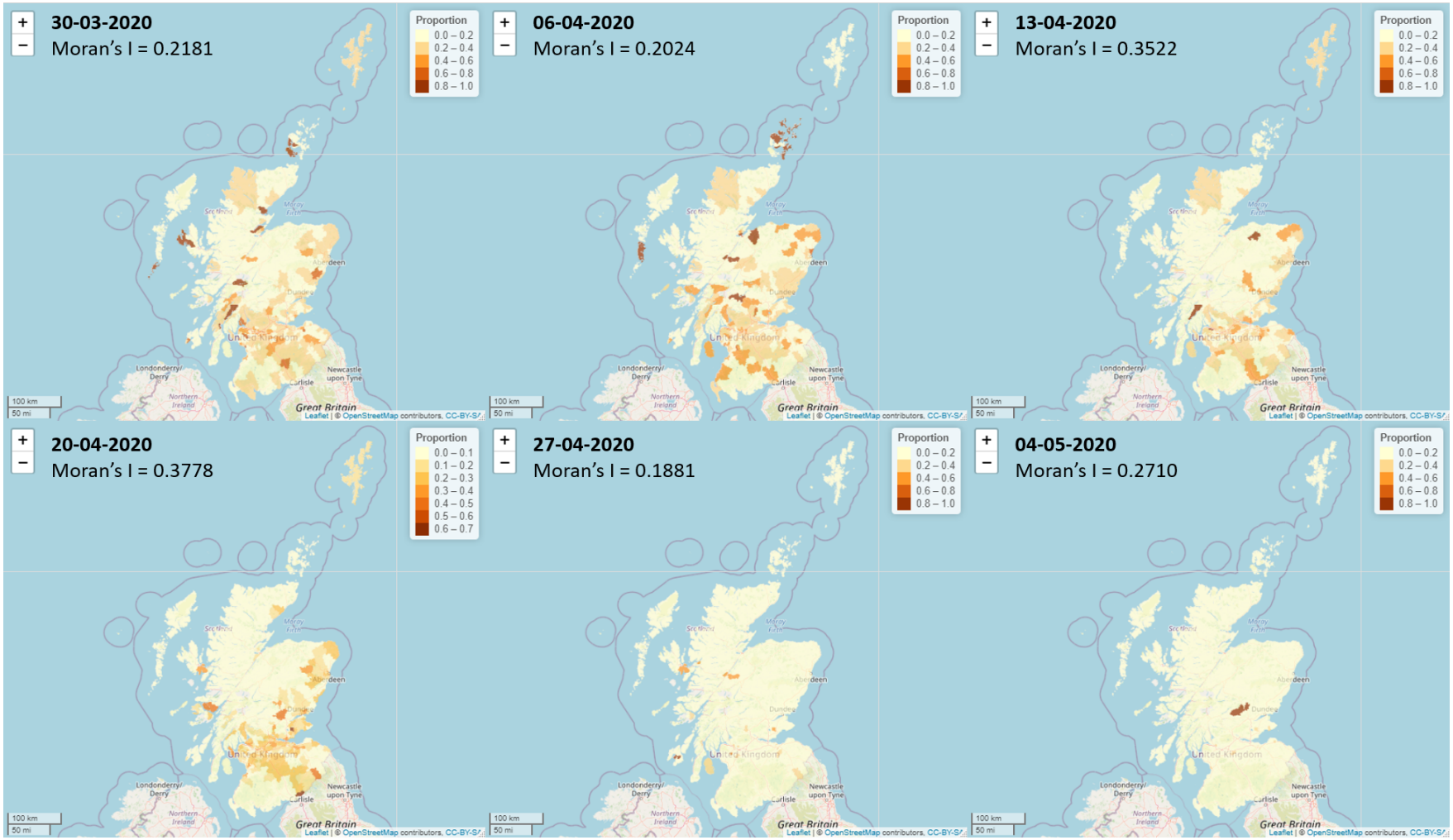


Figure 7.16: Areal maps of proportion of positive COVID-19 tests per week from: 30-03-2020 to 04-05-2020 with corresponding Moran's I statistic.

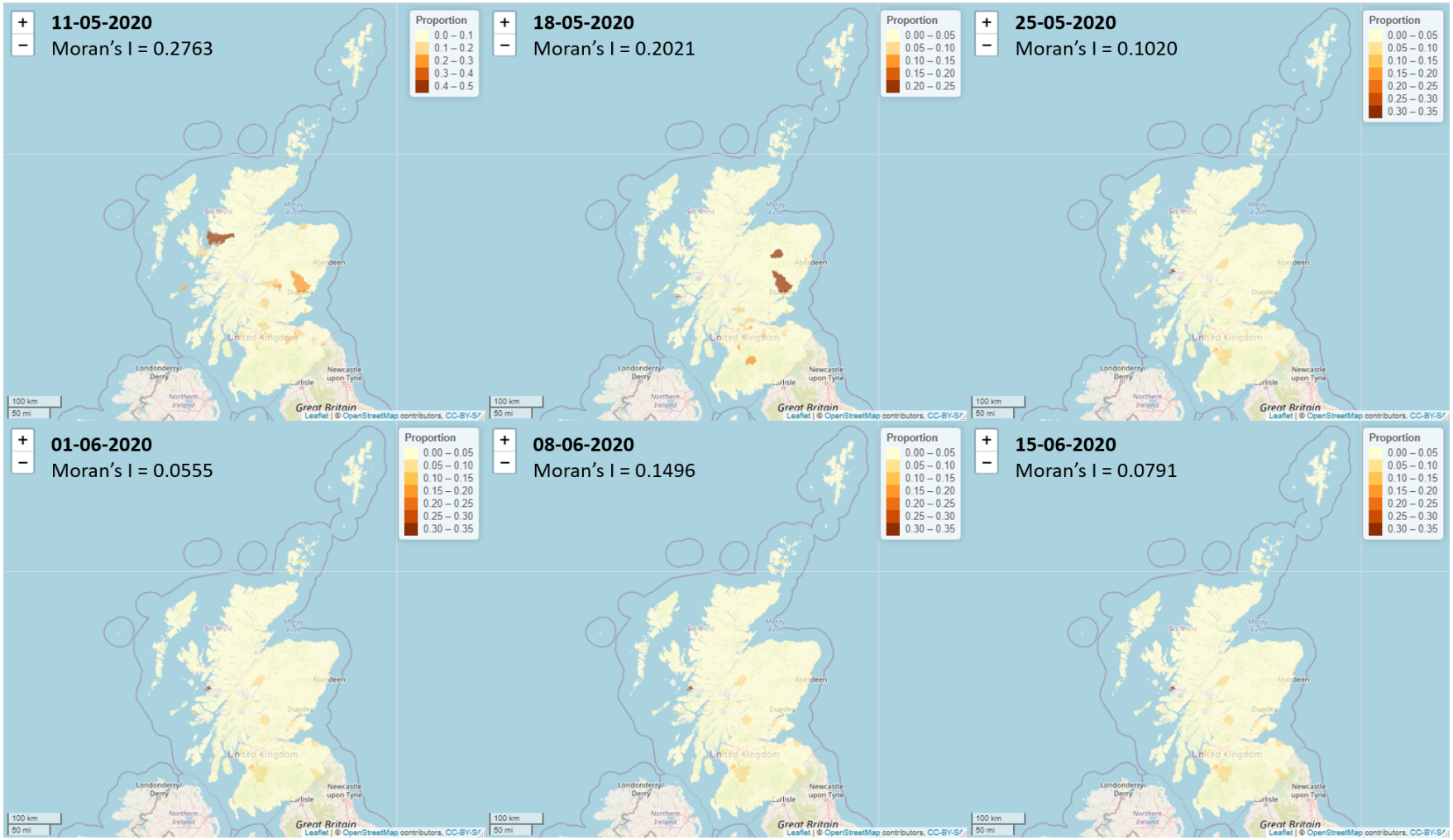


Figure 7.17: Areal maps of proportion of positive COVID-19 tests per week from: 11-05-2020 to 15-06-2020 with corresponding Moran's I statistic.

7.5 Discussion

The COVID-19 pandemic has highlighted the importance of disease surveillance using data from multiple sources which is imperative to help understand the spread of the virus across the UK and implement fast public health responses. This study explores the use of three different data sources of disease surveillance during the first wave of the COVID-19 pandemic in Scotland, showing NHS 24 and the CSS app data to be capable predictors for highlighting areas with increased COVID-19 positive testing. The national lockdown in Scotland was officially announced on 24th March 2020 and did not begin to ease until 29th May [75].

The primary spatial analysis showed strong positive spatial correlation between PCDs for COVID-19 positive testing (Moran's I = 0.537, $p<0.001$). Accounting for spatial autocorrelation in a spatial analysis and adjusting for population demographics, showed that increased percentage of NHS 24 COVID-19-related calls was associated with the proportion of positive COVID-19 testing (OR = 1.038, 95% Cred.In 1.024 - 1.052 per 1% increase): this implied a 45% increase in the odds of a positive COVID-19 test for a 10% increase in NHS 24 COVID-19 calls per PCD. The percentage of CSS app *predicted* COVID-19 cases also indicated a positive association with COVID-19 testing after adjusting for population demographics (OR= 1.014, 95% Cred. Int 0.974 - 1.056 per 1% increase), however, 95% CrI spanned 1. NHS 24 COVID-19 calls and CSS app *predicted* cases combined in to the same model gave similar effect sizes and credible intervals to the results stated previously. COVID-19 positive testing was associated with increased log population density, increased percentage of PCD population aged under 5 and over 84 years old. The adjusted model including NHS 24 COVID calls accounted for most spatial variability.

The adjusted spatio-temporal model in the secondary analyses showed a positive association between COVID-19 positive testing and the percentage of NHS 24 COVID calls (OR = 1.008, 95% Cred. Int 1.005 - 1.012): a 17% increased odd of testing for COVID-19 test for a 20% increase in COVID-19 related NHS 24 calls per PCD per week. Active CSS app users with *predicted* COVID-19 positive per week per PCD also showed an associated with COVID-19 positive testing (OR = 1.045, 95% Cred. Int 1.019 - 1.073), after adjusting for spatial covariates, which implied a 25% increase in the odds of testing positive for COVID-19 per PCD per week for a 5% increase in percentage of CSS app predicated COVID-19 cases per PCD per week. Combining both NHS 24 COVID-19 calls and active CSS app users *predicted* with COVID-19 into the same model produced comparable estimates and credible intervals while, accounting for more spatio-temporal variability than the separate models (τ_{both}=0.4893 vs τ_{NHS24}=0.5025 and τ_{CSS}=0.5117, implying a better fitting model.

The sensitivity analysis showed that the strength of the spatial autocorrelation fluctuated throughout the first wave of the pandemic, however allowing for a change in spatial association over time did not improve model fit. The contagious nature of COVID-19 implies that a level of spatial dependency between regions is expected however, reducing regional contact can lessen but does not eliminate, spatial contagion. The national lockdown encouraged the cross-border spread of COVID-19 to slow down in comparison to pre-lockdown rates, however, rates were seen to increase after lockdown eased [205].

Log population density was seen to have a positive association with COVID-19 positive testing in the spatially aggregated and spatio-temporal analyses. The transmission of COVID-19 is generally known to increase with the proximity of people, transmitting between those susceptible and those who are shedding the virus [206]. An association between population density and COVID-19 has been reported in many studies [206, 77, 201]. Areas of higher population density also tend to have higher numbers of

minority ethic residents and, increased deprivation and air pollution scores which are all risk factors of COVID-19 [201, 207]. This study also reported a positive association with percentage of population under 5 with COVID-19 testing. Although studies have suggested mild symptoms amongst children and reports of reopening of schools with half classes predicted that this was unlikely to increase the R number above 1 [208], a retrospective analyses in November 2020 stated that the reopening of schools in September 2020 had a significant impact on prevalence of COVID-19 in households [209]. An increased percentage of elderly population was also seen to be positively associated with proportions of positive COVID-19 testing. The risks of COVID-19 in elderly populations have been widely recognised, presumably due to lower immune responses and susceptibility [210].

These analyses identified a relationship between NHS 24 COVID-19-related calls and COVID-19 positive testing per PCD when aggregated over time, whereas this was not seen for the CSS app *predicted* COVID-19 cases per PCD (95% CrI spanned 1). Assessing these data by week found both the CSS app active users with predicted COVID-19 and NHS 24 calls to be predictors of positive COVID testing, although the association was stronger with CSS app. Comparing the distributions of COVID-19 testing, NHS 24 COVID-related calls and CSS active app users with *predicted* COVID-19 in figures 7.11, 7.12 and 7.13, show that the CSS study and COVID-19 testing follow a similar distribution from 30th March onwards. 30th March showed the peak of proportions across all three data streams. However, the proportions of positive COVID testing and CSS predicted app users decrease at a similar rate from 30th March onwards whereas, the median proportions of NHS 24 COVID-related calls remained high for approximately 5 weeks before decreasing slowly. This may describe the differences between the model estimates. NHS 24 COVID-19 calls are likely to pick up calls that are not truly related to COVID-19 through missclassifcation therefore introducing noise into the data [211], whereas the COVID-19 CSS predictions were more sparse (due to small numbers of

active app users in Scotland per week) but were more strongly related by week to the COVID testing data.

As two sources of disease surveillance, NHS 24 COVID-19 calls and CSS app COVID-19 prediction both provide information on real-time symptom reporting of COVID-19 within the community [166, 212, 211], however, the nature of these data are very different. NHS 24 is a service used by those looking for medical advice. The service is commonly used by those seeking advice regarding new symptoms or on behalf of someone else [213]. Therefore, NHS 24 is a method of surveillance by default but not by motive. For example, NHS 24 data has previously been used when monitoring the respiratory effects of the Icelandic volcanic ash plume in 2010. Symptoms of difficulty breathing, eye problems, coughing and rashes were noted to detect national exceedances [201].

Conversely, the CSS app is an altruistic data source with the primary goal of providing detailed information to better understand the COVID-19 pandemic. Population bias is expected with the CSS app data due to the lack of random sampling and is acknowledged as a limitation [200]. However, the richness and specific nature of these data are notably advantageous. A key example of the importance of real-time surveillance of COVID-19 was apparent during the introduction of localised lockdowns, with the first localised lockdown in Leicester on 30th June 2020. The CSS app data has been highlighted as a strong surveillance tool for detecting hotspots of COVID-19 outbreak [214]. Spatial modelling of these data at small-area resolution has been seen to produce strong near-real-time predictions of COVID-19 prevalence, showing its capacity to provide a safe and effective form of disease surveillance [166]. However, this is not to underestimate the potential predictive power of the NHS 24 data. A multivariate spatio-temporal (MVST) modelling of NHS 24 calls in Scotland during the first wave of the COVID-19 pandemic called attention to the temporal autoregressive nature of

these data and highlighted its suitability for making predictions of disease burden in the future [211].

Nevertheless, these fundamental differences between data sources may provide an explanation for the differing results seen in the spatial and spatio-temporal analyses, however, both surveillance sources were seen to be predictors of COVID-19 testing. The results from the spatio-temporal analysis also indicated that combining data sources accounted for the greatest spatio-temporal variability. This implied that using both data sources may provide strong predictability.

It is a noted limitation of this study was that the CSS app data were only available from the end of March until June, whereas data were available for NHS 24 and testing data from the beginning of March. Therefore, the data were subset to cover the same time period. In the first wave of the pandemic, the number of cases within the community began to decrease from April onwards, therefore these analyses assess the second half of the first wave [215] where the prevalence within the community was already decreasing. A further limitation of these analyses was that the access to COVID-19 testing was limited during this time period and therefore these data do not reflect the extent of the number of COVID-19 cases during this time. Furthermore, NHS 24 did not introduce a COVID-19 classification system until 14th April, therefore all classifications prior this time were back predicted (5-weeks) using a prediction model of NHS 24 calls from mid-April to the end of May relating to respiratory and gastrointestinal syndromes plus the patient's age. The prediction performance gave a specificity of 96% and a sensitivity of 75% with an area under the curve (AUC) of 0.88. These data were treated as observed data to ensure the peak of the first wave of the pandemic was included, however, the modelling of these data only include 2 weeks of predicted data.

Future work could utilise the predicted risks in each PCD and apply this as an explanatory variable to model COVID-19 related hospitalisations and deaths in the same PCD during the same time period, or in 2 - 4 weeks in the future. This could then be utilised as an external validation of this model and used as a means to identify areas of infection. Furthermore, future work may further investigate the assumptions made during these analyses such as the strict assumption of constant spatial association over time by allowing the spatial structure of COVID-19 to change over time. The relationship between covariates and COVID-19 positive testing is also assumed to be constant over time, however the relationship with covariates may change for different points during the pandemic. This could be assessed by including a random effect within the model and comparing between week variability or the inclusion of an interaction between covariates and time periods such as month.

In conclusion, The COVID Symptom Study (ZOE app) was shown in these data to be a predictor of sparse COVID-19 positive testing data during the first wave of the COVID-19 pandemic, therefore promoting its use more widely could be hugely beneficial in highlighting the spread of the virus and future outbreaks of disease. Nevertheless, pre-existing systems like NHS 24 should continue to be monitored as it is a well-trusted service that provides additional useful information. Both of these data sources provide high quality surveillance of COVID-19, however, utilising these data in parallel may provide an increasingly strong tool for COVID-19 surveillance to understand localised outbreaks, inform governance and identify future disease.

Chapter 8

Conclusion

8.1 Conclusion

The aim of this thesis was to explore the use of spatial and spatio-temporal methods in infectious diseases with particular application to *Clostridioides difficile* and COVID-19 infection. Routinely collected data were utilised throughout this thesis, on multiple spatial scales, which introduced a multilevel spatial data problem in some analyses. Individual level risk factors of *c-difficile* infections are well understood and highlighted in Chapter 1, such as the association with broad spectrum high-risk antibiotics (cephalosporins, co-amoxoclav, quinolones and clindamycin), however existing research lacks in-depth analyses of possible ecological risk factors and spatial components of the infection. Possible spatial determinants of *c-difficile* infection include cattle and population density and, therefore, the use of statistical spatial models was a focus of this thesis. The COVID-19 pandemic provided an opportunity to explore another infectious disease that poses a burden in hospital and community settings, these data were also explored using space-time risk models as the early part of the epidemic in Scotland showed strong spatial distributions with some parts of the country heavily affected and other parts relatively untouched.

This thesis began by exploring antibiotic prescribing by GP practices in Scotland between 2016 and 2018 in Chapter 3. The primary aim was to assess any spatial correlation between the rate of antibiotic prescribing by GP practices and how prescribing rates have changed over time. This study highlighted the disparity in antibiotic prescribing rates between areas of high and low deprivation and supports the need for deprivation adjusted antibiotic targets [150, 135]. Increased GP antibiotic prescribing was also seen for increased percentage of elderly practice populations, younger practice populations and for dispensing GP practice compared to non-dispensing practice. The secondary analyses of these data investigated the association between GP influenza vaccination uptake in patients aged over 65 and GP practice antibiotic prescribing rates, however there was no evidence of an association. These results indicate areas for potential intervention to aid the reduction of primary care antibiotic prescribing. Overall, total antibiotic prescribing rates were shown to decrease over time reflecting antibiotic stewardship efforts [216], however these trends varied between health boards and was less clear for prescriptions of high-risk antibiotic prescribing, therefore, these antibiotic classes may require more stewardship focus.

Chapter 4 then investigated the spatial and temporal distributions of CDI incidence in Scotland from 2014 to 2018. This analysis showed strong spatial correlation for total CDI incidence by intermediate zones (IZ), however, stratifying by healthcare-acquired and community-acquired CDI showed stronger spatial correlation for HA-CDI compared to CA-CDI. After adjusting for socio-demographic factors, some spatial correlation remained in these data and, therefore, were modelled using a conditional autoregressive (CAR) model to account for the spatial structure. The percentage of IZ population employment deprived was associated with an increase in total, healthcare-acquired and community-acquired CDI, with employment deprivation (%) accounting for 36% of the total spatial variability within the total CDI data. This relationship was supported by the literature [217, 158]. The overall temporal distribution of CDI appeared to de-

crease linearly over time, however there was some variation between financial quarters (seasonal effects), with evidence of an increase in summer months compared to winter. This was a noted disparity when compared to existing literature on CDI seasonality [218].

Unfortunately, this could not be investigated further in an extended spatio-temporal model, due to low counts of CDI at this level of spatial granularity and, therefore, these data were only investigated with yearly temporal variation. These analyses, again, showed an association between CDI incidence rates and a measure of increased deprivation, present for total, HA-CDI and CA-CDI. A spatio-temporal clustered trends model was also assessed, however, these data struggled to converge due to the low incidence rates across these time points. This problem was induced by the spatial granularity of these data as intermediate zones are a small areal unit and, therefore, lack variability over time with many zones having zero cases. Increasing the number of years of study, or increasing the spatial scale, would ensure larger CDI counts and, therefore, allow more adequate modelling of these data by quarter and to assess clustered trends. However, this is a trade-off as increasing the spatial scale may imply a loss of power to detect beneficial population-based risk factors.

The results from Chapter 3 and Chapter 4 then motivated an exploration of ecological risk factors on the casual pathway of CDI using spatial methods, although, these data were not available on the same intermediate zone spatial scale which introduced a multi-level spatial data problem. These analyses applied methods of spatial interpolations as a means to transform GP antibiotic prescribing point-location data, and incompatible areal-level environmental cattle density data, to match the CDI data by intermediate zones. This was achieved by making spatial predictions at the IZ centroids, to obtain a measure of GP antibiotic prescribing (including high-risk antibiotic prescribing) for 2016 to 2018 and cattle density by IZ. The spatio-temporal model from Chapter 4 was

applied to assess these ecological risk factors, adjusting for employment deprivation (%) which again showed a positive association with CDI incidence. There was no strong temporal variation estimated in this model, although, there were only three temporally varying points (2016 to 2018). However, the model did estimate strong spatial dependence within these data. Community-acquired CDI showed a positive association with cattle density which implied areas of high CA-CDI were associated with areas of high cattle density.

Similar environmental relationships have been previously reported but have previously not been observed in Scotland [171, 177]. HA-CDI showed a positive association with GP prescribing of clindamycin, however, an opposing relationship was seen for CA-CDI with a negative association with clindamycin and coamoxiclav. This was unexpected and difficult to explain as there are multiple individual-level studies supporting the associated use of these antibiotics with increased risks of CDI, particularly CA-CDI [58, 219, 195]. However, these findings must be considered in the contexts of ecological fallacy and further ecological assessment of this relationship would be required to determine if this is a repeatable observation, and if so try to understand possible explanations.

These results must also be handled sensitively due to a number of limitations presented during the transformation of these data. Firstly, the interpolation methods applied in these analyses are inherently subjective and it is difficult to completely eliminate biases. This transformation was also purely defined by distance from centroids as a measure of association, although, there is no guarantee the majority of a population reside in the centre of an IZ, similarly for the cattle density data. Additionally, the GP antibiotic prescribing data were defined as point-location spatial data set. A study of respiratory prescribing in Scotland discussed the spatial classification of GP antibiotic prescribing data [220], stating that these data are not entirely point-location data as its value is

representative of a surrounding population. However, these data are not strictly described at areal level due to overlapping of populations in urban areas. This issue was address in the thesis through a sensitivity analyses of varying interpolation smoothness. This showed consistent directionality of associations with varying interpolation smoothness suggesting that the conclusions around direction of the effect is relatively robust. Therefore, this method of transformation was appropriate for preliminary exploratory purposes.

Chapter 6 then explored CDI incidence by Welsh GP practices, to explore GP antibiotic prescribing as an ecological risk factor of total, inpatient and non-inpatient CDI. CDI cases were linked to GP practice with the definition of CDI cases being slightly different from those in Scotland: an inpatient CDI case was defined by a sample submitted from an inpatient hospital location irrespective of time in respect to admission date and non-inpatients were defined as samples submitted from non-inpatient locations (GP, AE or admission units) irrespective of time. This study design overcomes one of the issues with the analysis of the Scottish data: that of interpolating the GP prescribing data to an area level to link to area level data on CDI cases. However, these data were anonymised and as such, no formal spatial analyses could be conducted as locations of the GP practices were not available.

The results showed a positive association between GP prescribing of antibiotics, particularly the high risk antibiotic clindamycin, and the increased risk of CDI by GP practice. This highlighted that GP practice populations who were exposed to high prescribing of antibiotics, were at a predisposed increased risk of CDI infection. GP practice population demographics were also seen to have associations with CDI incidence: increased percentage of GP practice aged over 65 was associated with an increased risk of CDI and for an increased percentage of practice population with diabetes. There were also differences seen between the CDI incidences of Welsh health boards that could not be

explained by a measure of hospital antibiotic prescribing by health board, which may suggest an underlying spatial pattern and unaccounted for variation.

There were comparable aspects between the results from the Welsh and Scottish CDI studies such that they both displayed an association between CDI incidence and GP antibiotic prescribing of clindamycin [1], with both Chapter 6 and Chapter 3 reporting an isolated increase in GP antibiotic prescribing of clindamycin over time. A possible explanation for an increase in prescribing may be an association with penicillin allergies, an increase risk of CDI has been reported with the presence of a penicillin allergy with associated increase in clindamycin prescribing [193]. However, the CDI case definitions between these studies vary, making comparisons more complex. The Scottish study also highlighted an association with increased population deprivation, however, this was not seen for the percentage of Welsh GP practice populations residing in highly deprived areas. Although, the study in Wales did account for practice population health factors including diabetes, PPI prescribing, COPD and hypertension, and therefore, these difference may be partly described by a measure of health inequality. The Scottish CDI analyses assessed areal population demographics and the association with GP practice prescribing whereas the CDI Welsh study was entirely linked to GP practices and the demographics of those patients which is a fundamental difference in these studies.

Chapter 7 then presented an analyses of COVID-19 positive testing during the first wave of the 2020 coronavirus pandemic, and the association with two sources of disease surveillance. This analysis also handled a multilevel spatial data problem, transforming areal-to-areal data by proportion of postcodes. This transformation was possible due to nested link between postcodes, data zones and intermediate zones (similarly there is a link between postcodes and lower super output areas) which was not possible with the

cattle density data in Chapter 5. This study again highlights the burden of multilevel spatial scales in routinely collected data.

There was strong spatial correlation in the COVID-19 testing data during the first wave of the pandemic, however, assessing these data by week showed strong spatial correlation during the first few months of the pandemic, which decayed to represent spatial randomness into the summer months. This reflects the distribution of COVID-19 during a peak outbreak, however, the effects seen in later months are likely to be a result of lockdown, travel and social distancing restrictions that were put in place to minimise spread of infection between areas [75].

A spatial analysis showed that there was an association between positive COVID-19 testing by postcode district and increased percentage NHS 24 COVID-19 related calls. The percentage of COVID symptom study predictions of positive COVID-19 cases also indicated a positive association with positive COVID-19 testing by postcode district, although, the credible interval contained 1. These results indicate that areas reporting COVID-19 related symptoms to NHS 24 were associated with areas of high positive testing. There was also a positive association with increased population density, percentage of postcode district population aged over 84 years and under 5 years. A spatio-temporal analyses showed NHS 24 COVID-19-related calls and the COVID Symptom Study active app users predicted with positive COVID-19 per postcode district per week to be associated with positive COVID-19 testing. However, these associations weakened when lagging by one week with both symptom platforms. It would be of interest to explore the relationship between COVID-19 testing and spatial covariates over time, as it may be expected that the level of association would fluctuate similar to the spatial autocorrelation. This could be achieved with the inclusion of a temporal random effect

to compare between time-point variance or with an interaction term between spatial covariates and a measure of time such as months.

Spatial Modelling

This thesis has adopted the use of a CAR Leroux spatial generalised linear model throughout, with some analyses extended to include temporal random effects with an AR(1) process, seen in Chapters 4, 5 and 7. This class of spatial modelling is indicative of uncaptured (or *omitted*) spatially correlated variables, that if left ignored may affect the interpretation of results. As this thesis is primarily interested in understanding population-based risk factors of infectious diseases and the impact of these variables on spatial autocorrelation, this spatial modelling structure was, therefore, suited for these analyses.

The strength of spatial autocorrelation varies between the data sets assessed in this thesis. The COVID-19 data in Chapter 7 showed strong spatial autocorrelation (Moran's I = 0.54) for the crude COVID-19 testing data, whereas the CDI incidence data in Chapters 4 and 5, was much lower (Moran's I = 0.19). Assessing these spatial data sets using GLMs, and adjusting for population demographic information at an areal level, accounted for some of the spatial association within the data, however, both analyses displayed residual spatial autocorrelation: COVID-19 adjusted Moran's I = 0.30 and total CDI adjusted Moran's I = 0.15. This invalidated the independence assumption for the GLM residuals, providing a rationale for modelling these data spatially. For both of these analyses, model selection was carried out using a GLM structure, however, carrying these covariates forward into the spatial structure had important implications for the final choice of model covariates. This clearly illustrates the potential for misinterpretation if the spatial structure is not considered. In Chapter 4, forestry and fishery (%) was no longer seen a significant predictor in the CDI incidence analyses

when spatial random effects were included into the model structure, and similarly for the percentage of male population and Urban/ Rural classification in the COVID-19 analyses in Chapter 7. This is because the spatial GLM ensures more stable standard errors, which in turn, can affect the credible intervals of estimates. This highlights that, although the COVID-19 data displayed stronger autocorrelation compared to the CDI incidence data, it was equally important to account for the spatial autocorrelation in the residual.

Ecological vs. Individual-level Studies

Ecological studies are defined by the observational analyses of a population, or group of people, opposed to individual-level analyses. In an epidemiological setting, they are useful for defining prevalence or incidence in a population and investigate correlations with population exposures. This thesis has utilised aggregated routinely-collected data by area throughout, which lends itself to the benefits of ecological-based studies, that are simple, inexpensive and quick to conduct [221]. Individual-level studies are known to be more reliable in defining the strength of an association, particularly on a causal pathway, however, ecological studies are advantageous in defining population policy and highlighting specific area needs, especially when used in conjunction with a spatial framework. Ecological studies have also been shown to be stronger at determining population characteristics in context of population disease prevalence compared to individual-level studies [221].

This is notwithstanding the importance of ecological fallacy, which defines a type of confounding specific to ecological studies and warns that the relationships that exist for a population, or group of people, should not be assumed to be true at an individual-level. In Chapter 5, a negative association was seen between GP antibiotic prescribing of clindamycin and co-amoxiclav and CA-CDI incidence. These antibiotic classes are well-defined at an individual-level to be associated with a positive increase in the risk

of CA-CDI and, therefore, may be explained in the context of aggregation bias. This is not to infer that the ecological relationship is necessarily incorrect, however, population confounders are known to present differently to individual-level confounding which may provide an explanation. These differences may be accountable to aggregation-bias but given that the analyses in Chapter 6 is the first ecological study to assess this relationship, this result warrants further investigation at an ecological-level. Nonetheless, other results from this chapter were consistent with previous individual-level studies and with the results from the ecological study conducted in Wales, therefore these results are not to be disregarded.

The ecological studies discussed in this thesis are not intended to replace the work of individual-level studies, but to build on previous research and contribute to knowledge of these infectious diseases. The results of this thesis have implications on population health, which may aid public health policy decisions.

8.1.1 Future Work

Future work for the *Clostridiodies-difficile* analyses would be to obtain individual-level antibiotic prescribing linked to CDI cases by intermediate zone to reassess the association with antibiotic prescribing at an ecological level, allowing comparison to be made with the ecological results from the transformed GP antibiotic prescribing rates in Chapter 5. Conversely, linking CDI cases to GP practices, mirroring the analyses in Wales, would provide a more assured understanding of the risk of CDI to GP practice populations on prescribing in the community.

Methodologically there were limitation of these analyses which are highlighted by a similar study of respiratory GP prescribing and air pollution [220]. This study presented a novel spatio-temporal method of handling multilevel spatial data, particularly in reference to GP prescribing data. The novel process-convolution model introduced

in 2018, allows for a distance decay in spatial correlation based on surrounding GP prescribing rates, measured by euclidean distance. This allows some GP prescribing rates to be highly correlated and others to have no spatial dependence, whereas the analyses in this thesis assumed a constant spatial correlation determined by spatially close GP practices. An extension of the analyses in Chapter 5 would be to adopt this method and then compare model estimates to the current results. This would also provide a form of method comparison to assess the effectiveness of interpolation for assessing multilevel spatial data. A similar study of prescribing data in the UK adopted kernel density estimation as a method to smooth point-location GP data to be represented at an areal level. Kernel density estimation is a non-parametric smoothing method, which has similar properties to Kriging interpolation. The data in this study were similarly transformed based of the GP practices located nearby and noted the use of cross-validation to compare between results [157]. This method would have also been appropriate for these data, although the overall approach is comparable to the analyses already applied.

This study of GP respiratory prescribing [220] also has implications for the analyses in Chapter 3, which showed no evidence of spatial association at a GP practice level using Monte Carlo's test for spatial association. It would be of interest to apply this studies novel approach of defining the spatial structure and compare results when a strict distance decay is not enforced [220]. A study in England showed GP practice antibiotic prescribing have areal differences using a spatial cluster analyses, applying the getis-ord statistics for defining spatial clusters and modelling using a spatial cluster model [150]. This type of hot-spot analysis would have been appropriate for these data, and would be of interest to implement for future work. However, this approach was not adopted for this thesis as these analyses intended on utilising the variogram, as interpolation methods were applied later and, therefore, ensured continuity of methods.

The COVID-19 analyses could also be developed by using methods interpolation for transformation of incompatible spatial data to further explore the impact of multilevel spatial data transformations. Further work for the COVID-19 analyses would also include the assessment of model predictions. External validation of these data could be conducted by comparing to hospitalisations and mortality of COVID-19 during the same time period. Additionally, these analyses could be extended to include the second and third waves of the pandemic to explore the use of these symptom reporting platforms further into the pandemic. This would strengthen the knowledge of this model as a early detection method for future outbreaks.

Finally, future work that encompasses all of the topics described in this thesis would be to assess the impact of COVID-19 on the spread of CDI, and other healthcare associated infections, during the pandemic within community and hospital settings. COVID-19 is likely to continue as an endemic seasonal virus in the coming years, therefore it is crucial to use the knowledge gained throughout these first waves of the pandemic to minimise the rate of hospital-acquired COVID-19 infection [222]. GP antibiotic prescribing during the pandemic could be studied using the methods in this thesis. There are contradictory reports in regards to the rates of antibiotic prescribing during COVID-19. A study in Australia reported a 36% reduction in GP antibiotic prescribing after adjusting for appointment rate, highlighting areas of stewardship improvement [223], whereas a study in England reported a 15.5% decrease in prescription, although, after adjusting for absolute number of appointments this actually showed a 6.7% increase from previous years [70]. The amount of antibiotic prescribing in the community has a direct impact on the presence of healthcare associated and community spread infections like *c-difficile*. Therefore, changes in prescribing behaviours is likely to show changes in incidence.

The COVID-19 pandemic changed the day-to-day lives of people in the UK and worldwide. Social distancing, hand hygiene and mask wearing measures were implemented in the UK at the start of the pandemic to slow the rate of transmission of infection. It is reasonable to hypothesise that these measures might affect the rates of pre-existing community acquired and healthcare associated infections. It has been reported that rigorous hand hygiene practices in healthcare settings may have assisted in reducing HA-HAI's since the beginning of the pandemic, with one hospital reporting a 4-fold increase in the use of hand sanitizers [224]. Another study reported a significant reduction in the incidence of health-care associated CDI (HA-CDI). It was thought to be related to strategies that aimed to reduce the spread of microorganisms during the pandemic and highlighted the importance of maintaining "a level of attention" [225]. However, an increased risk of CDI was anticipated due to the large number of antibiotics that patients were prescribed in hospital due to COVID-19, with a particular risks for elderly patients [226]. A study in Rome observed an increase in HAIs, amongst patients with COVID-19, particularly related to hospital equipment [227]. There is limited research in these areas to date and there will be geographical difference in ecology, nonetheless these studies all highlight the importance of monitoring HAIs as they suggest rates of infections are susceptible to significant change with relatively small-scale differences in practice.

This work has displayed the beneficial use of spatial analyses in the context of *Clostridioes difficile* and COVID-19 infection, and the abilities to identify population risk factors to support the control and containment of infections over a larger spatial area. This thesis has also shown the capacities of routinely collected data for inferring on population health, however, it has simultaneously highlighted the challenges faced when handling multilevel spatial routinely-collected data.

Bibliography

[1] F. Tydeman, N. Craine, K. Kavanagh, H. Adams, R. Reynolds, V. McClure, H. Hughes, M. Hickman, and C. Robertson, "Incidence of Clostridioides difficile infection (CDI) related to antibiotic prescribing by GP surgeries in Wales," *Journal of Antimicrobial Chemotherapy*, 6 2021.

[2] K. V. Dalrymple, F. A. S. Tydeman, P. D. Taylor, A. C. Flynn, M. O'Keeffe, A. L. Briley, P. Santosh, L. Hayes, S. C. Robson, S. M. Nelson, N. Sattar, M. K. Whitworth, H. L. Mills, C. Singh, P. T. S. CStat, S. L. White, D. A. Lawlor, K. M. Godfrey, and L. Poston, "Adiposity and cardiovascular outcomes in three-year-old children of participants in UPBEAT, an RCT of a complex intervention in pregnant women with obesity," *Pediatric Obesity*, vol. 16, no. 3, p. e12725, 3 2021.

[3] M. Nana, F. Tydeman, G. Bevan, H. Boulding, K. Kavanagh, C. Dean, and C. Williamson, "Hyperemesis gravidarum is associated with increased rates of termination of pregnancy and suicidal ideation: results from a survey completed by $>$ 5000 participants," *American journal of obstetrics and gynecology*, vol. 224, no. 6, pp. 629–631, 6 2021.

[4] "Covid-19 in the UK: How many coronavirus cases are there in my area? - BBC News. Last accessed on 07/09/2021." [Online]. Available: https://www.bbc.co.uk/news/uk-51768274

[5] "Annual State of NHS Scotland Assets and Facilities Report for 2014 - gov.scot. Last accessed on 02/09/2021." [Online]. Available: https://www.gov.scot/publications/annual-state-nhsscotland-assets-facilities-report-2014/pages/3/

[6] "Semi-Variogram: Nugget, Range and Sill - GIS Geography. Last accessed on 02/09/2021." [Online]. Available: https://gisgeography.com/semi-variogram-nugget-range-sill/

[7] "How inverse distance weighted interpolation works—ArcGIS Pro — Documentation. Last accessed on 01/07/2021." [Online]. Available: https://pro.arcgis.com/en/pro-app/2.7/help/analysis/geostatistical-analyst/how-inverse-distance-weighted-interpolation-works.htm

[8] "Infection — definition of infection by Medical dictionary. Last accessed on 05/09/2021." [Online]. Available: https://medical-dictionary.thefreedictionary.com/infection

[9] "Infectious Diseases: Symptoms, Causes, Treatments. Last accessed on 05/09/2021." [Online]. Available: https://my.clevelandclinic.org/health/diseases/17724-infectious-diseases

[10] L. Luo, D. Liu, X.-l. Liao, X.-b. Wu, Q.-l. Jing, J.-z. Zheng, F.-h. Liu, S.-g. Yang, B. Bi, Z.-h. Li, J.-p. Liu, W.-q. Song, W. Zhu, Z.-h. Wang, X.-r. Zhang, P.-l. Chen, H.-m. Liu, X. Cheng, M.-c. Cai, Q.-m. Huang, P. Yang, X.-f. Yang, Z.-g. Han, J.-l. Tang, Y. Ma, and C. Mao, "Modes of contact and risk of transmission in COVID-19 among close contacts," 2020.

[11] "Impossible routes of HIV transmission — aidsmap. Last accessed on 05/09/2021." [Online]. Available: https://www.aidsmap.com/about-hiv/impossible-routes-hiv-transmission

[12] M. o. H. Government of Ontario and L.-T. Care, "E. coli Bacteria - Diseases and Conditions - Publications - Public Information - MOHLTC."

[13] "Infectious diseases - Symptoms and causes - Mayo Clinic. Last accessed on 06/08/2021." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351173

[14] R. J. Kim-Farley, "Principles of infectious disease control," *Oxford Textbook of Global Public Health*, pp. 1484–1506, 3 2015.

[15] "Chapter 1. What is epidemiology?. Last accessed on 05/09/2021." [Online]. Available: https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/1-what-epidemiology

[16] W. P. Hanage, "Pathogen Epidemiology," *Encyclopedia of Evolutionary Biology*, pp. 225–231, 4 2016.

[17] J. Murray and A. L. Cohen, "Infectious Disease Surveillance," *International Encyclopedia of Public Health*, p. 222, 10 2017.

[18] "New Zealand sees success in curbing Delta outbreak as new cases plunge — Reuters. Last accessed on 06/09/2021." [Online]. Available: https://www.reuters.com/world/asia-pacific/new-zealand-sees-success-curbing-delta-outbreak-new-cases-plunge-2021-09-03/

[19] D. Lewis, "Why many countries failed at COVID contact-tracing - but some got it right," *Nature*, vol. 588, no. 7838, pp. 384–387, 12 2020.

[20] "Antimicrobial Resistance and Healthcare-associated Infections Programme. Last accessed on 05/04/2021." *Control, European Centre for Disease Prevention and*. [Online]. Available: https://www.ecdc.europa.eu/en/about-us/who-we-are/disease-programmes/antimicrobial-resistance-and-healthcare-associated

[21] "WHO Coronavirus (COVID-19) Dashboard — WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. Last accessed on 05/09/2021." [Online]. Available: https://covid19.who.int/

[22] "Health Care-Associated Infections — health.gov. Last accessed on 08/08/2021." [Online]. Available: https://health.gov/our-work/health-care-quality/health-care-associated-infections

[23] "Preventing healthcare associated infection (HAI) - Better Health Channel. Last accessed on 06/08/2021." [Online]. Available: https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/infections-in-hospital-reduce-the-risk#what-are-healthcare-associated-infections

[24] "Hand hygiene helps reduce HCAIs. Last accessed on 06/08/2021." [Online]. Available: https://www.openaccessgovernment.org/hand-hygiene-2/66768/

[25] "Introduction — Healthcare-associated infections: prevention and control in primary and community care — Guidance — NICE. Last accessed on 08/06/2021." [Online]. Available: https://www.nice.org.uk/guidance/cg139

[26] "Healthcare-Associated Infections — Healthy People 2020. Last accessed on 08/06/2021." [Online]. Available: https://www.healthypeople.gov/2020/topics-objectives/topic/healthcare-associated-infections

[27] "Health Care Associated Infection (HCAI). Last accessed on 06/08/2021. ." [Online]. Available: https://www.blackpooljsna.org.uk/Living-and-Working-Well/Health-Protection/Health-Care-Associated-Infection.aspx

[28] "MRSA: Contagious, Symptoms, Causes, Prevention, Treatments. Last accessed on 06/08/2021." [Online]. Available: https://www.webmd.com/skin-problems-and-treatments/understanding-mrsa

[29] "What are superbugs and how can I protect myself from infection? - Mayo Clinic. Last accessed on 06/08/2021." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/infectious-diseases/expert-answers/superbugs/faq-20129283

[30] R. Barranco, L. V. B. D. Tremoul, and F. Ventura, "Hospital-Acquired SARS-Cov-2 Infections in Patients: Inevitable Conditions or Medical Malpractice?" *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, pp. 1–9, 1 2021.

[31] D. Oliver, "David Oliver: Deaths from hospital acquired covid are everyone's problem," *BMJ*, vol. 373, 6 2021.

[32] A. F. Monegro, V. Muppidi, and H. Regunath, "Hospital Acquired Infections," *Cambridge Handbook of Psychology, Health and Medicine, Second Edition*, pp. 736–738, 9 2020.

[33] "Disease Outbreak Control Division — Community acquired Infections. Last accessed on 05/09/2021." [Online]. Available: https://health.hawaii.gov/docd/disease-types/community-infections/

[34] E. Finn, F. L. Andersson, and M. Madin-Warburton, "Burden of Clostridioides difficile infection (CDI) - a systematic review of the epidemiology of primary and recurrent CDI," *BMC Infectious Diseases 2021 21:1*, vol. 21, no. 1, pp. 1–11, 5 2021.

[35] C. Robertson, J. Pan, K. Kavanagh, I. F. Ford, C. McCowan, M. Bennie, C. Marwick, and A. Leanord, "Cost burden of Clostridioides difficile infection to the health service: A population based case-control study in Scotland," *Journal of Hospital Infection*.

[36] S. Rhea, K. Jones, S. Endres-Dighe, B. Munoz, D. J. Weber, R. Hilscher, J. MacFarquhar, E. Sickbert-Bennett, L. DiBiase, A. Marx, J. Rineer, J. Lewis, and G. Bobashev, "Modeling inpatient and outpatient antibiotic stewardship interventions to reduce the burden of Clostridioides difficile infection in a regional healthcare network," *PLoS ONE*, vol. 15, no. 6, 6 2020.

[37] "Scottish antimicrobial use and resistance in humans in 2018. Last Accessed on 16/03/2021," *Health Protection Scotland; ISD Scotland; Scottish Antimicrobial Prescribing Group; Scottish Medicines Consortium.*

[38] D. A. Leffler and J. T. Lamont, "Clostridium difficile infection," pp. 1539–1548, 4 2015.

[39] J. Czepiel, M. Dróżdż, H. Pituch, E. J. Kuijper, W. Perucki, A. Mielimonka, S. Goldman, D. Wultańska, A. Garlicki, and G. Biesiada, "Clostridium difficile infection: review," 2019.

[40] S. Khanna, D. S. Pardi, S. L. Aronson, P. P. Kammer, R. Orenstein, J. L. St Sauver, W. S. Harmsen, and A. R. Zinsmeister, "The epidemiology of community-acquired clostridium difficile infection: A population-based study," *American Journal of Gastroenterology*, vol. 107, no. 1, pp. 89–95, 2012.

[41] "HPS Website - Clostridioides difficile infection (CDI). Last Accessed on 06/09/2021." [Online]. Available: https://www.hps.scot.nhs.uk/a-to-z-of-topics/clostridioides-difficile-infection/

[42] "The Vale of Leven Hospital Inquiry Report. Last Accessed on 06/09/2021," 2014. [Online]. Available: www.valeoflevenhospitalinquiry.org

[43] R. Cunningham and S. Dial, "Is over-use of proton pump inhibitors fuelling the current epidemic of Clostridium difficile-associated diarrhoea?"

[44] A. Y. Guh, S. H. Adkins, Q. Li, S. N. Bulens, M. M. Farley, Z. Smith, S. M. Holzbauer, T. Whitten, E. C. Phipps, E. B. Hancock, G. Dumyati, C. Concannon, M. A. Kainer, B. Rue, C. Lyons, D. M. Olson, L. Wilson, R. Perlmutter, L. G. Winston, E. Parker, W. Bamberg, Z. G. Beldavs, V. Ocampo, M. Karlsson, D. N. Gerding, and L. C. McDonald, "Risk Factors for Community-Associated Clostridium difficile Infection in Adults: A Case-Control Study," *Open Forum Infectious Diseases*, vol. 4, no. 4, 10 2017.

[45] L. C. Ahyow, P. C. Lambert, D. R. Jenkins, K. R. Neal, and M. Tobin, " Bed Occupancy Rates and Hospital-Acquired Clostridium difficile Infection: A Cohort Study ," *Infection Control & Hospital Epidemiology*, vol. 34, no. 10, pp. 1062–1069, 10 2013.

[46] K. E. Dingle, X. Didelot, T. P. Quan, D. W. Eyre, N. Stoesser, T. Golubchik, R. M. Harding, D. J. Wilson, D. Griffiths, A. Vaughan, J. M. Finney, D. H. Wyllie, S. J. Oakley, W. N. Fawley, J. Freeman, K. Morris, J. Martin, P. Howard, S. Gorbach, E. J. Goldstein, D. M. Citron, S. Hopkins, R. Hope, A. P. Johnson, M. H. Wilcox, T. E. Peto, A. S. Walker, D. W. Crook, C. Del Ojo Elias, C. Crichton, V. Kostiou, A. Giess, and J. Davies, "Effects of control interventions on Clostridium difficile infection in England: an observational study," *The Lancet Infectious Diseases*, vol. 17, no. 4, pp. 411–421, 4 2017.

[47] D. W. Eyre, M. L. Cule, D. J. Wilson, D. Griffiths, A. Vaughan, L. O'Connor, C. L. Ip, T. Golubchik, E. M. Batty, J. M. Finney, D. H. Wyllie, X. Didelot, P. Piazza, R. Bowden, K. E. Dingle, R. M. Harding, D. W. Crook, M. H. Wilcox, T. E. Peto, and A. S. Walker, " Diverse Sources of C. difficile Infection Identified on Whole-Genome Sequencing ," *New England Journal of Medicine*, vol. 369, no. 13, pp. 1195–1205, 9 2013.

[48] "Public Health England Annual CDI Report. Last Accessed on 16/03/2021." [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/724368/CDI_summary_2018.pdf

[49] D. P. Durham, M. A. Olsen, E. R. Dubberke, A. P. Galvani, and J. P. Townsend, "Quantifying Transmission of Clostridium difficile within and outside Healthcare Settings," *Emerging Infectious Diseases*, vol. 22, no. 4, p. 608, 4 2016.

[50] J. Jou, J. Ebrahim, F. S. Shofer, K. W. Hamilton, J. Stern, and J. H. Han, "Environmental Transmission of Clostridium difficile: Association between Hospital Room Size and C. difficile Infection," *Infection Control and Hospital Epidemiology*, vol. 36, no. 5, pp. 564–568, 2015.

[51] D. Z. Bliss, S. Johnson, K. Savik, C. R. Clabots, K. Willard, and D. N. Gerding, "Acquisition of Clostridium difficile and Clostridium difficile- associated diarrhea in hospitalized patients receiving tube feeding," *Annals of Internal Medicine*, vol. 129, no. 12, pp. 1012–1019, 12 1998.

[52] M. K. Shaughnessy, R. L. Micielli, D. D. DePestel, J. Arndt, C. L. Strachan, K. B. Welch, and C. E. Chenoweth, " Evaluation of Hospital Room Assignment and Acquisition of Clostridium difficile Infection ," *Infection Control & Hospital Epidemiology*, vol. 32, no. 3, pp. 201–206, 3 2011.

[53] D. J. Anderson, L. F. Rojas, S. Watson, L. P. Knelson, S. Pruitt, S. S. Lewis, R. W. Moehring, E. E. Sickbert Bennett, D. J. Weber, L. F. Chen, and D. J. Sexton, "Identification of novel risk factors for community-acquired Clostridium difficile infection using spatial statistics and geographic information system analyses," *PLoS ONE*, vol. 12, no. 5, 5 2017.

[54] L. S. Muñoz-Price, R. Hanson, S. Singh, A. B. Nattinger, A. Penlesky, B. W. Buchan, N. A. Ledeboer, K. Beyer, S. Namin, Y. Zhou, and L. E. Pezzin, "Association between Environmental Factors and Toxigenic Clostridioides difficile Carriage at Hospital Admission," *JAMA Network Open*, vol. 3, no. 1, p. 1919132, 1 2020.

[55] M. Jahangir Alam, S. T. Walk, B. T. Endres, E. Basseres, M. Khaleduzzaman, J. Amadio, W. L. Musick, J. L. Christensen, J. Kuo, R. L. Atmar, and K. W. Garey, "Community environmental contamination of toxigenic Clostridium difficile," *Open Forum Infectious Diseases*, 2017.

[56] S. E. . Manzoor, C. A. M. . Mcnulty, D. . Nakiboneka-Ssenabulya, D. M. . Lecky, K. . Hardy, and P. Hawkey, "Investigation of community carriage rates of Clostridium difficile and Hungatella hathewayi in healthy volunteers from four regions of England," *The Journal of hospital infection*, vol. 97, no. 2, pp. 153–155, 2017.

[57] A. Deshpande, V. Pasupuleti, P. Thota, C. Pant, D. D. K Rolston, T. J. Sferra, A. V. Hernandez, and C. J. Donskey, "Community-associated Clostridium difficile infection and antibiotics: a meta-analysis."

[58] K. Kavanagh, J. Pan, C. Marwick, P. Davey, C. Wiuff, S. Bryson, C. Robertson, and M. Bennie, "Cumulative and temporal associations between antimicrobial prescribing and community-associated Clostridium difficile infection: Population-based case-control study using administrative data," *Journal of Antimicrobial Chemotherapy*, 2017.

[59] J. Pan, K. Kavanagh, C. Marwick, P. Davey, C. Wuiff, S. Bryson, C. Robertson, and M. Bennie, "Residual effect of community antimicrobial exposure on risk of hospital onset healthcare-associated Clostridioides difficile infection: a case–control study using national linked data," *Journal of Hospital Infection*, vol. 103, no. 3, pp. 259–267, 11 2019.

[60] R. C. Owens, Jr., C. J. Donskey, R. P. Gaynes, V. G. Loo, and C. A. Muto, " Antimicrobial-Associated Risk Factors for Clostridium difficile Infection ," *Clinical Infectious Diseases*, vol. 46, no. s1, pp. S19–S31, 1 2008.

[61] D. Baur, B. P. Gladstone, F. Burkert, E. Carrara, F. Foschi, S. Döbele, and E. Tacconelli, "Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and Clostridium difficile infection: a systematic review and meta-analysis," *The Lancet Infectious Diseases*, vol. 17, no. 9, pp. 990–1001, 9 2017.

[62] A. Patton, P. Davey, S. Harbarth, D. Nathwani, J. Sneddon, and C. A. Marwick, "Impact of antimicrobial stewardship interventions on Clostridium difficile infection and clinical outcomes: segmented regression analyses," *Journal of Antimicrobial Chemotherapy*, vol. 73, no. 2, pp. 517–526, 2 2018.

[63] "Primary Care Antimicrobial Guidelines. Last Accessed on 16/03/2021." Tech. Rep., 2015.

[64] "Priority Areas - Stewardship: 4C Antimicrobials — CPD for General Practitioners. Last Accessed on 16/03/2021." [Online]. Available: https://gpcpd.heiw.wales/clinical/all-wales-national-prescribing-indicators-20-21/priority-areas-stewardship-4c-antimicrobials/

[65] "Scottish antimicrobial use and resistance in humans in 2015. Last Accessed on 16/03/2021." 2015. [Online]. Available: http://www.isdscotland.org/Health-Topics/Prescribing-and-Medicines/Publications/2016-08-30/2016-08-30-SAPG-2015-Report.pdf

[66] "An Official Statistics publication for Scotland. Last Accessed on 08/08/2021." [Online]. Available: https://www.statisticsauthority.gov.uk/osr/code-of-practice/

[67] C. Llor and L. Bjerrum, "Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem," *Therapeutic Advances in Drug Safety*, vol. 5, no. 6, p. 229, 2014.

[68] "Respiratory Tract Infections - Antibiotic Prescribing. Last Accessed on 07/08/2021." *NICE Clinical Guidelines 69*, vol. 69, no. July, pp. 1–240, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK53632/

[69] W. Malcolm, R. A. Seaton, G. Haddock, L. Baxter, S. Thirlwell, P. Russell, L. Cooper, A. Thomson, and J. Sneddon, "Impact of the COVID-19 pandemic on community antibiotic prescribing in Scotland," *JAC-Antimicrobial Resistance*, vol. 2, no. 4, 10 2020.

[70] R. Armitage and L. B. Nellums, "Antibiotic prescribing in general practice during COVID-19," *The Lancet Infectious Diseases*, vol. 21, no. 6, p. e144, 6 2021.

[71] A. Cole, "GPs feel pressurised to prescribe unnecessary antibiotics, survey finds," *BMJ*, vol. 349, 8 2014.

[72] "Antibiotic Awareness. Last Accessed on 07/09/2021." [Online]. Available: https://www.sapg.scot/education-resources/antibiotic-awareness/

[73] "Timeline - WHO's COVID-19 response. Last Accessed on 26/08/2021." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline/#!

[74] "NHSGGC : COVID 19 - March 2020. Last Accessed on 15/02/2021." [Online]. Available: https://www.nhsggc.org.uk/your-health/public-health/public-health-protection-unit-phpu/covid-19-march-2020/#

[75] "Timeline of Coronavirus (COVID-19) in Scotland – SPICe Spotlight — Solas air SPICe. Last Accessed on 15/02/2021." [Online]. Available: https://spice-spotlight.scot/2021/02/12/timeline-of-coronavirus-covid-19-in-scotland/

[76] "Symptoms of COVID-19 — CDC. Last Accessed on 07/08/2021." [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[77] I. Bray, A. Gibson, and J. White, "Coronavirus disease 2019 mortality: a multivariate ecological analysis in relation to ethnicity, population density, obesity, deprivation and pollution," *Public Health*, vol. 185, pp. 261–263, 8 2020.

[78] N. I. Lone, J. McPeake, N. I. Stewart, M. C. Blayney, R. C. Seem, L. Donaldson, E. Glass, C. Haddow, R. Hall, C. Martin, M. Paton, A. Smith-Palmer, C. T. Kaye, and K. Puxty, "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*, vol. 1, p. 100005, 2 2021.

[79] O. Byambasuren, M. Cardona, K. Bell, J. Clark, M.-L. McLaws, and P. Glasziou, "Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis," *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada*, vol. 5, no. 4, pp. 223–234, 12 2020.

[80] S. M. Moghadas, M. C. Fitzpatrick, P. Sah, A. Pandey, A. Shoukat, B. H. Singer, and A. P. Galvani, "The implications of silent transmission for the control of COVID-19 outbreaks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 30, pp. 17 513–17 515, 7 2020.

[81] "COVID-19: How is this wave different from the first? – SPICe Spotlight — Solas air SPICe. Last Accessed on 09/06/2021." [Online]. Available: https://spice-spotlight.scot/2020/10/08/covid-19-how-is-this-wave-different-from-the-first/

[82] "Enhanced Surveillance of COVID-19 in Scotland - Population-based seroprevalence surveillance 11 August 2021 - Enhanced Surveillance of COVID-19 in Scotland - Publications - Public Health Scotland. Last Accessed on 07/09/2021." [Online]. Available: https://publichealthscotland.scot/publications/enhanced-surveillance-of-covid-19-in-scotland/enhanced-surveillance-of-covid-/19-in-scotland-population-based-seroprevalence-surveillance-11-august-2021/

[83] H. J. Miller, "Tobler's First Law and Spatial Analysis," *Source: Annals of the Association of American Geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[84] "Mapping the Cholera Epidemic of 1854 — National Geographic Society. Last Accessed on 07/09/2021." [Online]. Available: https://www.nationalgeographic.org/activity/mapping-cholera-epidemic-1854/

[85] L. A. Waller and B. P. Carlin, "Disease mapping," *Chapman & Hall/CRC handbooks of modern statistical methods*, vol. 2010, p. 217, 3 2010.

[86] S. I. Hay, D. B. George, C. L. Moyes, and J. S. Brownstein, "Big Data Opportunities for Global Infectious Disease Surveillance," *PLOS Medicine*, vol. 10, no. 4, p. e1001413, 4 2013.

[87] G. Chowell and R. Rothenberg, "Spatial infectious disease epidemiology: On the cusp," p. 192, 10 2018.

[88] R. S. Evans, "Electronic Health Records: Then, Now, and in the Future," *Yearbook of Medical Informatics*, no. Suppl 1, p. S48, 5 2016.

[89] "Electronic Health Records — NHS Research Scotland — NHS Research Scotland. Last Accessed on 07/09/2021." [Online]. Available: https://www.nhsresearchscotland.org.uk/research-in-scotland/data/sub-page-4

[90] "ISD Scotland — Information Services Division. Last Accessed on 07/09/2021." [Online]. Available: https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?ID=128&Title=CHI%20Number

[91] S. de Lusignan and C. van Weel, "The use of routinely collected computer data for research in primary care: opportunities and challenges," *Family Practice*, vol. 23, no. 2, pp. 253–263, 4 2006.

[92] L. G. Hemkens, D. G. Contopoulos-Ioannidis, and J. P. Ioannidis, "Routinely collected data and comparative effectiveness evidence: promises and limitations," *CMAJ : Canadian Medical Association Journal*, vol. 188, no. 8, p. E158, 5 2016.

[93] "ISD Services — Electronic Data Research and Innovation Service (eDRIS) — Data for Research — ISD Scotland. Last Accessed on 07/08/2021." [Online]. Available: https://www.isdscotland.org/Products-and-Services/EDRIS/Data-for-Research/

[94] "Welcome - Scottish Health and Social Care Open Data. Last Accessed on 07/08/2021." [Online]. Available: https://www.opendata.nhs.scot/

[95] "R: The R Project for Statistical Computing. Last Accessed on 02/09/2021." [Online]. Available: https://www.r-project.org/

[96] R. Janipella, V. Gupta, and R. V. Moharir, "Application of geographic information system in energy utilization," *Current Developments in Biotechnology and Bioengineering: Waste Treatment Processes for Energy Generation*, pp. 143–161, 1 2019.

[97] P. r Edzer, R. Bivand, B. Rowlingson, V. Gomez-Rubio, R. Hijmans, M. Sumner, D. MacQueen, J. Lemon, J. O'Brien, and J. O'Rourke, "Classes and Methods for Spatial Data," *CRAN*.

[98] "Convert UK Postcode to Latitude / Longitude. Last Accessed on 02/09/2021." [Online]. Available: https://www.freemaptools.com/convert-uk-postcode-to-lat-lng.htm

[99] "Organisations – Scotland's Health on the Web. Last Accessed on 02/09/2021." [Online]. Available: https://www.scot.nhs.uk/organisations/

[100] "Intermediate Zone Boundaries 2011 - data.gov.uk. Last Accessed on 02/09/2021." [Online]. Available: https://data.gov.uk/dataset/133d4983-c57d-4ded-bc59-390c962ea280/intermediate-zone-boundaries-2011

[101] D. Lee, A. Rushworth, and G. Napier, "CARBayesST version 2.2: An R Package for Spatio-temporal Areal Unit Modelling with Conditional Autoregressive Priors." [Online]. Available: http://www.hscic.gov.uk/indicatorportal

[102] "ISD Services — Geography, Population and Deprivation Analytical Support Team — Geography — ISD Scotland. Last Accessed on 02/09/2021." [Online]. Available: https://www.isdscotland.org/Products-and-Services/GPD-Support/Geography/

[103] R. S. Bivand, E. Pebesma, and V. Gómez-Rubio, *Applied Spatial Data Analysis with R*.

[104] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev, "Why the Monte Carlo method is so important today," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 386–392, 11 2014.

[105] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers and Geosciences*, vol. 34, no. 9, pp. 1044–1055, 9 2008.

[106] N. Hofstra, M. Haylock, M. New, P. Jones, and C. Frei, "Comparison of six methods for the interpolation of daily, European climate data," *Journal of Geophysical Research Atmospheres*, vol. 113, no. 21, p. 21110, 11 2008.

[107] "Choosing the Right Interpolation Method - GIS Resources. Last Accessed on 04/08/2021." [Online]. Available: https://www.gisresources.com/choosing-the-right-interpolation-method_2/

[108] M. Gentile, F. Courbin, and G. Meylan, "Interpolating point spread function anisotropy," 10 2012.

[109] "Kriging Interpolation Explanation — Columbia Public Health. Last Accessed on 02/09/2021." [Online]. Available: https://www.publichealth.columbia.edu/research/population-health-methods/kriging-interpolation

[110] D. Lee and T. Neocleous, "Bayesian quantile regression for count data with application to environmental epidemiology," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 59, no. 5, pp. 905–920, 11 2010.

[111] "Applied Spatial Statistics for Public Health Data. Last Accessed on 02/09/2021." [Online]. Available: https://web-b-ebscohost-com.proxy.lib.strath.ac.uk/ehost/ebookviewer/ebook/bmxlYmtfXzExNzA3M19fQU41?sid=ed9bd884-736a-4f24-86bd-73e89eee3295%40sessionmgr101&vid=0&format=EB&rid=1

[112] D. Lee, "CARBayes version 5.2.3: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors," Tech. Rep. [Online]. Available: http://www.r-inla.org/

[113] Smyth, *Springer Texts in Statistics Generalized Linear Models With Examples in R*.

[114] "4.3 GLM, GAM and more — Interpretable Machine Learning. Last Accessed on 28/04/2020." [Online]. Available: https://christophm.github.io/interpretable-ml-book/extend-lm.html

[115] J. Maindonald, "Smoothing Terms in GAM Models. Last accessed on 31/01/2022," 2010. [Online]. Available: http://wwwmaths.anu.edu.au/%7Ejohnm/r-book/xtras/lm-compute.pdf.

[116] "Package 'mgcv' Title Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Last accessed on 31/01/2022," 2021.

[117] "Bayes' Theorem (Stanford Encyclopedia of Philosophy). Last Accessed on 04/08/2021." [Online]. Available: https://plato.stanford.edu/entries/bayes-theorem/

[118] M. A. B.-R. MARTINEZ-BENEITO, *DISEASE MAPPING : from foundations to multidimensional modeling*. CRC PRESS, 2021.

[119] V. Roy, "Convergence diagnostics for Markov chain Monte Carlo. Last Accessed on 04/09/2021." 2019. [Online]. Available: www.annualreviews.org

[120] A. Gelman, D. Lee, and J. Guo, "Stan: A probabilistic programming language for Bayesian inference and optimization. Last Accessed on 04/09/2021." 2015. [Online]. Available: http://mc-stan.org/

[121] D. Lee, A. Rushworth, G. Napier, and W. Pettersson, "CARBayesST version 3.2: Spatio-Temporal Areal Unit Modelling in R with Conditional Autoregressive Priors." [Online]. Available: http://seer.cancer.gov

[122] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics 1990 43:1*, vol. 43, no. 1, pp. 1–20, 3 1991.

[123] D. Lee, "A comparison of conditional autoregressive models used in Bayesian disease mapping," *Spatial and Spatio-temporal Epidemiology*, vol. 2, no. 2, pp. 79–89, 6 2011.

[124] H. S. Stern and N. Cressie, "Posterior predictive model checks for disease mapping models," *STATISTICS IN MEDICINE Statist. Med*, vol. 19, pp. 2377–2397, 2000.

[125] B. G. Leroux, X. Lei, and N. Breslow, "Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence," pp. 179–191, 2000.

[126] S. K. Sahu and D. Böhning, "Bayesian spatio-temporal joint disease mapping of Covid-19 cases and deaths in local authorities of England," *Spatial Statistics*, 2021.

[127] L. Knorr-Held, "Knorr-Held: Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk," *Sonder for schungsbereich*, vol. 386, 1999.

[128] G. Napier, D. Lee, C. Robertson, and A. Lawson, "A Bayesian space-time model for clustering areal units based on their disease trends," *Biostatistics*, vol. 20, no. 4, pp. 681–697, 10 2019.

[129] "Local surveillance of antimicrobial use: framework for Antimicrobial Management Teams. Last Accessed on 25/08/2021." *Scottish Antimicrobial Prescribing Group and Health Improvement Scotland*, 2015.

[130] "HPS Website - Scottish One Health Antimicrobial Use and Antimicrobial Resistance in 2016. Last Accessed on 14/08/2021." [Online]. Available: https://www.hps.scot.nhs.uk/web-resources-container/scottish-one-health-antimicrobial-use-and-antimicrobial-resistance-report-2016/

[131] "HPS Website - Scottish One Health Antimicrobial Use and Antimicrobial Resistance in 2018. Last Accessed on 14/08/2021." [Online]. Available: https://www.hps.scot.nhs.uk/web-resources-container/scottish-one-health-antimicrobial-use-and-antimicrobial-resistance-in-2018/

[132] K. B. Pouwels, F. C. K. Dolk, D. R. M. Smith, T. Smieszek, and J. V. Robotham, "Explaining variation in antibiotic prescribing between general practices in the UK," *Journal of Antimicrobial Chemotherapy*, vol. 73, no. suppl_2, pp. ii27–ii35, 2 2018.

[133] B. Goldacre, C. Reynolds, A. Powell-Smith, A. J. Walker, T. A. Yates, R. Croker, and L. Smeeth, "Do doctors in dispensing practices with a financial conflict of interest prescribe more expensive drugs? A cross-sectional analysis of English primary care prescribing data," *BMJ Open*, vol. 9, no. 2, p. e026886, 2 2019.

[134] I. Wales, M. Heginbothom, M. Cronin, R. Howe, and E. Davies, "Antibacterial Usage in Primary Care Status: Final v1," Tech. Rep., 2013.

[135] J. R. Covvey, B. F. Johnson, V. Elliott, W. Malcolm, and A. B. Mullen, "An association between socioeconomic deprivation and primary care antibiotic prescribing in Scotland," *Journal of Antimicrobial Chemotherapy*, vol. 69, no. 3, pp. 835–841, 3 2014.

[136] K. Thomson, R. Berry, T. Robinson, H. Brown, C. Bambra, and A. Todd, "An examination of trends in antibiotic prescribing in primary care and the association with area-level deprivation in England," *BMC Public Health 2020 20:1*, vol. 20, no. 1, pp. 1–9, 8 2020.

[137] E. Y. Klein, E. Schueller, K. K. Tseng, D. J. Morgan, R. Laxminarayan, and A. Nandi, "The Impact of Influenza Vaccination on Antibiotic Use in the United States, 2010–2017," *Open Forum Infectious Diseases*, vol. 7, no. 7, 7 2020.

[138] L. R. Rodgers, A. J. Streeter, N. Lin, W. Hamilton, and W. E. Henley, "Impact of influenza vaccination on amoxicillin prescriptions in older adults: A retrospective cohort study using primary care data," *PLOS ONE*, vol. 16, no. 1, p. e0246156, 1 2021.

[139] J. C. Kwong, S. Maaten, R. E. G. Upshur, D. M. Patrick, and F. Marra, "The Effect of Universal Influenza Immunization on Antibiotic Prescriptions: An Ecological Study," *Clinical Infectious Diseases*, vol. 49, no. 5, pp. 750–756, 9 2009.

[140] "Prescriptions in the Community. Last Accessed on 16/03/2021." *NHS*. [Online]. Available: https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community

[141] "Prescribing Data: BNF Codes — Blog — Oxford DataLab. Last Accessed on 14/08/2021." [Online]. Available: https://www.thedatalab.org/blog/161/prescribing-data-bnf-codes/

[142] "BNF 5.1: Antibacterial drugs — OpenPrescribing. Last Accessed on 14/08/2021." [Online]. Available: https://openprescribing.net/bnf/0501/

[143] "GP Practice Population Demographics. Last Accessed on 16/03/2021." *NHS*. [Online]. Available: https://www.opendata.nhs.scot/dataset/gp-practice-populations

[144] "General Practice - GP Workforce and practice list sizes 2010 - 2020 - General Practice - GP workforce and practice list sizes - Publications - Public Health Scotland. Last Accessed on 09/09/2021." [Online]. Available: https://publichealthscotland.scot/publications/general-practice-gp-workforce-and-practice-list-sizes/general-practice-gp-workforce-and-practice-list-sizes-2010-2020/

[145] "Sankey plot — the R Graph Gallery. Last Accessed on 14/08/2021." [Online]. Available: https://www.r-graph-gallery.com/sankey-diagram.html

[146] "Prescribing Indicators 2016/17 in general practice. Last Accessed on 16/03/2021." *Lothian Prescribing Bulletin*, 2016. [Online]. Available: https://www.ljf.scot.nhs.uk/PrescribingBulletins/PrescribingBulletins/PIssupplement2016-17FINAL.pdf

[147] D. Kahle, H. W. [aut], S. J. [aut], and M. K. [ctb], "Spatial Visualization with ggplot2. Last Accessed on 16/03/2021." *CRAN*. [Online]. Available: https://cran.r-project.org/web/packages/ggmap/ggmap.pdf

[148] "prcomp function - RDocumentation. Last Accessed on 14/08/2021." [Online]. Available: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp

[149] "National Records of Scotland." [Online]. Available: https://www.nrscotland.gov.uk/

[150] A. Mölter, M. Belmonte, V. Palin, C. Mistry, M. Sperrin, A. White, W. Welfare, and T. Van Staa, "Antibiotic prescribing patterns in general medical practices in England: Does area matter?" *Health & Place*, vol. 53, pp. 10–16, 9 2018.

[151] V. Adekanmbi, H. Jones, D. Farewell, and N. A. Francis, "Antibiotic use and deprivation: an analysis of Welsh primary care antibiotic prescribing data by socioeconomic status," *Journal of Antimicrobial Chemotherapy*, vol. 75, no. 8, pp. 2363–2371, 8 2020.

[152] L. G. González-Ortiz and G. Masiero, "Disentangling spillover effects of antibiotic consumption: a spatial panel approach," *Applied Economics*, vol. 45, no. 8, pp. 1041–1054, 3 2011.

[153] M. FILIPPINI, G. MASIERO, and K. MOSCHETTI, "Small area variations and welfare loss in the use of outpatient antibiotics," *Health Economics, Policy and Law*, vol. 4, no. 1, pp. 55–77, 2009.

[154] O. Nitzan, M. Low, I. Lavi, A. Hammerman, S. Klang, and R. Raz, "Variability in outpatient antimicrobial consumption in Israel," *Infection*, vol. 38, no. 1, pp. 12–18, 2 2010.

[155] M. Filippini, F. Heimsch, and G. Masiero, "Antibiotic consumption and the role of dispensing physicians," *Regional Science and Urban Economics*, vol. 49, pp. 242–251, 11 2014.

[156] D. Lee, "A locally adaptive process-convolution model for estimating the health impact of air pollution," *The Annals of Applied Statistics*, vol. 12, no. 4, pp. 2540–2558, 12 2018.

[157] B. Rowlingson, E. Lawson, B. Taylor, and P. J. Diggle, "Mapping English GP prescribing data: a tool for monitoring health-service inequalities," *BMJ Open*, vol. 3, no. 1, p. e001363, 1 2013.

[158] A. Lal, A. Swaminathan, and T. Holani, "Spatial clusters of Clostridium difficile infection and an association with neighbourhood socio-economic disadvantage in the Australian Capital Territory, 2004–2014," *Infection, Disease & Health*, vol. 25, no. 1, pp. 3–10, 2 2020.

[159] L. Furuya-Kanamori, J. Robson, R. J. Soares Magalhães, L. Yakob, S. J. McKenzie, D. L. Paterson, T. V. Riley, and A. C. Clements, "A population-based spatio-temporal analysis of Clostridium difficile infection in Queensland, Australia over a 10-year period," *Journal of Infection*, vol. 69, no. 5, pp. 447–455, 11 2014.

[160] S. van Dorp, M. Hensgens, O. Dekkers, A. Demeulemeester, A. Buiting, P. Bloembergen, S. de Greeff, and E. Kuijper, "Spatial clustering and livestock exposure as risk factor for community-acquired Clostridium difficile infection," *Clinical microbiology and infection*, vol. 25, no. 5, pp. 607–612, 5 2019.

[161] "NHS Scotland performance against LDP standards - gov.scot. Last Accessed on 22/06/2021." [Online]. Available: https://www.gov.scot/publications/nhsscotland-performance-against-ldp-standards/pages/clostridium-difficile-infections/

[162] E. Ofori, D. Ramai, M. Dhawan, F. Mustafa, J. Gasperino, and M. Reddy, "Community-acquired Clostridium difficile: epidemiology, ribotype, risk factors, hospital and intensive care unit outcomes, and current and emerging therapies," pp. 436–442, 8 2018.

[163] L. Furuya-Kanamori, S. J. McKenzie, L. Yakob, J. Clark, D. L. Paterson, T. V. Riley, and A. C. Clements, "Clostridium difficile infection seasonality: Patterns across hemispheres and continents - A systematic review," 3 2015. [Online]. Available: /pmc/articles/PMC4361656//pmc/articles/PMC4361656/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4361656/

[164] P. Polgreen, M. Yang, L. Bohnett, and J. Cavanaugh, "A time-series analysis of clostridium difficile and its seasonal association with influenza," *Infection control and hospital epidemiology*, vol. 31, no. 4, pp. 382–387, 4 2010.

[165] "HPS Website - Protocol for the Scottish Surveillance Programme for Clostridium difficile Infection. User Manual. Last Accessed on 20/08/2021." [Online]. Available: https://www.hps.scot.nhs.uk/web-resources-container/protocol-for-the-scottish-surveillance/programme-for-clostridium-difficile-infection-user-manual/

[166] R. Fry, J. Hollinghurst, H. R. Stagg, D. A. Thompson, C. Fronterre, C. Orton, R. A. Lyons, D. V. Ford, A. Sheikh, and P. J. Diggle, "Real-time spatial health surveillance: Mapping the UK COVID-19 epidemic," 8 2020.

[167] H. Kato, "Development of a Spatio-Temporal Analysis Method to Support the Prevention of COVID-19 Infection: Space-Time Kernel Density Estimation Using GPS Location History Data," pp. 51–67, 2021.

[168] "Clostridium difficile - NHS. Last Accessed on 16/03/2021." *NHS*. [Online]. Available: https://www.nhs.uk/conditions/c-difficile/

[169] P. Bandelj, R. Blagus, F. Briski, O. Frlic, A. Vergles Rataj, M. Rupnik, M. Ocepek, and M. Vengust, "Identification of risk factors influencing Clostridium difficile prevalence in middle-size dairy farms," *Veterinary Research 2016 47:1*, vol. 47, no. 1, pp. 1–11, 3 2016.

[170] N. E. Hopman, E. C. Keessen, C. Harmanus, I. M. Sanders, L. A. van Leengoed, E. J. Kuijper, and L. J. Lipman, "Acquisition of Clostridium difficile by piglets," *Veterinary Microbiology*, vol. 149, no. 1-2, pp. 186–192, 4 2011.

[171] T. Peláez, L. Alcalá, J. Blanco, S. -P. Anaerobe, and u. 2013, "Characterization of swine isolates of Clostridium difficile in Spain: a potential source of epidemic multidrug resistant strains?" *Elsevier*.

[172] "Agricultural Parishes. Last Accessed on 30/06/2021." [Online]. Available: https://spatialdata.gov.scot/geonetwork/srv/api/records/c1d34a5d-28a7-4944-9892-196ca6b3be0c

[173] "Agriculture maps - gov.scot. Last Accessed on 21/08/2021." [Online]. Available: https://www.gov.scot/publications/agriculture-maps/

[174] N. Siu Ngan Lam, "Spatial interpolation methods: A review," *American Cartographer*, vol. 10, no. 2, pp. 129–150, 1 1983.

[175] P. C. Kyriakidis, "A Geostatistical Framework for Area-to-Point Spatial Interpolation," *Geographical Analysis*, vol. 36, no. 3, pp. 259–289, 7 2004.

[176] "Agricultural Parishes - data.gov.uk. Last Accessed on 21/08/2021." [Online]. Available: https://data.gov.uk/dataset/939fdd5e-7322-4ab7-9dc9-bbfc538c4477/agricultural-parishes

[177] C. Rodriguez, H. Hakimi, R. Vanleyssem, B. Taminiau, J. Van Broeck, M. Delmée, N. Korsak, and G. Daube, "Clostridium difficile in beef cattle farms, farmers and their environment: Assessing the spread of the bacterium," *Veterinary microbiology*, vol. 210, pp. 183–187, 10 2017.

[178] C. A. Gotway and L. J. Young, "Combining Incompatible Spatial Data," *Journal of the American Statistical Association*, vol. 97, pp. 632–648, 2002.

[179] E. Hallisey, E. Tai, A. Berens, G. Wilt, L. Peipins, B. Lewis, S. Graham, B. Flanagan, and N. B. Lunsford, "Transforming geographic scale: a comparison of combined population and areal weighting to other interpolation methods," *International Journal of Health Geographics 2017 16:1*, vol. 16, no. 1, pp. 1–16, 8 2017.

[180] "Annual epidemiological commentary: Gram-negative bacteraemia, MRSA bacteraemia, MSSA bacteraemia and C. difficile infections, up to and," Tech. Rep.

[181] "HPS Website - Healthcare Associated Infection Annual Report 2017. Last Accessed on 11/05/2021." [Online]. Available: https://www.hps.scot.nhs.uk/web-resources-container/healthcare-associated-infection-annual-report-2017/

[182] A. Banks, E. K. Moore, J. Bishop, J. E. Coia, D. Brown, H. Mather, and C. Wiuff, "Trends in mortality following Clostridium difficile infection in Scotland, 2010–2016: a retrospective cohort and case–control study," *Journal of Hospital Infection*, vol. 100, no. 2, pp. 133–141, 10 2018.

[183] B. G. Mitchell and A. Gardner, "Mortality and Clostridium difficile infection: a review," *Antimicrobial Resistance and Infection Control*, vol. 1, no. 1, p. 20, 5 2012.

[184] *Together for Health Tackling antimicrobial resistance and improving antibiotic prescribing A Delivery Plan for NHS Wales and its partners. Last Accessed on 16/03/2021*, 2016. [Online]. Available: https://gov.wales/sites/default/files/publications/2019-01/together-for-health-tackling-antimicrobial-resistance-/and-improving-antibiotic-prescribing.pdf

[185] "Clostridium Difficile Annual reports - Public Health Wales. Last Accessed on 29/01/2021." [Online]. Available: https://phw.nhs.wales/services-and-teams/harp/healthcare-associated-infections-hcai/clostridium-difficile-accordian/clostridium-difficile-annual-reports/

[186] "General Medical Services Contract: Quality and Outcomes Framework Statistics for Wales, 2017-18," Tech. Rep.

[187] N. Salkind, "Last Observation Carried Forward," in *Encyclopedia of Research Design*. SAGE Publications, Inc., 10 2012.

[188] D. Xiang, "Fitting Generalized Additive Models with the GAM Procedure," Tech. Rep.

[189] Public Health Data Science, "Public Health England: Technical Guide Confidence Intervals. Last Accessed on 16/03/2021," 2018. [Online]. Available: file:///C:/Users/rjb13173/Downloads/PHDSGuidance-ConfidenceIntervals.pdf

[190] "Diabetic foot infections, antibacterial therapy — Treatment summary — BNF content published by NICE. Last Accessed on 06/05/2021." [Online]. Available: https://bnf.nice.org.uk/treatment-summary/diabetic-foot-infections-antibacterial-therapy.html

[191] M. E and C. R, "Health care use and serious infection prevalence associated with penicillin "allergy" in hospitalized patients: A cohort study," *The Journal of allergy and clinical immunology*, vol. 133, no. 3, pp. 790–796, 3 2014.

[192] R. M. West, C. J. Smith, S. H. Pavitt, C. C. Butler, P. Howard, C. Bates, S. Savic, J. M. Wright, J. Hewison, and J. A. T. Sandoe, "'Warning: allergic to penicillin': association between penicillin allergy status in 2.3 million NHS general practice electronic health records, antibiotic prescribing and health outcomes," *Journal of Antimicrobial Chemotherapy*, vol. 74, no. 7, pp. 2075–2082, 7 2019.

[193] K. G. Blumenthal, N. Lu, Y. Zhang, Y. Li, R. P. Walensky, and H. K. Choi, "Risk of meticillin resistant Staphylococcus aureus and Clostridium difficile in patients with a documented penicillin allergy: population based matched cohort study," *BMJ (Clinical research ed.)*, vol. 361, p. k2400, 6 2018.

[194] C. Thomas and R. Boldero, "Antimicrobial Stewardship Forum: National Prescribing Indicators," Tech. Rep.

[195] R. Dantes, Y. Mu, L. A. Hicks, J. Cohen, W. Bamberg, Z. G. Beldavs, G. Dumyati, M. M. Farley, S. Holzbauer, J. Meek, E. Phipps, L. Wilson, L. G. Winston, L. C. McDonald, and F. C. Lessa, "Association Between Outpatient Antibiotic Prescribing Practices and Community-Associated Clostridium difficile Infection," *Open Forum Infectious Diseases*, vol. 2, no. 3, 9 2015.

[196] K. A. Brown, N. Khanafer, N. Daneman, and D. N. Fisman, "Meta-analysis of antibiotics and the risk of community-associated Clostridium difficile infection," *Antimicrobial Agents and Chemotherapy*, vol. 57, no. 5, pp. 2326–2332, 5 2013.

[197] "BREATHE - Health Data Research Hub for Respiratory Health — The University of Edinburgh. Last Accessed on 06/09/2021." [Online]. Available: https://www.ed.ac.uk/usher/breathe

[198] M. Woodward, S. A. Peters, and K. Harris, "Social deprivation as a risk factor for COVID-19 mortality among women and men in the UK Biobank: Nature of risk and context suggests that social interventions are essential to mitigate the effects of future pandemics," *Journal of Epidemiology and Community Health*, vol. 0, pp. 1–6, 4 2021.

[199] "COVID Symptom Study - Help slow the spread of COVID-19. Last Accessed on 09/06/2021." [Online]. Available: https://covid.joinzoe.com/

[200] D. A. Drew, L. H. Nguyen, C. J. Steves, C. Menni, M. Freydin, T. Varsavsky, C. H. Sudre, M. Jorge Cardoso, S. Ourselin, J. Wolf, T. D. Spector, and A. T. Chan, "Rapid implementation of mobile technology for real-time epidemiology of COVID-19," *Science*, vol. 368, no. 6497, pp. 1362–1367, 6 2020.

[201] P. Tammes, "Social distancing, population density, and spread of COVID-19 in England: A longitudinal study," *BJGP Open*, vol. 4, no. 3, 8 2020.

[202] "Scottish Index of Multiple Deprivation 2020 - gov.scot. Last Accessed on 26/08/2021." [Online]. Available: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/

[203] "Coronavirus. Last Accessed on 05/09/2021." [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1

[204] "Coronavirus: New case of Covid-19 confirmed on Skye - BBC News. Last Accessed on 16/03/2021." [Online]. Available: https://www.bbc.co.uk/news/uk-scotland-highlands-islands-52769088

[205] D. Laroze, E. Neumayer, and T. Plümper, "COVID-19 does not stop at open borders: Spatial contagion among local authority districts during England's first wave," *Social Science and Medicine*, vol. 270, p. 113655, 2 2021.

[206] K. I, H. Budhwani, and B. Podbielski, "Evaluating Population Density as a Parameter for Optimizing COVID-19 Testing: Statistical Analysis," *JMIRx Med 2021;2(1):e22195 https://xmed.jmir.org/2021/1/e22195*, vol. 2, no. 1, p. e22195, 2 2021.

[207] R. Elson, T. M. Davies, I. R. Lake, R. Vivancos, P. B. Blomquist, A. Charlett, and G. Dabrera, "The spatio-temporal distribution of COVID-19 infection in England between January and June 2020," *Epidemiology & Infection*, vol. 149, p. e73, 3 2021.

[208] M. J. Keeling, M. J. Tildesley, B. D. Atkins, B. Penman, E. Southall, G. Guyver-Fletcher, A. Holmes, H. McKimm, E. E. Gorsich, E. M. Hill, and L. Dyson, "The impact of school reopening on the spread of COVID-19 in England," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1829, 7 2021.

[209] "SAGE Return to School. Last Accessed on 28/08/2021," 11 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/935125/tfc-covid-19-children-transmission-s0860-041120.pdf

[210] Z. Sun, B. Yang, R. Zhang, and X. Cheng, "Influencing Factors of Understanding COVID-19 Risks and Coping Behaviors among the Elderly Population," *International Journal of Environmental Research and Public Health 2020, Vol. 17, Page 5889*, vol. 17, no. 16, p. 5889, 8 2020.

[211] D. Lee, C. Robertson, and D. Marques, "Quantifying the small-area spatio-temporal dynamics of the Covid-19 pandemic in Scotland during a period with limited testing capacity," *Spatial Statistics*, p. 100508, 4 2021.

[212] F. Service, "Syndromic Surveillance Summary. Last Accessed on 10/03/2021," Tech. Rep., 2021. [Online]. Available: https://www.gov.uk/government/collections/syndromic-

[213] A. McAteer, P. C. Hannaford, D. Heaney, L. D. Ritchie, and A. M. Elliott, "Investigating the public's use of Scotland's primary care telephone advice service (NHS 24): A population-based cross-sectional study," *British Journal of General Practice*, vol. 66, no. 646, pp. e337–e346, 5 2016.

[214] T. Varsavsky, M. S. Graham, L. S. Canas, S. Ganesh, J. Capdevila Pujol, C. H. Sudre, B. Murray, M. Modat, M. Jorge Cardoso, C. M. Astley, D. A. Drew, L. H. Nguyen, T. Fall, M. F. Gomez, P. W. Franks, A. T. Chan, R. Davies, J. Wolf, C. J. Steves, T. D. Spector, and S. Ourselin, "Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study," *The Lancet Public Health*, vol. 6, no. 1, pp. e21–e29, 1 2021.

[215] "Covid in numbers: The story of Scotland's pandemic - BBC News. Last Accessed on 28/08/2021." [Online]. Available: https://www.bbc.co.uk/news/uk-scotland-56399043

[216] D. Nathwani, J. Sneddon, A. Patton, and W. Malcolm, "Antimicrobial stewardship in Scotland: impact of a national programme," *Antimicrobial Resistance and Infection Control 2012 1:1*, vol. 1, no. 1, pp. 1–3, 2 2012.

[217] C. A. Marwick, N. Yu, M. C. Lockhart, C. C. McGuigan, C. Wiuff, P. G. Davey, and P. T. Donnan, "Community-associated Clostridium difficile infection among older people in Tayside, Scotland, is associated with antibiotic exposure and care home residence: cohort study with nested case–control," *Journal of Antimicrobial Chemotherapy*, vol. 68, no. 12, pp. 2927–2933, 12 2013.

[218] L. Furuya-Kanamori, S. J. McKenzie, L. Yakob, J. Clark, D. L. Paterson, T. V. Riley, and A. C. Clements, "Clostridium difficile Infection Seasonality: Patterns across Hemispheres and Continents – A Systematic Review," *PLOS ONE*, vol. 10, no. 3, p. e0120730, 3 2015.

[219] T. Lawes, J. M. Lopez-Lozano, C. A. Nebot, G. Macartney, R. Subbarao-Sharma, K. D. Wares, C. Sinclair, and I. M. Gould, "Effect of a national 4C antibiotic stewardship intervention on the clinical and molecular epidemiology of Clostridium difficile infections in a region of Scotland: a non-linear time-series analysis," *The Lancet Infectious Diseases*, vol. 17, no. 2, pp. 194–206, 2 2017.

[220] D. Lee, "A spatio-temporal process-convolution model for quantifying health inequalities in respiratory prescription rates in Scotland," 4 2017.

[221] K. A. Levin, "Study Design VI-Ecological Studies," *Evidence-Based Dentistry*, vol. 7, pp. 60–61, 2003.

[222] J. M. Read, C. A. Green, E. M. Harrison, A. B. Docherty, S. Funk, J. Harrison, M. Girvan, H. E. Hardwick, L. Turtle, J. Dunning, J. S. Nguyen-Van-Tam, P. J. Openshaw, J. K. Baillie, and M. G. Semple, "Hospital-acquired SARS-CoV-2 infection in the UK's first COVID-19 pandemic wave," *The Lancet*, vol. 0, no. 0, 8 2021.

[223] M. B. Gillies, D. P. Burgner, L. Ivancic, N. Nassar, J. E. Miller, S. G. Sullivan, I. M. F. Todd, S.-A. Pearson, A. L. Schaffer, and H. Zoega, "Changes in antibiotic prescribing following COVID-19 restrictions: Lessons for post-pandemic antibiotic stewardship," *British Journal of Clinical Pharmacology*, 2021.

[224] R. Roshan, A. S. Feroz, Z. Rafique, and N. Virani, "Rigorous Hand Hygiene Practices Among Health Care Workers Reduce Hospital-Associated Infections During the COVID-19 Pandemic:," *https://doi.org/10.1177/2150132720943331*, vol. 11, 7 2020.

[225] E. Bentivegna, G. Alessio, V. Spuntarelli, M. Luciani, I. Santino, M. Simmaco, and P. Martelletti, "Impact of COVID-19 prevention measures on risk of health care-associated Clostridium difficile infection," *American Journal of Infection Control*, vol. 49, no. 5, pp. 640–642, 5 2021.

[226] P. Spigaglia, "COVID-19 and Clostridioides difficile infection (CDI): Possible implications for elderly patients," *Anaerobe*, vol. 64, p. 102233, 8 2020.

[227] V. Baccolini, G. Migliara, C. Isonne, B. Dorelli, L. C. Barone, D. Giannini, D. Marotta, M. Marte, E. Mazzalai, F. Alessandri, F. Pugliese, G. Ceccarelli, C. De Vito, C. Marzuillo, M. De Giusti, and P. Villari, "The impact of the COVID-19 pandemic on healthcare-associated infections in intensive care unit patients: a retrospective cohort study," *Antimicrobial Resistance & Infection Control 2021 10:1*, vol. 10, no. 1, pp. 1–9, 6 2021.