A spatial and temporal analysis of Shiga-toxin producing *Escherichia coli* O157 and severe acute respiratory syndrome coronavirus 2 infection in England.

Richard Elson MSc

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy to the University of East Anglia, School of Environmental Sciences.

.

December 2022

Abstract

Incorporating the domains of space and time to the analysis of infectious diseases can reveal unseen structure that may elucidate the mechanisms leading to infection. Spatial statistical methods have been available for many years, but they are not used routinely for surveillance purposes or for risk assessment during outbreaks. The primary aims of this thesis were to identify high or low risk areas of STEC O157 and SARS-CoV-2 in England; examine the spatial relationship between STEC O157 case density and environmental and socio-demographic risk factors and investigate the relationship between individual exposure to risk factors and residence in areas considered high risk for STEC O157. This was achieved using non-parametric smoothing techniques and multivariable negative binomial and logistic regression models

We identified areas of England where the risk of STEC and SARS-CoV-2 infection was significantly increased accounting for the underlying population at risk. For SARS-CoV-2, we describe the highly dynamic spatio-temporal risk at the start of the pandemic and show that widespread transmission was underway prior to lockdown in March 2020. For STEC O157, the risk of infection was greatest in the North, North West and South West of England.

Compared to baseline, STEC O157 risk was associated with cattle (Incidence rate ratio (IRR) 2.2, $p<0.001$) and sheep (IRR 1.7, $p<0.001$) density, rural residence (IRR 1.6, $p<0.001$) and presence of private water supplies (IRR 1.4 $p=0.02$) and we identified a novel association between sheep density and STEC O157 PT21/28 (IRR 2.8, $p<0.001$). Socio-economic status appeared to modify the risk related to travel outside the UK. Direct contact with the environment (Population attributable risk (PAR) 14%) and contact with dogs (PAR 12%) were important risk factors for residents of high-risk areas. Indirect contact with the environment (PAR 44%) and daytrips (PAR 37%) were more important for travellers. Residents of high-risk areas were less likely to report travel (adjusted Odds Ratio 0.56, $p<0.001$) suggesting that they acquired their infection close to home.

These results highlight the importance of considering spatial location and mobility when considering risks of infection. Identifying geographical areas that present an increased risk of infection allows public health messages to be targeted at residents and visitors.

List of contents

List of tables

List of figures

List of supplementary materials

Acknowledgements

I'd like to thank the following people who all played a part in producing this thesis.

- Professor Iain Lake for his unwavering support and supervision throughout.
- Dr. Tilman Davies for writing the R package sparr and for supporting me through my first coding steps.
- Dr Roberto Vivancos for financial support and supervision and Professor Jeremy Hawker for his support and encouragement during the initial stages of my application.
- Professor Sarah O'Brien for support and encouragement since 2000 (and spotting double stops) and Professor Goutam Adak (aka Bob) for his support.
- Dr. Claire Jenkins for expert advice on STEC.
- Ros Lewis for reminding me to complete my timesheets and Nalini Purohit, Rhada Patel, Kirsten Glen and Mike Harte for data entry and administration.
- Iain Gillespie, Lisa Byrne, Naomi Launders, Bhavita Vishram, Saira Butt and Sanjay Chaudhary for creating and managing NESSS since its inception in 2009.
- All local authority staff who provide surveillance data and all laboratory staff who process, type and sequence samples.
- My Mum, Irene and sister Claire for always showing interest in my PhD and supporting me throughout.
- Lucia and Lilly for their never-ending support and patience throughout. Sorry it took so long, and I will repay my childcare debt Luce!
- My Dad, David Elson whose death from the effects of Idiopathic Pulmonary Fibrosis in 2013 motivated me to undertake this work.

1 Introduction

1.1 Brief overview of Shiga-toxin producing *Escherichia coli*

Shiga-toxin producing *Escherichia coli* (STEC) are a group of bacteria associated with human disease and are defined by the presence of one or both phage encoded Shiga toxin (Stx) genes; Stx1 and Stx2 (1). These toxins are amongst the most potent bacterial toxins known (2). First recognised as a human pathogen in 1982 (3), STEC are now globally distributed (4). The infectious dose is low and the mean incubation period ranges from 3.5 to 8.1 days (5).

1.2 Nomenclature

Throughout this thesis, we use the term STEC or STEC O157 to refer to STEC O157:H7, STEC may also be referred to in the literature as verotoxigenic E. coli (VTEC), because of the cell line (Vero) on which their cytotoxicity was first demonstrated, or enterohemorrhagic *E. coli* (EHEC), because they often cause bloody diarrhoea.

1.3 Brief overview of severe acute respiratory syndrome coronavirus 2

Coronaviruses (CoV) are found globally in humans and many different animal species. They are classified in the Orthocoronaviridae subfamily (order: Nidovirales, subordination: Cornidovirineae, family: Coronaviridae).2 CoV can be grouped into 4genera, including α-/β-/γ-/δ-CoV and α- and β-CoV can infect mammals, while γ- and δ-CoV primarily infect birds (6).

CoV have a high mutation rate and homologous recombinations often occur (7). These properties have contributed to a great diversity of CoV in nature, which enables these viruses to infect numerous species. The SARS pandemic, in 2002–2003, led to an increasing number of studies in wild animals on all continents and the greatest diversity of CoV is seen in bats (8). The most recent introductions to humans are thought to be bat viruses, spread via an intermediate animal (eg, the Himalayan palm civet for SARS-CoV

and the dromedary camel for the Middle East respiratory syndrome [MERS]-CoV). However, data on the prevalence of zoonotic CoV in wild animal populations are patchy, particularly in economically and/or politically unstable regions of the world (8).

In late December 2019, several cases of pneumonia of unknown origin were reported from China, which in early January 2020 were announced to be caused by a novel coronavirus. The virus was later denominated severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and defined as the causal agent of Coronavirus Disease 2019 (COVID-19). Despite attempts to contain the disease in China, the virus spread globally, and COVID-19 was declared a pandemic by the World Health Organization (WHO) in March 2020.

SARS-CoV-2 is efficiently transmitted from person-to-person and spread rapidly across all continents in our globalized world. In the resulting COVID-19 pandemic, 766,895,075 people have been infected and 6,935,889 patients have died so far (as of 24th May 2023, source: WHO Coronavirus (COVID-19) Dashboard).

1.4 Surveillance of infectious diseases and infectious intestinal diseases (IID) in England

1.4.1 The Health Protection (Notification) Regulations 2010

The Health Protection (Notification) Regulations 2010 (9) provide the legal basis for the notification of certain infectious diseases and the organisms that cause them. The diseases and causative agents are listed in Schedule 1 and Schedule 2 respectively (9). The main purpose of notification is to enable cases or outbreaks of infection to be investigated and also to provide information for surveillance purposes.

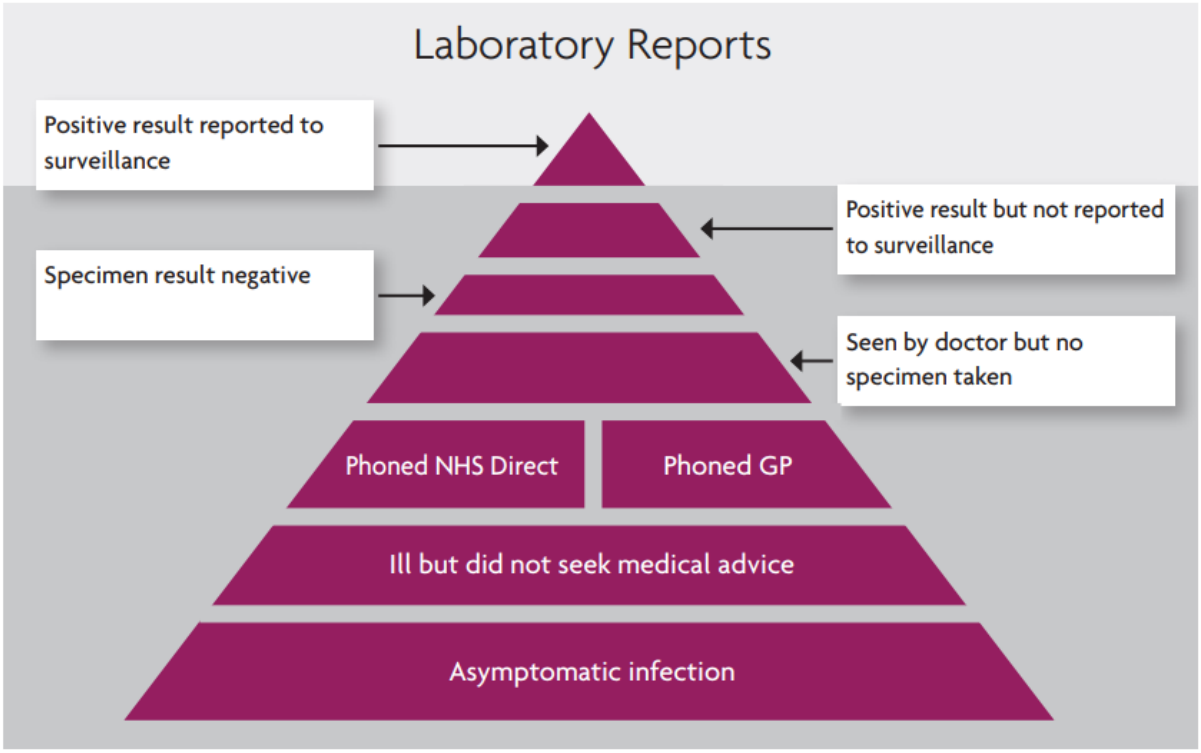
1.4.1.1 Duty to make notifications

Where a registered medical practitioner suspects a patient or dead person has a notifiable disease listed in Schedule 1, they are legally required to notify the proper officer of the local authority where the person usually resides or, in the case of travellers, the port health authority where the patient disembarked (9). Where diagnostic laboratories identify a Schedule 2 causative agent in a human sample, they must report this to UKHSA as a laboratory confirmed case. COVID-19, food poisoning, Haemolytic Uraemic Syndrome (HUS) and infectious bloody diarrhoea are all notifiable diseases (9). SARS-CoV-2 and Verocytotoxigenic Escherichia coli (including *E. coli* O157) are notifiable causative agents (9).

1.4.2 The surveillance pyramid for IID

Despite being very common in the community, not all cases of infectious intestinal disease (IID) present to a healthcare professional, and not all cases that present are reported to national surveillance systems. For a laboratory confirmed case to be included in national surveillance, the case must feel unwell enough to seek medical help, the clinician must request a sample which should then be submitted to a laboratory. The laboratory must examine the sample using appropriate methods and, if a pathogen is identified, report the result to a national surveillance system. Reports of laboratory confirmed IID pathogens therefore represent only a fraction of the true incidence in the community, and because it relies upon cases being unwell enough to seek medical attention, likely reflects the severe end of the disease spectrum. Surveillance data in the UK therefore underestimate the total burden of IID and this pattern of ascertainment is commonly described schematically as the surveillance pyramid presented in Figure 1.1 taken from the second study of infectious intestinal disease in the community (10).

Figure 1.1 The surveillance pyramid



1.5 Surveillance of STEC in England

1.5.1 Microbiological detection, confirmation and typing

In England, isolates of STEC identified by local pathology laboratories are sent for confirmation and typing at the UK Health Security Agency (UKHSA) Gastrointestinal Bacterial Reference Unit (GBRU). Detection and confirmation of STEC includes biochemical identification and serotyping of bacterial isolates. Since 1989, STEC strains have been further differentiated using a phage typing (PT) scheme developed in Canada (11). Retrospective real-time polymerase chain reaction (PCR) targeting Stx 1 or Stx 2 and the intimin (*eae*) gene, associated with intimate attachment of the bacteria to the host gut mucosa, was introduced in 2012 (11). Since 2015, all isolates have been routinely sequenced allowing identification of genetic lineage/sub-lineage and *stx* subtypes (12, 13).

1.5.2 The National Enhanced Surveillance System for STEC (NESSS)

The National Enhanced Surveillance System for STEC (NESSS) was introduced in England in 2009. The system collects clinical and epidemiological information for each laboratory confirmed case using a standardised questionnaire, a copy of which is provided in the supplementary material (SM 1). This includes details about whether they had travelled abroad or within the UK prior to their illness onset and the residential postcode of each case (an alphanumeric reference developed by the UK Post Office to facilitate the delivery of mail, each containing around 15 addresses). This information is linked to reference microbiology information including PT, presence of virulence factors and whole genome sequence data [1].

For the purposes of surveillance, cases are defined as follows:

- **Primary case**: The individual who introduced the disease into a group or the population (not necessarily the index case or the first case diagnosed).
- **Co-primary case**: Case whose date of onset is within one incubation period (4 days) of the primary case, that is a case thought to have been exposed to the same risk factor(s) as the primary case.
- **Secondary case**: Case whose date of onset is more than one incubation period (4 days) after the primary case or whose risk factor is believed to be "exposure to a primary case".
- **Travel-related case**: Case whose date of onset is within one exposure period (7 days) of having been outside of the UK.
- **Asymptomatic case**: a person identified through contact screening procedures, who has not had any symptoms consistent with VTEC infection within one exposure period (7 days) of the symptomatic contact. They are still a case (as they

are shedding bacteria). It is expected that an asymptomatic case does not have an onset date at all.

Cases also present themselves in the following ways:

- **Outbreak related case**: Generally speaking, outbreaks occur when observed case numbers exceed those expected and are epidemiologically and microbiologically linked, usually in time and space. Cases are carefully defined during outbreak investigations based on microbiological and epidemiological similarities.
- **Clustered case**: Cases that are microbiologically similar but where no common epidemiological link can be established. These cases may also cluster in space and/or time and are usually within expected numbers.
- **Sporadic case**: where the case is not microbiologically or epidemiologically linked to another known case. These cases may acquire their infection outside the UK or may be single cases of infection acquired in the UK.

Most cases in England are considered sporadic (14), however, the introduction of whole genome sequencing (WGS) from 2015 greatly improved our ability to identify cases that are phylogenetically linked, thus reducing the number of cases considered to be sporadic. Isolates that fall within a 5- Single Nucleotide Polymorphism (SNP) cluster are considered to have been exposed to the same source of infection but because these clusters tend to be small, establishing the route by which people were exposed to the source remains a challenge (13, 15, 16). This 5-SNP threshold is based upon a pairwise SNP distance distribution of isolate pairs with a known epidemiological link, which showed that no pair had >5 SNP differences with a mean of 1 SNP in isolates from the same household or known common source of infection (13).

1.6 Overview of STEC

1.6.1 Microbiological characteristics and typing methods

STEC has multiple genetic and phenotypic features that contribute to its pathogenicity or are used for detection and identification. The primary virulence factor defining the STEC group is production of Stx1, Stx2, or both. The genes encoding the toxins, Stx1 and Stx2 are the targets of commercial and in-house diagnostic PCR assays. Both toxins can be divided into several subtypes, *stx*1a–1d and *stx*2a–2g (17).

The inability to ferment sorbitol along with other factors differentiates STEC O157 from ≈90% of other gastrointestinal bacteria (17), facilitating the detection and identification of STEC O157 on selective media (17).

For many years, further characterisation was provided using a phage typing (PT) scheme (18) developed in Canada (19). Bacteriophages are viruses that infect bacteria and cause bacterial lysis and cell death, but also play a role in bacterial genome evolution (18). In simple terms, phage typing involves culturing the bacterial isolate and then inoculating this with different phages. If lysis occurs with Phage A, for example, the isolate would be classified (or typed) as Phage Type (PT) A.

More recently, the introduction of multi-locus variable number tandem repeat (VNTR) analysis (MLVA) and whole genome sequencing (WGS) for routine typing of STEC has revealed previously undetected phylogenetic and evolutionary relationships (12). WGS is highly discriminatory and demonstrates unparalleled sensitivity and accuracy in identifying linked cases coupled with phylogenetic clustering of how strains are related over time and space. Its ability to accurately define sporadic cases over time enables better characterisation of the population at risk and assessment of the relative importance of exposures leading to sporadic infections and those that are genetically or epidemiologically linked (13).

1.6.2 Clinical features and severity of STEC infection

Symptoms range from asymptomatic infection through to serious systemic disease and, rarely, death. Clinical illness is characterised by diarrhoea which can range from mild and self-limiting to more severe bloody diarrhoea. Unlike other bacterial GI infections, fever is uncommon and if present, tends to be mild. This lack of fever combined with bloody diarrhoea is used to indicate STEC infection, particularly in children.

During infection, STEC release Shiga- toxins (Stx), the primary virulence factor responsible for the most serious clinical outcomes. The Stx target cells expressing the glycolipid lobotriaosylceramide (Gb3), disrupting protein synthesis and resulting in apoptotic cell death. Renal epithelial cell membranes are enriched for Gb3 meaning that kidneys bear the brunt of Stx toxicity but damage to cardiovascular and neurological systems can also occur (20).

Between 6% and 14% of STEC cases go on to develop haemolytic uraemic syndrome (HUS) (20). HUS is defined as microangiopathic haemolytic anaemia, thrombocytopaenia and acute kidney injury (20) and usually occurs 5–13 days after initial diarrhoeal symptoms (20). Children under the age of 5 are at greatest risk of HUS, (20, 21) and a study of paediatric HUS cases in the UK and Ireland found that STEC infection was the cause of 80% HUS cases (21).

Strains of STEC O157 encoding the stx2-only toxin, specifically the *stx* subtype *stx*2a, are significantly associated with progression to HUS (1, 11, 20). Antibiotic usage is generally contraindicated for use in cases of STEC infection, due to the possibility that bacterial DNA damage may upregulate the production of Stx, particularly the *stx*2 subtype (20), therefore increasing the risk of HUS. Observational epidemiological studies, and analysis

of outbreak data have shown that clinical presentation and the use of antibiotics are risk factors for HUS (20).

1.6.2.1 Burden of STEC

Compared to other bacterial pathogens, STEC is a relatively rare infection in many parts of the world but is of public health concern due its low infectious dose (<100 bacteria) (1) and potential to cause severe disease (1, 22-25). Worldwide, it is estimated that there are around 2.8 million cases annually, leading to 3,890 cases of HUS and 230 deaths (26). The Europe wide rate of infection is estimated to be 2.2 cases per 100,000 population but reported rates vary between countries (Range: <0.1 to 16.3 cases per 100,000 population) (27). This may be due to genuine differences in prevalence but may also be a result of the differences in the methods of detection and surveillance used across the European Union.

The O157 STEC serogroup is most commonly associated with human disease in the UK however, other serogroups are seen more frequently in other European countries (27). Rates of infection in England have remained fairly constant for many years (around (11). Europe showed a similar pattern until an increase since 2011 attributed to wider use of molecular methods following a large outbreak linked to sprouted fenugreek seeds (28).

There are around 800-1,000 cases of STEC reported each year in England but underreporting means that the real number of cases in the community is under-ascertained by a factor of 8:1 (29). Rates of infection are highest in children (1, 27) and most cases occur in the late summer, at least in temperate areas, and this pattern is seen universally (30).

Roberts et al. (31) assessed the economic costs of a serious STEC outbreak that occurred in Central Scotland in 1994 (32). The average cost per HUS case was £62,353, a case of thrombotic thrombocytopenic purpura (TTP) case cost £21,422, non-HUS and non-TTP

cases cost £1,030. The costs of investigating and controlling the Central Scotland outbreak were £171,848 and the costs of cases projected over 30 years were £11.9 million, or £168,032 per case (31). More recent figures show that the total societal burden of foodborne STEC O157 infection in the UK to be £3.9 million each year (33).

1.6.3 Geographical and temporal distribution

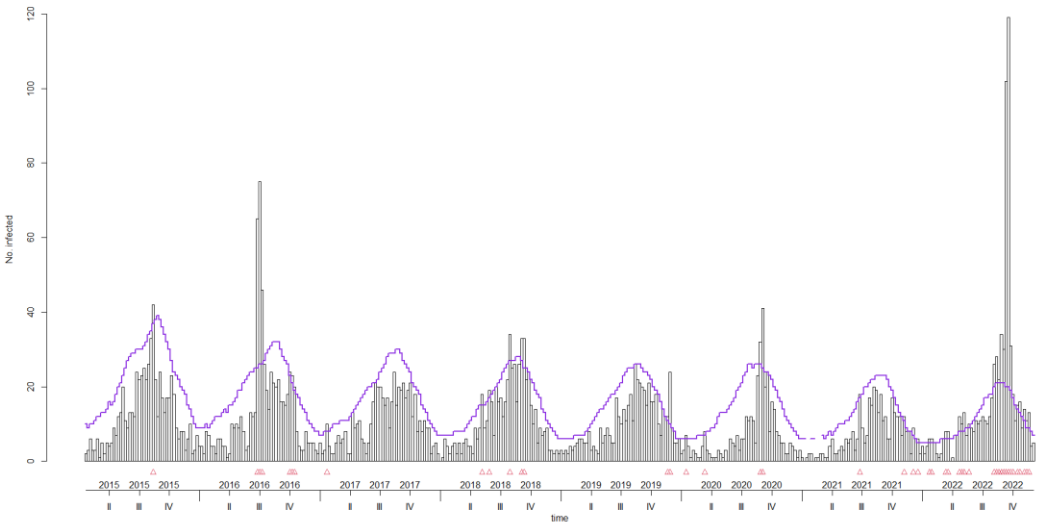
There is evidence that a common ancestor of STEC was introduced to countries around the world on a number of occasions in the past, likely due to international transport of animals and/or contaminated animal feed (34). Following introduction, localised genetic variation has occurred leading to a patchwork of strains that are related at the global level but show distinct geographical differences. Within the United Kingdom, rates of STEC infection in Scotland are more than twice that of England (35) . Within England, rates of infection vary considerably from 0.40 to 1.34 cases per 100,000 person years in London and the North respectively (11, 14) and there is evidence that this relates to living in areas with high densities of farmed animals (36).

However, the strains infecting humans are not always the same as those circulating in the 'local' ruminant reservoir (37, 38). The reasons for this are unclear but may be due to widespread exposure to a remote source of infection, or localised exposure to a source where the availability of comparative microbiological information is scant (14). Conversely, evidence from outbreak investigations shows that transmission of highly related strains can occur via multiple routes from geographically restricted sources (15, 39).

Figure 1.2 shows the weekly numbers of STEC fitted with an exceedance algorithm developed by Farrington et al (40) and improved by Noufaily et al (41). The algorithm uses the preceding six years of data (2009-2015) to provide an indication of expected numbers,

superimposed by the purple line in the figure. Observed numbers that exceed the expected represent large national outbreak events responsible for much of the inter year variation, but smaller outbreaks and phylogenetic clusters occur within endemic levels throughout the year.

Figure 1.2 Observed and expected (purple line) weekly counts of STEC in England 2015-2022

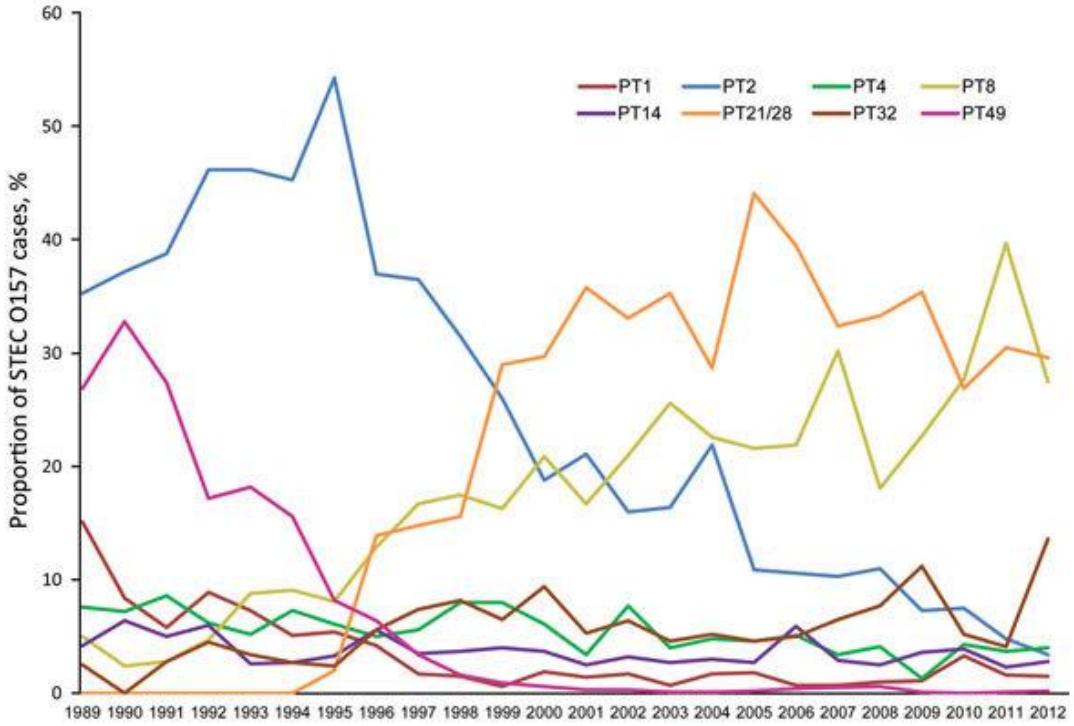


The distinct seasonal pattern is seen from year to year; low levels during the winter months are followed by an increase in infections during March and April, and a late summer peak. This pattern is pronounced in rural areas (14) and is seen globally (30), likely the result of multiple factors including farming practices, animal movements, human behaviour (including travel) and climatic factors.

There have been considerable changes in the strains circulating in England since surveillance began. Adams et al (11) described the changes in phage types over time presented in Figure 1.3. This shows that the predominant strains circulating until the late 1990s were PT2 and PT1 until they were replaced by PT 8 and 21/28. This was also seen

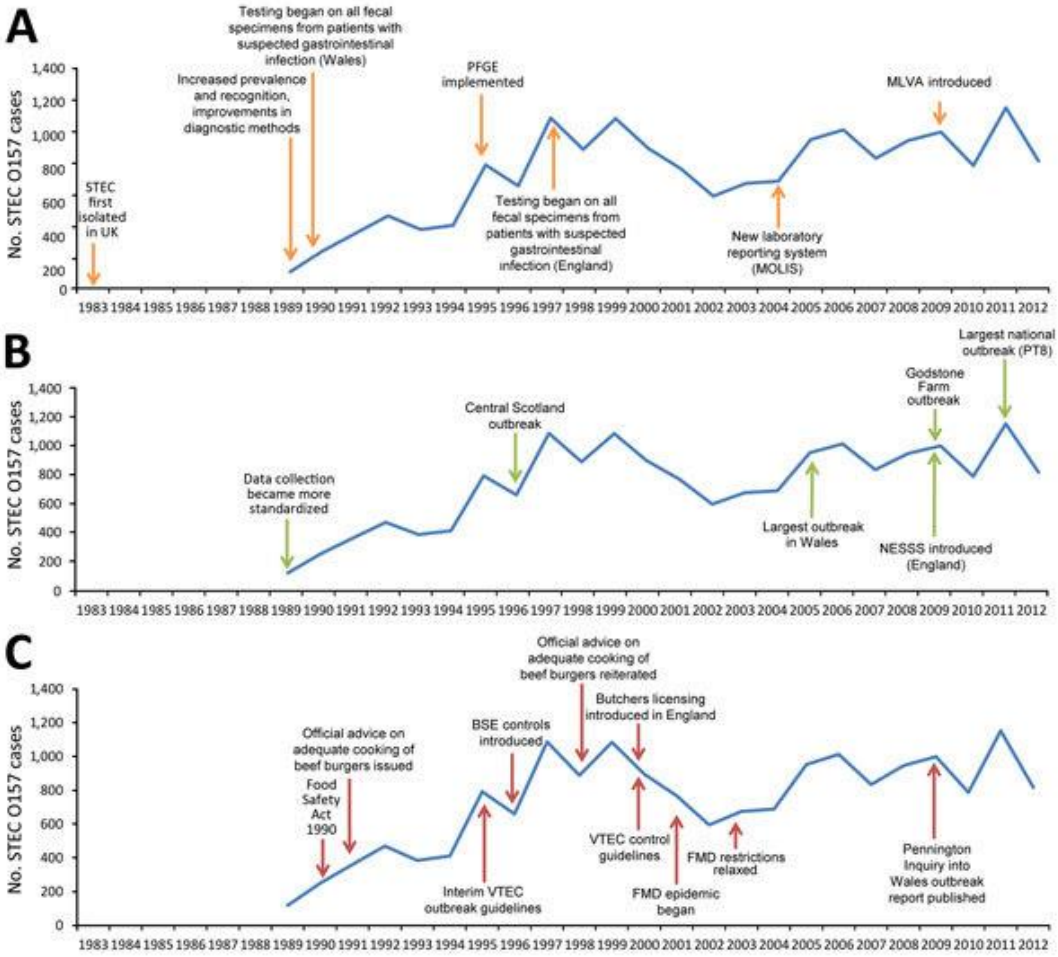
in Scotland (42), suggesting a strain replacement event. The reasons for this change are unclear but the destruction and restocking of UK cattle herds after concerns about bovine spongiform encephalopathy and foot and mouth disease during the study period may be contributing factors (11).

Figure 1.3 Proportions of common phage types (PTs) of Shiga toxin–producing *Escherichia coli* O157 identified, England and Wales, 1989–2012



More recently, phylogenetic analysis of strains circulating in humans and UK cattle during 2014 (12) described three distinct lineages (I, II and I/II) descended from a common ancestor. Lineage I contains PT 21/28 and PT32; strains encoding Stx2 only and associated with more severe disease. Lineage II contains PT8 and Lineage I/II PT2. Isolates from humans and UK cattle are closely related suggesting that PT8 and, in particular, PT21/28,

Figure 1.4 Timeline of key events influencing the epidemiology (A), microbiology (B), and guidance and control (C) of STEC O157, England and Wales, 1983–2012. Numbers before 1989 are available only as an aggregate for that period and therefore cannot be presented by year. BSE, bovine spongiform encephalopathy; FMD, foot and mouth disease; MLVA, multilocus variable-number tandem-repeat analysis; MOLIS, Modular Open Laboratory Information System; NESSS, National Enhanced Surveillance Scheme for STEC; PFGE, pulsed-field gel electrophoresis; PT, phage type; STEC, Shiga toxin–producing Escherichia coli; VTEC, verocytotoxin–producing E. coli.



have a domestic source and are domestically acquired (12). With the advent of routine WGS, it is now possible to identify links between cases that previously appeared sporadic in nature. These cases may exhibit spatial clustering, sometimes over long periods of time, suggesting geographically restricted transmission of highly related strains (15) and extended persistence in the zoonotic reservoir.

1.7 Modes of transmission

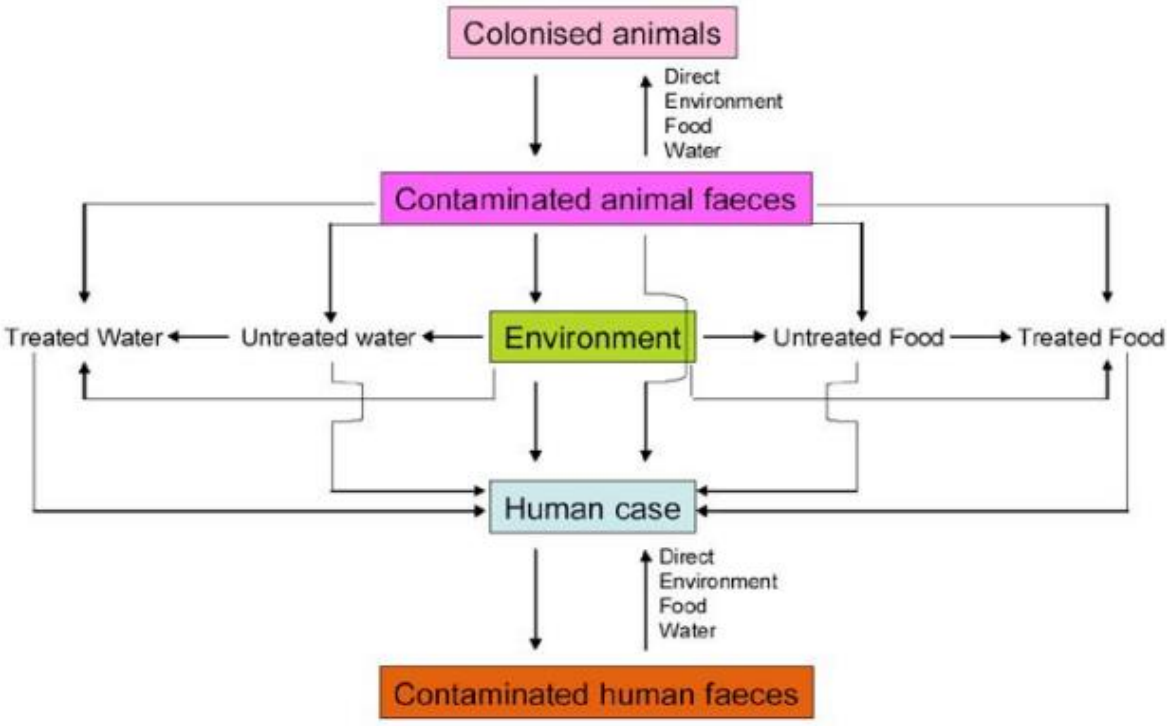
Figure 1.4, taken from the report and action plan of the Scottish *E. coli* O157 Task Force (43) describes the numerous routes by which STEC can be transmitted to humans from the zoonotic reservoir.

The predominant routes leading to human infections are via food (including water), person- to -person contact or contact with the ruminant reservoir and/or environments contaminated with their faeces. Irrespective of the route of transmission, cases present themselves sporadically (occurring independently of other cases) or as part of outbreaks where cases are microbiologically and epidemiologically linked in some way e.g., by eating at the same fast-food outlet.

Direct contact with the farming environment or ruminants, such as in open farms or petting zoos, are important risk factors for STEC infection (1, 44). One outbreak associated with a petting farm in South-East England in 2009 resulted in 93 infections, 27 admissions to hospital and 17 cases of HUS (45). Indirect contact with animals or environments contaminated with their faeces is also of importance but the actual process leading to infection is less well understood. The role of fomites has been considered for STEC O157 transmission between animals (46) but compared to other food borne pathogens (47-49), there is little published information specific to STEC O157 and fomites. Heavy rainfall and flooding events can lead to contamination of fresh and marine water systems (50),

beaches (51-55) and food crops. Poorly managed private water supplies present a particular risk in areas of high animal density (56, 57).

Figure 1.4 Overview of modes of transmission leading to colonisation of animals and infection of humans



1.8 Risk factors

Infection with STEC is the result of a complex set of interactions between distal and proximal risk factors related to the reservoir, the environment, the pathogen, the host and opportunities for transmission (30). Distal risk factors are those factors that have a remote or indirect causal influence on the disease outcome and include age, sex, ethnicity and socio-economic status. Proximal risk factors are those in the causal chain that precipitate disease and include the virulence of the infecting organism, food, water, exposure to animals and the rural environment as well as travel within the UK and abroad.

1.8.1 Risks presented by animals and the environment

The association between increased cattle density and rates of infection has been demonstrated in a number of studies in the UK and Europe. These are summarised in Table 1.1.

In the UK, most work has been carried out in Scotland on account of their persistently higher rates of infection compared to other countries. A spatial and temporal analysis in Scotland between 1996 and 1999 showed significant regional variation, a distinct temporal pattern linked to season and increased rates of infection from West to East and South to North. Spatial effects appeared to be distant rather than locally driven; that is the attributable risk of living at specified distances from an existing case was at its highest between 50 and 100km. The study also found cattle population density; human population density and the number of cattle per person were variously significant depending on geographical location (35). In Europe, the association has been examined in the Netherlands (58, 59), France (60), Germany (61), Sweden (62) and Finland (63).

Taken together, these studies all show that the risk of infection and/or poor clinical outcome (e.g. HUS) is associated with cattle or sheep density (14, 35, 59-64), farm density/rural residence (14, 59, 65, 66) and summer months (59) (14).

Table 1.1 Summary of studies exploring the role of the agricultural environment on STEC infection and HUS

Country	Dependent variable(s)	Significant risk factors	Statistical methods	Year	Reference
Sweden	Mean annual incidence 1995–1998	Farm density Cattle density	Multiple linear regression	2004	(62)
France	HUS incidence in children aged 15 years or under.	Total cattle density Ratio of calves to children	Multivariate Poisson regression	2005	(60)
Scotland			Poisson GLM, Scan statistic, K-function analysis.	2005	(35)
Germany	Counts of STEC O157 by serogroup	Cattle density Age	Bayesian Poisson regression	2008	(61)
Scotland	Incidence of STEC O157 stratified by phage type	Farm level prevalence of STEC in cattle stratified by phage type. Season Geography	Comparison of proportions using the Cochran Mantel Haenzel test.	2009	(42)
Netherlands	STEC O157 incidence	Cattle density Age Sex Season	Integrated nested Laplace approximation (INLA) with fixed and random effects.	2011	(59)
Finland	Counts of STEC	Bulls per population Adult education levels	Bayesian Poisson complementary-log log (clog log) hurdle model with and without spatial correlation variable	2011	(63)

Country	Dependent variable(s)	Significant risk factors	Statistical methods	Year	Reference
		Disadvantaged households with children			
		Farm density			
		Fresh water coverage			
Republic of Ireland	Spatial presence/absence of ≥ 1 STEC O157 or O26 case.	Livestock density	Multivariate logistic regression	2017	(64)
		Septic tank density			
		Private wells			
		Serotype			
		Age			
		Sex			

Swimming or paddling in fresh or seawater has been linked to outbreaks (51-54, 67). Swimming pools and paddling pools have also been identified as risk factors or the source of outbreaks (68-72). The proportion of land covered with freshwater was positively associated with risk in Finland (63).

Animal based attractions (petting farms, agricultural shows and zoos) have been associated with many outbreaks worldwide (45, 73-77). One outbreak in south-east England in 2009 affected at least 93 people (45) was subject to an independent investigation that resulted in over 40 recommendations focused on minimising visitor contact with animal faeces, raising public awareness about this risk, developing an approved code of practice for open farms, getting regulators to work together, and performing research on rapid diagnosis and the reduction of the carriage of STEC in animals (78).

1.8.2 Risks presented by food and water

Raw (unpasteurised) drinking milk (RDM) and products made with unpasteurised milk (cheese, yoghurt, cream etc.) may be contaminated with STEC (79) and have been associated with outbreaks worldwide (22, 80-89). The sale of RDM is illegal in Scotland and is regulated by the Food Standards Agency and the Dairy Hygiene Inspectorate in England.

A 2016 systematic review showed that of 32 studies that included consumption of pink or raw meat as a risk factor, 20 (62.5%) found this to be associated with STEC infection (90). Random-effects meta-analysis provided pooled odds ratios and population attributable fraction (PAF) of 19% for undercooked/raw meat, followed by person-to-person transmission at 15%. Contact with animals and visiting farm environments had PAFs of 14% and 12% respectively (90) although these results differed geographically.

Watercress was identified as a risk factor in a case control study in England (44). Outbreaks associated with leafy greens are becoming more common (11, 91) and have also been identified as a risk factor for specific STEC lineages (92).

The risk associated with mains drinking water is considered to be low (57) as STEC is susceptible to the range of contemporary treatments used. However, outbreaks linked to drinking water do occur, most notably that which occurred in Walkerton, Ontario in 2000 associated with over 2,000 cases and at least seven deaths (93). There is extensive evidence that private water supplies are a risk factor for STEC infection (14, 68, 90, 91, 94-109). The use of septic tanks to treat and dispose of sewage in rural areas was identified as a risk factor in the Republic of Ireland (64).

1.8.3 Socio-demographic risk factors

A study in England showed that Accident and Emergency attendance and hospitalisation because of STEC illness was higher amongst the most disadvantaged group compared to the least, suggesting potential lower ascertainment of milder cases or delayed care-seeking behaviour in disadvantaged groups (110). Advantaged individuals were significantly more likely to report salad/fruit/vegetable/herb consumption (110), non-UK or UK travel (14, 65, 110) and environmental exposure (walking in a paddock, soil contact) suggesting that other risks, such as person-to-person transmission, could be more important in the most disadvantaged groups (110).

An ethnographic study showed that in disadvantaged areas, gastrointestinal infections spread to multiple households within a small radius, compared to advantaged areas where illness was confined to one household or dispersed over long distances (111). These differences were shaped by historical, social, and economic contrasts in housing; social networks and childcare arrangements; employment and household income. Educational

attainment in disadvantaged households has also been suggested as a risk factor for STEC in low -income households in Finland (63).

Travel outside the UK is generally regarded as a risk factor for IID. Some Lineage II strains of STEC. are particularly associated with travel outside the UK (1, 12, 14).

Rates of infection are consistently higher amongst younger age groups (1, 11, 20, 66). This may be a surveillance artefact (parents are more likely to seek medical attention for their children than themselves), a reflection of testing protocols (112, 113) or, indeed, a real feature of the disease.

In those aged 18 or under, rates are comparable between the sexes. However, rates in women are higher across all adult age groups (1, 114) and women and children are more likely to progress to HUS (20).

The relative importance of these risk factors may also vary at different spatial scales (14). For example, the same seasonal distribution of cases is seen in countries separated by large distances and this is thought to reflect the presence of similar agricultural and climatic risk factors (30). However, these factors alone are unlikely to explain the considerable variation of infection rates between (115-118), and within (119), countries around the world, particularly when considering the comparable levels of carriage by cattle in those countries (120).

1.9 Overview of spatial methods

Identifying geographical areas with significantly higher or lower rates of infection has the potential to provide important clues on the presence of environmental or socio-demographic risk factors in particular areas compared to others. These clues can then be used to inform the design of epidemiological studies to generate the evidence base needed for sound public health policies designed to reduce morbidity. Routine integration of

spatial information with infectious disease surveillance data is increasingly common and statistical methods that allow precise delineation of high and low risk areas are widely available. These methods include area-based studies; global, local, and focused tests for spatial clustering; estimates of spatially varying risk; and spatiotemporal modelling.

1.9.1 Ecological studies

Area-based studies compare disease rates or counts between different populations, often combined with other data, to examine the effect of risk factors (62, 121).

1.9.2 Global, local and focused tests for spatial clustering

Global tests for spatial clustering, such as Moran's I (122) and the Diggle-Chetwynd statistic (123), identify whether there is a general tendency for cases to occur more closely together than would be expected compared to the underlying population at risk. Local and focused tests for clustering, such as Local Indicators of Spatial Association (LISA)(122) and Kuldorff's scan statistic (124), are used to identify specific concentrations of disease that are statistically significant and may require further investigation.

Global statistics help to describe the spatial structure of data (i.e., is it clustered, dispersed or uniform? Does autocorrelation exist?). However, they do not identify the location of clusters or quantify how spatial dependency varies from one place to another (125).

Local statistics quantify spatial autocorrelation and clustering within small areas that together comprise the study area. They quantify spatial dependency by identifying clusters of high or low values or outliers in a given locality (125).

Focused statistics quantify clustering around a specific location or focus and are particularly useful to identify and explore possible clusters of disease near potential sources of environmental pollutants (125).

1.9.3 Spatially varying risk

Methods to estimate spatial variation in risk are used to describe the change in risk over a given study area and include kernel smoothing, which forms a key component in the estimation of the kernel density-ratio or relative risk function (126-129), and spatial interpolation methods such as inverse distance weighting (130) and kriging (131).

1.9.4 Spatio-temporal modelling

Modelling approaches can either take the form of empirical or mechanistic models that consider the effect of space and time alongside other factors (132-134).

The mechanistic approach models individuals or populations as moving between discrete states (e.g., susceptible, infected, recovered) to represent the specific mechanisms of infection. The empirical (or statistical) approach uses associations to estimate or predict transmission variables or disease risk and includes so-called forecasting models. Both forms are used to project disease burden and to test competing hypotheses on what mechanisms best explain observed disease dynamics (135).

1.9.5 Application to STEC and other infectious intestinal diseases

Smith et al. (136) systematically reviewed the use of spatial methods in infectious disease outbreaks between 1979 and 2013. Most reports were from the United Kingdom and a range of techniques was used, including simple dot maps, cluster analyses and modelling approaches. Spatial methods were used in only 0.4% of the total number of published outbreaks, predominately for environmental or waterborne infections, and were applied in only one foodborne outbreak. Since 2013, spatial methods have been applied specifically to infectious intestinal disease data and have included tests for global (137-141) and local (137, 138, 141-147) clustering, spatial variation in risk (118, 139, 148), modelling and other approaches (118, 134, 138, 139, 143, 146, 149-152).

1.10 Contribution

This thesis explores the applied use of spatial and spatio-temporal statistical techniques to enhance the understanding of risk factors associated with STEC in England.

It explores hypothesised links from previous research conducted within the UK and further afield using a novel geodatabase that linked disparate information on risk factors (animal density, private water supplies, distance to coast, freshwater coverage) spatially linked to detailed information on human infections at a fine spatial scale.

For the first time in England, we investigated the effects of hypothesised risk factors on case incidence, described geographical differences in the distribution of STEC cases using phenotypic and genetic information, and explored the relationship between residential location and the risk of infection. We focused on sporadic infection, the major fraction of cases, and the role of residential location and the environment at fine spatial scales.

We also used a novel approach to explore the relationship between risky behaviour and likely place of exposure. We did this by comparing behaviours reported by cases living in high-risk areas with those living in low-risk areas as well as comparing the differences between those living in low-risk areas who had travelled within the UK compared to those who stayed at home.

We also demonstrate the flexibility of kernel density estimation during the first six months of the COVID-19 pandemic in England and applied the methodology to the routine surveillance of new variants of SARS-CoV-2 during 2021.

1.11 Thesis structure

The core of this thesis is divided into four data chapters. Each chapter is structured as a paper for publication in an academic journal.

Chapter 2 aimed to identify high or low risk areas of STEC O157 in England, and the lineages associated with these areas, between 2009 and 2015. We used kernel smoothing techniques to produce continuous risk surfaces, space-time slices and spatio-temporal animations. We identified specific areas of England where the risk of infection was significantly elevated and explored differences in the spatial distribution of genomic lineages associated with severe disease outcomes.

A similar version of this chapter has been published as: Elson R, Davies TM, Jenkins C, Vivancos R, O'Brien SJ, Lake IR. Application of kernel smoothing to estimate the spatio-temporal variation in risk of STEC O157 in England. Spatial and Spatiotemporal Epidemiology. 2020 Feb; 32:100305. Authors' contributions: RE conceived and designed the study. RE collected and analysed the data with technical advice from TMD. RE interpreted the data with support from IRL, TMD and RV. RE produced the draft and critical revisions were received from RE, IRL, RV, CJ and TMD.

Chapter 3 examined the spatial relationship between the occurrence of cases and the presence of environmental and socio-demographic risk factors associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England between 2009 and 2015. We used geographical information systems (GIS) to create a geodatabase linking human surveillance data with a range of information (animal density, type of domestic water supply, fresh water coverage, residential distance from coast, urban/rural setting) at a fine spatial scale.

A similar version of this chapter has been published as: Elson R, Grace K, Vivancos R, Jenkins C, Adak GK, O'Brien SJ, Lake IR. A spatial and temporal analysis of risk factors associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England between 2009 and 2015. Epidemiology and Infection. 2018 Nov;146(15):1928-1939. Authors' contributions: RE conceived and designed the study with KG. RE collected and

analysed the data. RE interpreted the data with support from IRL and RV. RE produced the draft and critical revisions were received from RE, IRL, RV, CJ and GKA.

Chapter 4 investigates the relationship between individual exposure to risk factors and residence in high-risk areas associated with STEC. Using the boundaries describing high risk areas from Chapter 2, we compared the exposures reported by cases living inside and outside areas considered high risk. Multivariable logistic regression was used to identify differences between risk factors for those living in high-risk areas as well as those travelling away from home during their incubation period.

A similar version has been drafted for submission as: Elson R, Davies TM, Jenkins C, Vivancos R, O'Brien SJ, Lake IR. Using spatial relative risk to identify modifiable risk factors for STEC O157 infection in England. Authors' contributions: RE conceived and designed the study. RE collected and analysed the data with technical advice from TMD. RE interpreted the data with support from IRL, TMD and RV. RE produced the draft and critical revisions were received from RE, IRL, RV, CJ and TMD.

Chapter 5 estimates the spatial-temporal variation of COVID-19 in England in the first six months of 2020. We applied the methodology described in Chapter 2 to demonstrate its application to a different disease. The work describes the introduction, spread and establishment of SARS CoV2 during a public health emergency. Our analysis shows the volatile nature of the initial stages of the pandemic and the subsequent increased risk observed in some large urban areas of England.

A similar version of this paper has been published as: Elson R, Davies TM, Lake IR, Vivancos R, Blomquist PB, Charlett A, Dabrera G. The spatio-temporal distribution of COVID-19 infection in England between January and June 2020. Epidemiology and Infection. 2021 Mar 8;149:e73. Authors' contributions: RE conceived and designed the

study. RE collected the data with GD and analysed the data with advice on bandwidth selection from TMD. RE interpreted the data with support from TMD and RV. RE produced the draft and critical revisions were received from IRL, TMD, PBB, AC.

2 Application of kernel smoothing to estimate the spatio-temporal variation in risk of STEC O157 in England

2.1 Abstract

Identifying geographical areas with significantly higher or lower rates of infectious diseases can provide important aetiological clues to inform the development of public health policy and interventions designed to reduce morbidity. We applied kernel smoothing to estimate the spatial and spatio-temporal variation in risk of STEC O157 infection in England between 2009 and 2015, and to explore differences between the residential locations of cases reporting travel and those not reporting travel. We provide evidence that the distribution of STEC O157 infection in England is non-uniform with respect to the distribution of the at-risk population; that the spatial distribution of the three main genetic lineages infecting humans (I, II and I/II) differs significantly and that the spatio-temporal risk is highly dynamic. Our results also indicate that cases of STEC O157 reporting travel within or outside the UK are more likely to live in the south/south-east of the country, meaning that their residential location may not reflect the location of exposure that led to their infection. We suggest that the observed variation in risk reflects exposure to sources of STEC O157 that are geographically prescribed. These differences may be related to a combination of changes in the strains circulating in the ruminant reservoir, animal movements (livestock, birds or wildlife) or the behaviour of individuals prior to infection. Further work to identify the importance of behaviours and exposures reported by cases relative to residential location is needed.

2.2 Introduction

Shiga-toxin producing *E. coli* (STEC) are a group of bacteria associated with human disease and are defined by the presence of one or both phage encoded Shiga toxin genes; Stx1 and Stx2 (1). The main reservoir is ruminant animals, particularly cows and sheep.

First recognised as a human pathogen in 1982 (3), STEC are now globally distributed (4). There is evidence that a common ancestor of STEC was introduced to countries around the world on a number of occasions in the past, likely due to international transport of animals and/or contaminated animal feed (34). Following introduction, localised genetic variation has occurred leading to a patchwork of strains that are related at the global level but show distinct geographical differences.

Infection with STEC is the result of a complex set of interactions between distal and proximal risk factors related to the reservoir, the environment, the pathogen, the host and opportunities for transmission (30). The relative importance of these factors may vary at different spatial scales (14). For example, the same seasonal distribution of cases is seen in countries separated by large distances and this is thought to reflect the presence of similar agricultural and climatic risk factors (30). However, these factors alone are unlikely to explain the considerable variation of infection rates between (115, 116, 153-156), and within (119), countries around the world, particularly when considering the comparable levels of carriage by cattle in those countries (120).

Within the United Kingdom, rates of STEC infection in Scotland are more than twice that of England (35). Within England, rates of infection vary considerably from 0.40 to 1.34 cases per 100,000 person years in London and the North respectively (11, 14) and there is evidence that this relates to living in areas with high densities of farmed animals (14). However, the strains infecting humans are not always the same as those circulating in the 'local' ruminant reservoir (37, 38). The reasons for this are unclear but may be due to widespread exposure to a remote source of infection, or localised exposure to a source where the availability of comparative microbiological information is scant (14, 37). Conversely, evidence from outbreak investigations shows that transmission of highly

related strains can occur via multiple routes from geographically restricted sources (15, 157).

Identifying geographical areas with significantly higher or lower rates of infection therefore has the potential to provide important aetiological clues. These can then be used to inform the design of epidemiological studies to generate the evidence base needed for sound public health policies designed to reduce morbidity. Routine integration of spatial information with infectious disease surveillance data is increasingly common and statistical methods that allow precise delineation of high and low risk areas are widely available. These methods include area-based studies; global, local and focused tests for spatial clustering; estimates of spatially varying risk; and spatiotemporal modelling.

Area-based studies compare disease rates or counts between different populations, often combined with other data, to examine the effect of risk factors. Global tests for spatial clustering, such as Moran's I (158) and the Diggle-Chetwynd statistic (123), identify whether there is a general tendency for cases to occur more closely together than would be expected compared to the underlying population at risk. Local and focused tests for clustering, such as Local Indicators of Spatial Association (LISA) (122) and Kuldorff's scan statistic (124), are used to identify specific concentrations of disease that are statistically significant and may require further investigation. Methods to estimate spatial variation in risk are used to describe the change in risk over a given study area and include kernel smoothing, which forms a key component in the estimation of the kernel density-ratio or relative risk function (126-129), and spatial interpolation methods such as inverse distance weighting (130) and kriging (131). Modelling approaches can either take the form of empirical or mechanistic models that consider the effect of space and time alongside other factors (132-134).

Smith et al. (136) systematically reviewed the use of spatial methods in infectious disease outbreaks between 1979 and 2013. Most reports were from the United Kingdom and a range of techniques was used, including simple dot maps, cluster analyses and modelling approaches. Spatial methods were used in only 0.4% of the total number of published outbreaks, predominately for environmental or waterborne infections, and were applied in only one foodborne outbreak. Since 2013, spatial methods have been applied specifically to infectious intestinal disease data and have included tests for global (137-140, 159) and local (137, 138, 140, 141, 143, 145-147, 152) clustering, spatial variation in risk (118, 139, 148, 153), modelling and other approaches (118, 134, 138, 139, 143, 146, 149, 151, 152).

The aims of this study were to identify areas presenting a higher or lower risk of infection with STEC O157 in England by estimating the space-time variation in risk over the study period and exploring any difference between the residential locations of cases reporting travel and those not reporting travel.

2.3 Methods

In England, isolates of *E. coli* O157 identified locally are sent for confirmation and typing at the Gastrointestinal Bacterial Reference Unit (GBRU). Detection and confirmation of STEC includes biochemical identification and serotyping of bacterial isolates. Since 1989, strains belonging to *E. coli* O157 have been further differentiated using a phage typing (PT) scheme developed in Canada (11). Retrospective real-time polymerase chain reaction (PCR) targeting *stx*1 or *stx*2 and the intimin (*eae*) gene, associated with intimate attachment of the bacteria to the host gut mucosa, was introduced in 2012 (11). Since 2015, all isolates have been routinely sequenced allowing identification of genetic lineage/sub-lineage and *stx* subtypes (13, 160).

The National Enhanced Surveillance System for STEC (NESSS) was introduced in England in 2009. The system collects clinical and epidemiological information for each

laboratory confirmed case using a standardised questionnaire. This includes details about whether the case had travelled abroad or within the UK prior to their illness onset and the residential postcode of each case (an alphanumeric reference developed by the UK Post Office to facilitate the delivery of mail, each containing around 15 addresses). This information is linked to reference microbiology information including PT, presence of virulence factors and whole genome sequence data (1).

Case selection

We selected primary cases of STEC O157 with valid postcodes reported to the NESSS between 2009 and 2015. Strains of STEC O157 circulating in humans fall into three distinct lineages (I, II and I/II) descended from a common ancestor. Lineage I contains PT 21/28 and PT32; strains encoding *stx*2 only and associated with more severe disease. Lineage II contains PT8 and Lineage I/II PT2 (12). Cases were categorised into these Lineages and Lineage II was further divided into sub-lineages IIa, IIb and IIc. Because routine whole genome sequencing (WGS) was not introduced until 2015, we extrapolated the phenotypic characteristics of PT and *stx* of strains identified by whole genome sequencing to isolates falling into Lineage II. This was not possible for isolates in Lineage I because sub-lineages are identified using the *stx* subtype which is inferred from the sequence result. The categorisation and numbers of strains are presented in Table 2.1.

The NESSS categorises cases into primary, co-primary, secondary, or unknown. This categorisation is given at the time of the case interview and is quality checked when the data are entered into the system. Primary cases are either those that are not epidemiologically linked to other cases or, in the case of household outbreaks, the case that developed symptoms first. We selected primary cases only and cases linked to known outbreaks were excluded.

Control selection

Controls were randomly sampled from the National Population Database (NPD) (161). The NPD is a point-based Geographical Information System (GIS) dataset that combines locational information from providers like the Ordnance Survey with population information about those locations, mainly sourced from UK government statistics. It consists of a number of dataset layers, including population data from the 2011 Census (162). Data are provided in a 100-metre by 100-metre grid situated on a centroid of the square with the population generalised to this level (161, 163). Four control locations per case were drawn without replacement. The probability of a location being sampled was weighted by the summed population of each grid square to reflect the spatially varying nature of the underlying population at risk.

Analytical strategy

We chose the kernel smoothing method because our primary interest was to identify large scale variation in risk as opposed to small-scale localised clustering (164). This method is also well suited to studying the occurrence of cases relative to the heterogeneous nature of the underlying at-risk population present in our data and the tools with which to perform the analyses are free and easily accessible (165).

The data used to estimate a particular relative risk surface are given as two distinct samples of planar points assumed to originate from (unknown, possibly equivalent) density functions f (cases) and g (controls) (166). A fixed or adaptive (167, 168) bandwidth determines the spread of smoothing kernels centred on each point, producing a nonparametric density estimate that can be evaluated at all locations within the spatial study region. The ratio of case density to control density is calculated to provide a continuous estimate of relative risk which can then be plotted on a map. Where $f > g$ there

is a peak in the surface (indicative of heightened risk); where $f \cong g$, the surface is flat (no difference in risk); and where $f < g$, there is a trough in the surface (lower risk). Specialised coordinate-wise hypothesis tests permit detection of statistically significant departures of these peaks and troughs from uniformity, and any such sub-regions can be delineated by drawing associated tolerance contours upon the risk surface in question (128, 129, 167).

Spatially varying risk

To estimate the spatially varying risk we created case-control datasets for all PTs, Lineages I, II and I/II and Sub-Lineages IIa, IIb and IIc. For all PTs, we included cases that reported travel abroad or within the UK in the seven days prior to the onset of symptoms. For the Lineage and Sub-Lineage analysis, only cases who reported no travel were included. The same control dataset described earlier was used for each analysis.

For all spatial risk surfaces we used adaptive kernel estimation following Abramson's square-root rule (169). This adaptation reduces the smoothing in areas of high point density (to capture more detail in the final estimate where we have an abundance of data), while increasing the smoothing in areas where the observations are relatively sparse (reflecting our greater uncertainty in areas where we do not have as much information). Such an approach has been shown to work extremely well for applications in geographical epidemiology (166-168, 170), but the issue of bandwidth selection is more complicated than in the fixed bandwidth case; we require selection of both a "pilot" and a "global" bandwidth value to initialise the estimator for a single density estimate. To simplify the selection problem, recent work has shown constraining these two values to be equal, as well as following an established practice of choosing equal values between both the case and control density estimates (128) offers both theoretical and practical benefits for the resulting risk function estimate.

As such, we followed these guidelines in producing all spatial risk surfaces in this work, calculated as symmetric adaptive risk function estimates using the pooled case/control data set to compute the variable bandwidth factors (168), using equal global and pilot bandwidths chosen simultaneously via the likelihood cross-validation methodology described in (171). The global bandwidth value was used for the fixed estimate in the sensitivity analyses. The far-right hand column of Table 2.1 reports the common case/control bandwidth found for each estimate.

All estimates are edge-corrected to account for kernel weight lost over the boundary of the study region (172, 173) and results are reported as log-relative risk surfaces $\log f - \log g$ for symmetry around the 'null' log risk value of zero. Finally, corresponding asymptotic p-value surfaces were estimated for each surface (167, 168), and contours were superimposed at the 5% significance level to delineate areas of significantly higher or lower risk.

To estimate the spatial effect of reported travel, we created a dataset containing case data only. Cases were marked with the following travel status categories: 'Foreign travel' (cases reporting travel outside the UK in the seven days prior to onset); 'Any travel' (cases reporting foreign travel and/or travel within the UK in the seven days prior to onset) and; 'No travel' (cases reporting no travel either in the UK or abroad in the seven days prior to onset). We calculated the spatial relative risk for reported foreign travel by comparing cases in the 'Foreign travel' category to those falling into the "Any" and "No" travel categories. To produce the risk surface for 'Any travel', we compared cases falling into the 'Any travel' category with those in the 'No travel category'.

Rural residence is known to be associated with an increased risk of STEC infection in England (14). To explore the potential confounding effects of this on our analysis, we conducted two sensitivity analyses using both fixed and adaptive bandwidths. The first was

restricted to rural areas only and the second used data stratified by urban/rural residence. For both these analyses we compared fixed to adaptive bandwidths to explore whether they produce similar results.

Spatio-temporal risk

Creating a dataset containing all cases marked with the month of disease onset as a temporal event permits exploration of the temporal variation in the spatial risk of STEC O157. However, estimation of spatio-temporal relative risk is somewhat more complicated than purely spatial risk, and the properties of adaptive kernel estimators for such functions have not yet been studied in sufficient detail in the statistical literature. Thus, we approach these estimates using the Fernando-Hazelton fixed bandwidth kernel estimator (174). Each spatio-temporal density estimate requires a separate smoothing bandwidth for the spatial and the temporal margins of the data. As in the purely spatial setting, it is recommended to choose the same values of these bandwidths between the case and control estimates. For the sake of comparison, we produced fixed-bandwidth relative risk surfaces (174) using two bandwidth prescriptions. The first used the maximal smoothing principle proposed by Terrell (175) applied separately to the spatial and temporal margins of the data. The second used the fixed bandwidth cross-validated likelihood method (171) to produce a risk surface with less smoothing. Estimates were edge corrected using the same methodology as mentioned earlier and results are reported as raw-risk estimates for ease of interpretation. Asymptotic p-value contours are again superimposed to identify areas of elevated risk only at the 5%, 1% and 0.01% significance levels.

Data preparation was performed using ArcMap v10.2 (176). All subsequent analyses were performed using the contributed packages sparr (166) and spatstat (177, 178) in the R language (179). Bandwidth selections were performed using cases and/or controls falling within a simplified polygon of the mainland boundary of England.

2.4 Results

The spatial locations of all unmarked cases and controls are shown in Figure 2.1. A total of 3,592 cases and 14,392 controls were considered for analysis. The majority of cases fell into Lineages I and II (Table 2.1). Just over half of all cases (1,942; 54%) reported no travel in the seven days preceding the onset of their symptoms, 29% (1,029) reported foreign travel and 17% (621) reported travelling within the UK (Table 2.1). Over half of the cases (2011;56%) were female and most (2157; 60.1%) were adults aged over 18 years or more. One fifth of cases (735; 20.1%) were children aged five years or less and the remainder (700; 19.5%) were children aged between 6 and 18 years.

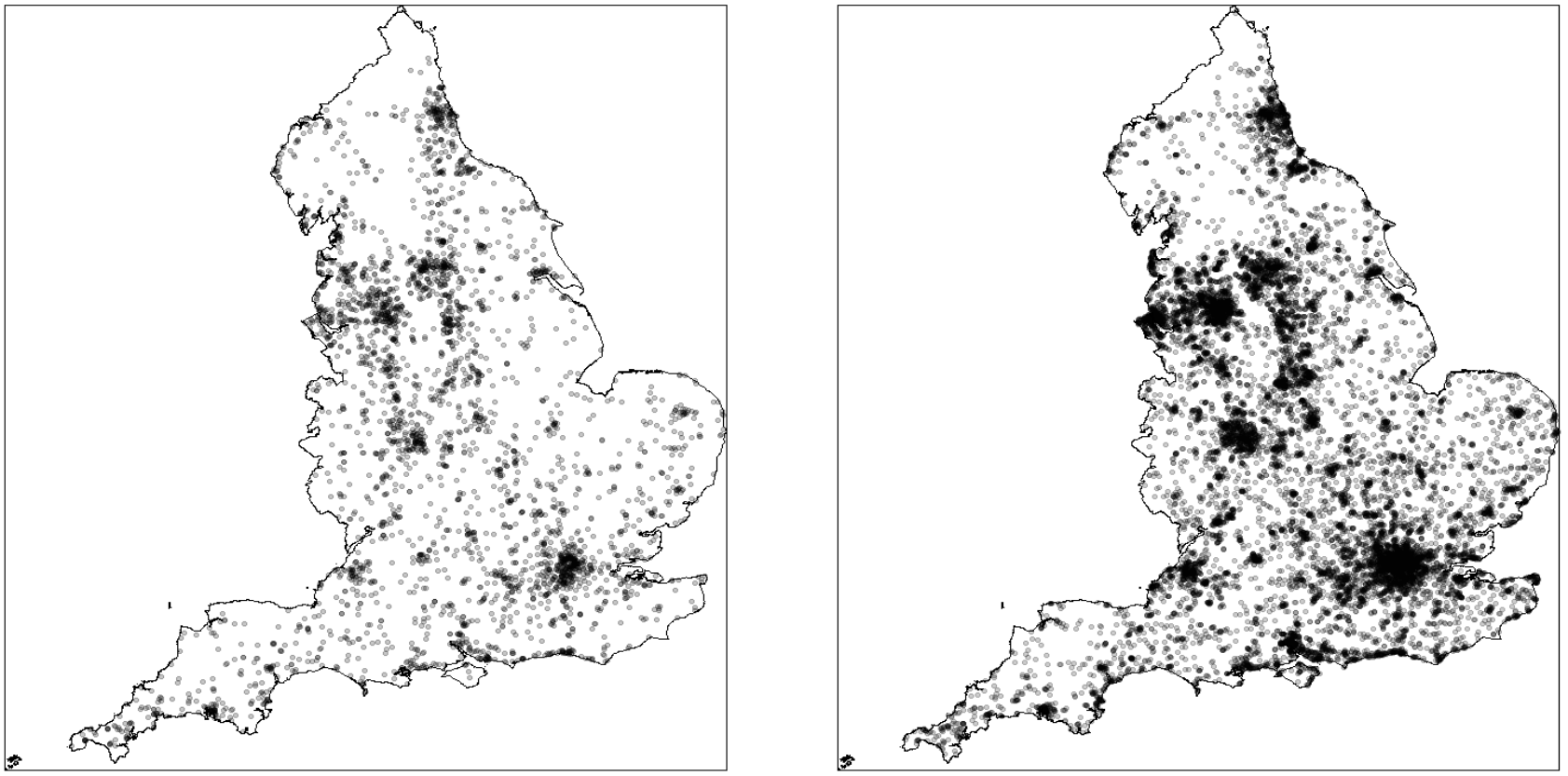
The relative risk surface for all cases (including those reporting travel) is shown in Figure 2.2. There were three main areas where risk was significantly higher compared to the underlying population at risk. These were in the North/North West of the country and the South West. Areas of significantly lower risk were largely confined to the South.

Table 2.1 Case selection criteria and associated common case-control bandwidths

Case details		PTs	stx	n	Common smoothing bandwidth (km)
All cases		-	-	3,592	9.39
Reporting foreign travel		-	-	1,029	31.84
Reporting any travel		-	-	1,650	31.84
Lineage I*		21/28, 32	2	752	12.37
Lineage II*		4,8,34,54	-	778	18.10
	Sublineage IIa	34,54	2	134	92.79
	Sublineage IIb	4,8	2	140	60.15
	Sublineage IIc	8,54	1&2	493	20.31
Lineage I/II*		2	2	120	21.69
Others		1,14,31,33,46,51,8 (*stx*1),4(*stx*1&2)	-	652	-

*- cases reporting no travel

Figure 2.1 Spatial location of 3,592 STEC O157 cases (left panel) and 14,392 randomly selected controls (right panel)



The relative risk surfaces for Lineages I, II and I/II are presented in Figure 2.3. For Lineage I, the greatest risk was largely seen in the North West and South West of the country. Areas of lower risk were confined to the Midlands and South as well as a small urban area in the North West.

Compared to Lineage I, the risk surface for Lineage II was more uniform across the country. Areas of significantly elevated risk for Lineage II were confined to the North and North West, and two areas in the South West of the country. Areas of significantly lower risk were largely restricted to the extreme South and South East of the country.

For Lineage I/II, areas of significantly higher risk were restricted to the North, the East and the far South West of the country. Areas of significantly lower risk were located in the South East.

The relative risk surfaces for Sub-Lineages IIa, IIb and IIc are presented in Figure 2.4. For Sub-Lineage IIc, areas of significantly elevated risk appeared in the North West and the South West. Areas of significantly lower risk were located in the South and the far South East. The risk for IIa appeared highest in the far South West and for IIb across the North and South West of the country but these were not statistically significant.

The results of the spatiotemporal analysis are best viewed in the animation provided in the supplementary material (SM2). This shows that the spatio-temporal risk was largely confined to the north and South West of the country but was highly dynamic within and between these areas. The over-smoothed surface (left panel in the animation) showed an area of elevated risk largely restricted to the far North West. In late 2010, this area expanded to the East and South and persisted across the North of England for two years before disappearing towards the end of 2013.

In the South West, risk was similar to the North but lower between 2010 and 2013, after which the highest risk areas were seen in this area. Compared to the North, the areas of high risk were more mobile and appeared in different areas from year to year.

Figure 2.5 shows the two risk surfaces for cases reporting foreign travel and for those reporting foreign travel or travel within the UK in the seven days preceding onset of symptoms. Cases reporting travel were significantly more likely to live in the South and South East of the country than cases who reported no travel, who were more likely to live in the North or South West.

The results of the sensitivity analysis comparing the main results with those of the rural areas only and the analysis stratified by urban/rural residence are presented in the supplementary material as SM 3. Each analysis identified broadly the same areas of higher and lower risk identified by the main analysis. When compared to the adaptive surfaces, those produced using the fixed bandwidths were ‘noisier’, even though both generally agree on areas of heightened and lowered risk. This is likely the result of simultaneous over- and under-smoothing in different areas of the study region; a common symptom of fixed-bandwidth estimation (180).

Figure 2.2 Estimated log relative risk for all cases of STEC O157 (including cases reporting travel). Tolerance contours are superimposed as solid lines at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk

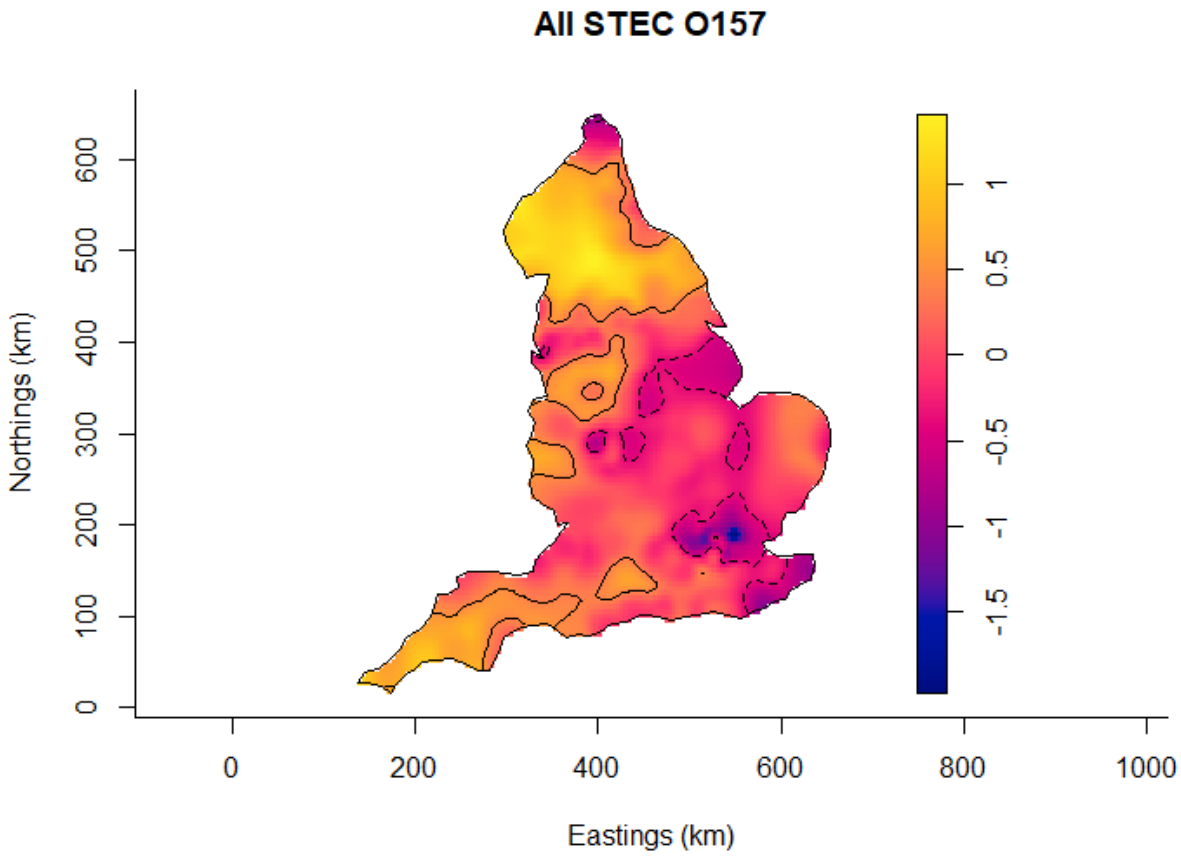


Figure 2.3. Estimated log relative risk for STEC O157 Lineages I, II and I/II in cases reporting no travel. Darker areas indicate areas of lower risk. Tolerance contours are superimposed as solid lines at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk

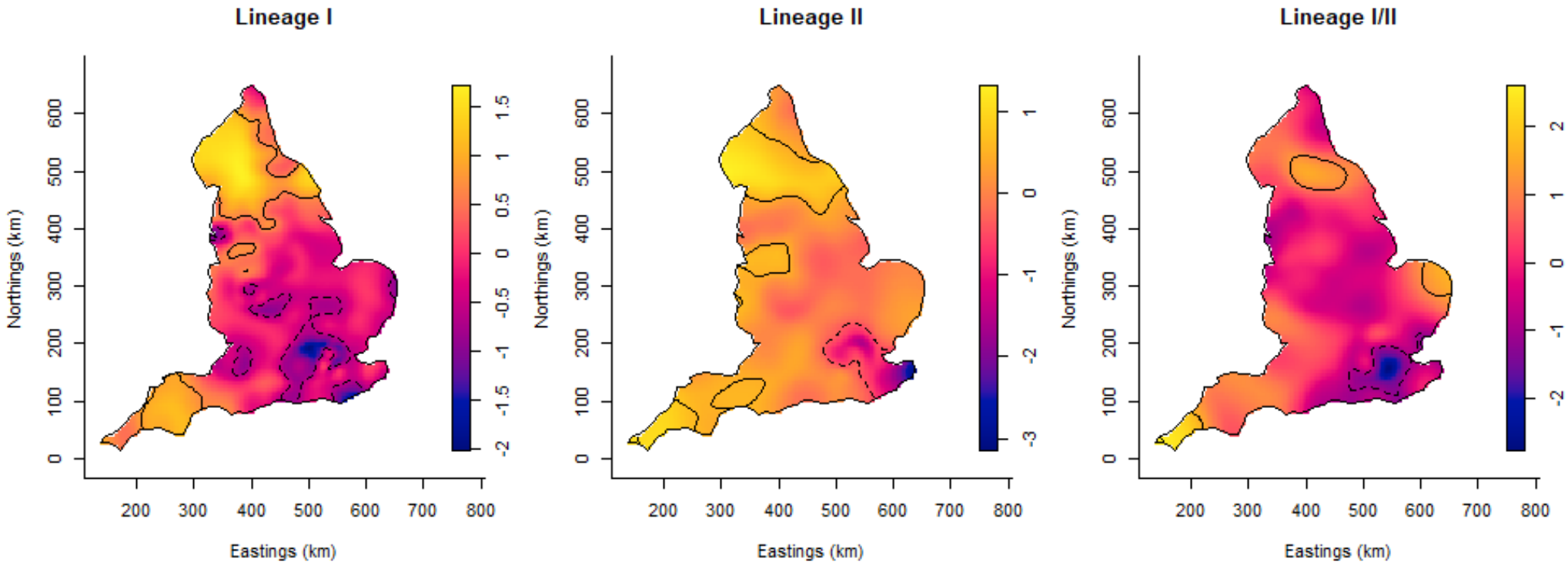


Figure 2.4 Estimated log relative risk for STEC O157 sub-lineages of Lineage II in cases reporting no travel. Darker areas indicate areas of lower risk. Tolerance contours are superimposed as solid lines at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk higher risk and dashed lines indicate areas of lower risk

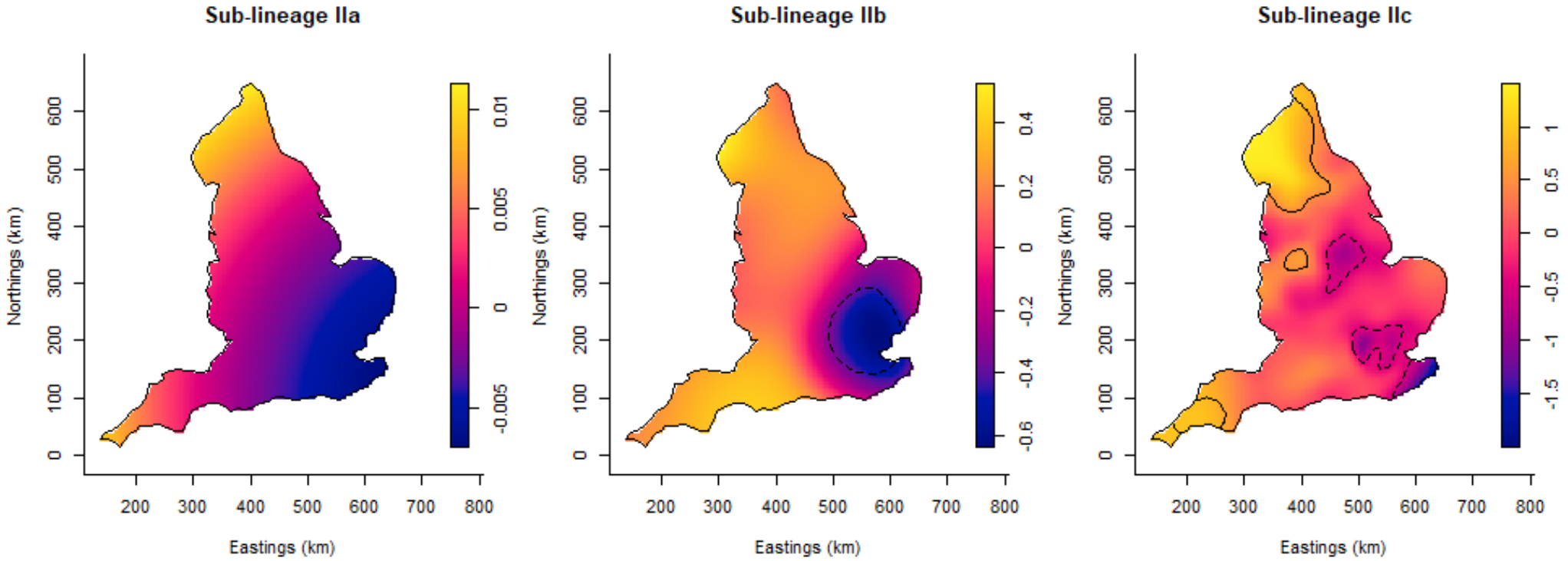
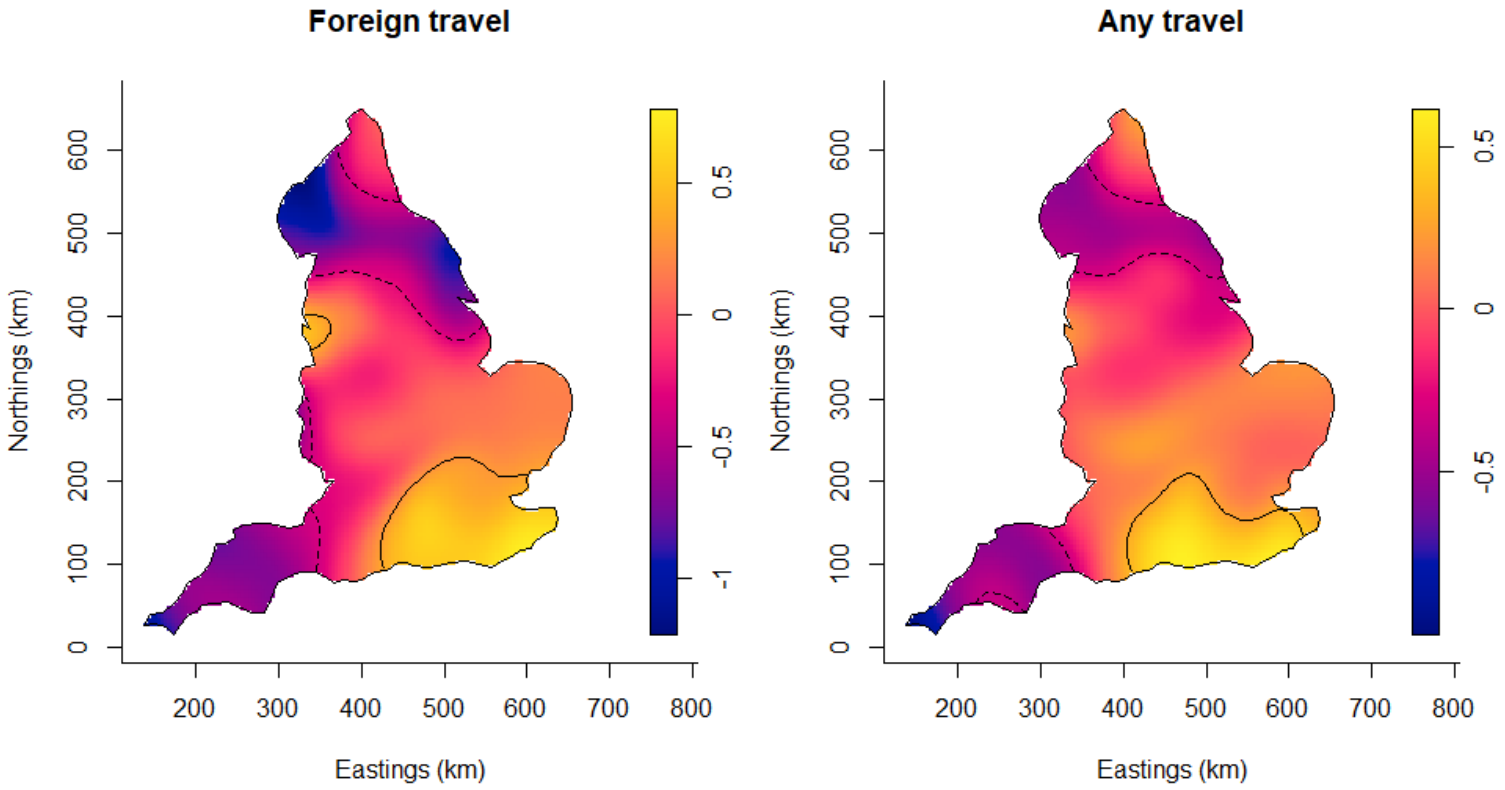


Figure 2.5 Estimated log relative risk for cases of STEC O157 reporting foreign travel (left panel) and those reporting any travel. Tolerance contours are superimposed at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk higher risk and dashed lines indicate areas of lower risk.



2.5 Discussion

Our analysis provides evidence that the distribution of STEC O157 infection in England is non-uniform with respect to the distribution of the at-risk population; that the spatial distribution of the three main genetic lineages infecting humans differs significantly and that the spatio-temporal risk is highly dynamic. We also provide evidence that cases of STEC O157 reporting travel within or outside the UK are more likely to live in the south/south-east of the country, meaning that their residential location may not reflect the location of exposure that led to their infection. We propose that the observed variation in risk is likely to reflect a differential exposure to a source of STEC O157 that is geographically prescribed.

Comparison with other studies

Contact with the agricultural environment is a known risk factor for STEC infection (14, 44, 90, 181). Within the British Isles, increased risk of STEC O157 infection is associated with rural areas where there are high densities of animals (particularly cattle and sheep) and less likely to be served by mains water supplies (14, 64, 182, 183). There is evidence that the spatial distribution and relative importance of risk factors differ by pathogen sub-type (14, 64, 118) and similar findings have been produced from Northern European countries (35, 59-61, 63, 184), the United States (118), Canada (184, 185) and New Zealand (119).

Our analysis is exploratory and therefore inference regarding causation cannot be drawn. However, the areas of elevated risk presented here are consistent with findings from other studies in that they are predominately rural areas with sparse populations, high densities of farmed animals and with greater numbers of private water supplies (14). They also share similar locations to national parks; popular destinations for day trips for local residents and longer holidays, particularly for those living in the south and south east

of England (186). In contrast to most farmland in England, public access to National Parks is largely unrestricted and visitors often camp, walk or cycle in areas where animals and/or their faeces are present (186, 187).

The importance of the pathways through which pathogens are transferred from the environment to humans is subject to debate (187).However, because of their low infectious dose, widespread prevalence in farmed animals and their ability to survive in the environment for extended periods of time STEC are well suited to environmental transmission. Recent studies using boot sock sampling over wide geographical areas demonstrate that *Campylobacter* (187) and STEC (188) can be recovered from boots following recreational walks in the countryside. The rate of recovery for both pathogens was highest in North West England (47% for Campylobacter and 25% for STEC) and is likely a reflection of high densities of cattle and sheep in this part of the country (187).

Spatial variation in risk at Lineage and Sub-Lineage level

Strains falling into Lineage I/II were the dominant strain infecting humans in England for many years but are now uncommon (11) and our analysis demonstrates that these strains are also spatially restricted. Lineages I and II have dominated since the late nineties (11) and this is reflected in the geographically widespread areas of elevated risk seen in broadly similar areas of the country. However, at regional level, the spatial distribution of the three lineages differed. Increased risk of infection with STEC in England is generally associated with residential proximity to high densities of farmed animals, however, risk of infection with Lineage I strains is particularly associated with sheep density (14). This suggests that the presence of particular lineages in the environment is uneven and dependent, at least to some extent, on the underlying distribution of the zoonotic reservoir. This finding is consistent with the distribution of *Campylobacter* sp. in the environment relative to the presence of different animal species in England (187).

Spatio-temporal relative risk

The two versions of the animated spatiotemporal risk surface provide the opportunity to critically appraise the detected sub-regions of significantly elevated risk. For example, a large area that remains significant over an extended period of time in the over smoothed estimate (left panel in animation SM 2) could, to a certain extent, be a methodological artefact arising from too generous a bandwidth. However, if certain smaller pockets within such a sub-region persist for noticeable periods in the noisy ("less-smooth") estimate (right panel in animation SM2), this indicates that anomalies in the infection rates are genuine, in turn suggesting these are a result of geographically restricted source. This was indeed the case, particularly in the North and South West. The appearance, persistence and decline of an area of very high risk in the north of England between 2010 and 2013 appeared distinct to activity elsewhere in the country and corresponds with an unexplained decline in Lineage 1 strains, particularly in rural areas (14).

Bandwidth selection for kernel estimation

Choosing an optimal bandwidth is important for making reliable inference from relative risk surfaces. Even with tailored bandwidth selection methods (189), classical fixed bandwidth estimators can be unstable and do not cope well with the smoothing requirements of highly heterogeneous patterns (165, 167, 180). However, choosing appropriate smoothing parameters for the more sophisticated adaptive estimator is far more difficult, and this is an active area of research (165, 166, 180).

We used a recently developed likelihood-based selection strategy for the purely spatial analyses (171), and while theoretically valid, further research into how well this type of simultaneous global/pilot bandwidth selection might perform in practice is warranted. This bandwidth selection method did not identify an optimal bandwidth within a scale-

appropriate range for the risk surfaces of Sub-lineages IIa and IIb, erring toward excessive smoothing. Such a result is suggestive of spatial uniformity of risk, though the relatively low numbers of cases falling into these sub-lineages may, at least in part, be to blame in these instances. Of note is that Sub-lineage IIb (an unusual clone of PT8 encoding stx2), only emerged in significant numbers following an outbreak towards the end of 2015 (157, 190) and so fell outside the scope of our analysis. Further work on the recent spatio-temporal nature of this event is recommended.

Cases reporting foreign travel or travel within the UK

To provide the best estimate of indigenous risk, our study design at Lineage and Sub-lineage level did not consider cases reporting travel and did not therefore capture the possible location of exposure related to foreign or UK travel. Notwithstanding this, the inclusion of cases reporting travel made little difference to the overall results suggesting that the distribution of these cases is broadly similar to the underlying population at risk. However, when considering spatial relative risk *between* cases, those who did report travel were significantly more likely to live in the south and south-east of the country. This is consistent with previous findings that for these cases, exposure to risk factors not present in their residential environment are important when considering the source of their infections (14).

Data quality and potential limitations

One potential limitation to our study is that for every STEC O157 infection reported to national surveillance systems in England, there are an estimated 7.4 in the community (29). The reasons for this are likely to be related to severity of disease, health seeking behaviours and whether a clinician takes a sample and requests a microbiological

examination from a laboratory. It is unknown whether these reporting biases vary geographically and hence would affect the spatial patterns presented in this paper.

There were no changes to laboratory methods or surveillance systems during the study period (11). However, a large petting farm outbreak in 2009 (45) attracted media attention and prompted a review of national guidelines for the public health management of STEC which had the potential to improve case ascertainment and follow-up from 2010 onwards, as well as reducing risk.

In addition, the Health Protection (Notification) Regulations 2010 (9) came into force during the study period. This legislation introduced the mandatory reporting of STEC as a causative agent, and haemolytic uraeamic syndrome (HUS) as a notifiable disease. Our results do not suggest that these events created a reporting differential based on severity of disease because risk is elevated in similar geographical areas for Sub-lineage IIc strains that tend to be associated with less severe symptoms than those falling into Lineage I.

We also considered the effect of rural versus urban residence in our sensitivity analysis, the results of which suggest that the observed spatial variation is unlikely to be explained by rural residence alone, and that the adaptive bandwidths used in this paper do not produce different results to fixed bandwidths.

Conclusion

To conclude, the risk of sporadic infection with STEC O157 varies significantly across England. We suggest that this is due to differential exposure of the population to geographically restricted risk factors. The appearance, expansion and decline of an area of significantly elevated risk in the north of England between 2010 and 2013 corresponds with an overall reduction of STEC O157 in England, seen most acutely in PT21/28

reported in rural areas (14). Cases reporting travel prior to onset of illness were more likely to live in south of England.

These differences could be related to a combination of changes in the strains circulating in the ruminant reservoir, animal movements (livestock, birds, or wildlife), contaminated animal feed or the behavior of individuals prior to infection. Further work to identify the importance of behaviours and exposures reported by cases relative to residential location is needed. Statistically speaking, designing a semi-parametric, generalised additive style of model (see for example (191, 192) is one way we could build in extraneous predictors and estimate any associated effects on infection risk in such an analysis. We anticipate the findings in this work will help guide such future research endeavours.

3 A spatial and temporal analysis of risk factors associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England between 2009 and 2015

3.1 Abstract

Infection with STEC O157 is relatively rare but has potentially serious sequelae, particularly for children. Large outbreaks have prompted considerable efforts designed to reduce transmission primarily from food and direct animal contact. Despite these interventions, numbers of infections have remained constant for many years and the mechanisms leading to many sporadic infections remain unclear.

Here we show that two thirds of all cases reported in England between 2009 and 2015 were sporadic. Crude rates of infection differed geographically and were highest in rural areas during the summer months. Living in rural areas with high densities of cattle, sheep or pigs and those served by private water supplies were associated with increased risk. Living in an area of lower deprivation contributed to increased risk but this appeared to be associated with reported travel abroad. Fresh water coverage and residential proximity to the coast were not risk factors.

To reduce the overall burden of infection in England, interventions designed to reduce the number of sporadic infections with STEC should focus on the residents of rural areas with high densities of livestock and the effective management of non-municipal water supplies. The role of sheep as a reservoir and potential source of infection in humans should not be overlooked.

3.2 Introduction

Shiga toxin-producing Escherichia coli (STEC) are a group of bacteria associated with human disease and are defined by the presence of one or both phage encoded Shiga toxin genes; Stx1 and Stx2. Compared to other bacterial pathogens, it is a relatively rare

infection in many parts of the world but is of public health concern due its low infectious dose (<100 bacteria) (1) and potential to cause severe disease (27).

Worldwide, it is estimated that there are around 2.8 million cases annually, leading to 3,890 cases of haemolytic uraemic syndrome (HUS) and 230 deaths (26). The Europe wide rate of infection is estimated to be 1.4 cases per 100,000 population but reported rates vary between countries (Range: <0.1 to 12.4 cases per 100,000 population) (193). The O157 STEC serogroup is most commonly associated with human disease in the UK however, other serogroups are seen more frequently in other European countries (27, 193). Rates of infection in England have remained fairly constant for many years (around 1.65 (95% CI 1.49–1.81) cases/100,000 person-years) (11). Europe has shown a similar pattern with an increase since 2011 attributed to wider use of molecular methods following a large outbreak linked to sprouted fenugreek seeds. Rates of infection are highest in children and most cases occur in the late summer, at least in temperate areas, and this pattern is seen universally (30) .

Healthy cattle are the main reservoir of STEC although they are also carried by sheep and other animals (194). Animals shed a range of phage types (PTs) with the most prevalent in UK cattle being PT21/28, 8 and 34, PT 4 in sheep and PT2 in pigs (195, 196). STEC O157 survives well in the environment, remaining viable for many months in temperate conditions (197, 198).

Transmission to humans occurs through multiple routes. Cases present themselves sporadically (occurring independently of other cases) or as part of small outbreaks due to person-to-person spread in closed settings, particularly childcare facilities. The low infectious dose of STEC means that once in the population, person-to-person spread is common (1). Larger outbreaks tend to be associated with foodborne transmission, with an increasing trend towards salad vegetables and away from meat and dairy products (11).

Direct contact with the farming environment or ruminants, such as in open farms or petting zoos (1), are important risk factors for STEC infection (44). Indirect contact with animals or environments contaminated with their faeces is also of importance but the actual process leading to infection is less well understood. Heavy rainfall and flooding events can lead to contamination of fresh and marine water systems (50), beaches (51-55) and food crops. Poorly managed private water supplies present a particular risk in areas of high animal density (56)(57).

Phylogenetic analysis of strains circulating in humans and UK cattle during 2014 described three distinct lineages (I, II and I/II) descended from a common ancestor. Lineage I contains PT 21/28 and PT32; strains encoding Stx2 only and associated with more severe human disease. Lineage II contains PT8 and Lineage I/II PT2. Isolates from humans and UK cattle are closely related suggesting that PT8 and PT21/28, have a domestic source and are domestically acquired (12). With the advent of routine whole genome sequencing (WGS), it is now possible to identify links between cases that previously appeared sporadic in nature. These cases may exhibit spatial clustering, sometimes over long periods of time, suggesting that geographically restricted transmission of highly related strains can occur (15).

Ecological studies in the UK, Europe and further afield demonstrate a spatial association between rates of infection, cattle density, and other factors and all describe a seasonally driven picture with rates highest in the summer (59-61, 63, 64, 153, 183, 185). There are limitations to these studies. Sheep were only considered in one study (64), despite being a known reservoir. The study populations were restricted to children (60) or focussed upon severe disease (60). Some studies may also have included cases linked to outbreaks which is not ideal as the source of their infection may have differed from sporadic cases (185). Cases reporting foreign travel were included in some studies (60, 61, 153, 185), but not

others (35, 59, 63, 153). Finally, all these studies were performed at serogroup level only, even though different subgroups may have different sources and hence potentially different risk factors.

In this study we overcome these limitations using enhanced surveillance of STEC which has been performed in England since 2009. These data arguably represent the most comprehensive dataset for STEC infections in the world. This allows accurate identification of cases who have been part of an outbreak (and so are not representative of all cases) and those who report travel abroad or within the UK

This study had three aims. The first was to describe the spatial and temporal distribution of sporadic STEC O157 cases in England, the second was to test the relationship between the numbers of infections and hypothesised risk factors, the third was to test whether these risks differed by STEC subtype. Finally, we explored how these risks varied between all sporadic cases and sporadic cases when those reporting national or foreign travel are excluded.

3.3 Methods

Isolates of STEC O157 identified locally are sent for confirmation and typing at the Gastrointestinal Bacterial Reference Unit (GBRU). Detection and confirmation of STEC includes biochemical identification and serotyping of bacterial isolates. Since 1989, strains belonging to STEC O157 have been further differentiated by using a phage typing scheme developed in Canada (11).

The National Enhanced Surveillance System for STEC (NESSS) was introduced in England in 2009. The system collects clinical and epidemiological information for each laboratory confirmed case using a standardised questionnaire. This information is linked to

reference microbiology information including PT, presence of virulence factors and whole genome sequence data.

The case definition for the purpose of this study was a sporadic case of STEC, confirmed by GBRU and reported to the national enhanced surveillance system for STEC (NESSS) between January 1st 2009 and December the 31st 2015. An overview of the data selection process is shown in Figure 3.1.

The main aim of our analysis was to estimate the effect of hypothesised risk factors on the occurrence of sporadic cases (i.e., those occurring independently of each other). We therefore excluded cases linked to known outbreaks because their residential location rarely reflects exposure to the source of their infection, particularly for large outbreaks linked to widely distributed foodstuffs. Cases linked to household outbreaks were also excluded. Household outbreaks were identified as those where at least two cases had isolates of the same serotype and phage type that were collected within six months of each other, processed by a laboratory in the same Health Protection Team area and sharing the same surname and/or UK postcode.

The postcode (an alphanumeric reference developed by the UK Post Office to facilitate the delivery of mail and each containing around 15 addresses) for each case was geocoded to provide a spatial reference, allow visual display of the location of cases and enable the details of each case to be spatially joined to other datasets at the Lower Super Output Area (LSOA) level defined by the Office for National Statistics (162). LSOAs were chosen because they provide the most homogenous unit in terms of geographical size (Mean 3.9 km^2, Range 0.02-684 km^2) and population size (Mean 1,613 persons, Range 985-8,300 persons).

Crude incidence rates were calculated using the total population denominator data for each LSOA drawn from the last ONS Census performed in 2011 (162).

For each LSOA, a dependent variable, indicating the number of cases that occurred during the study period was created. Because STEC is an uncommon infection, the majority (90.1%) of LSOAs had no cases, 9.3% had one case and less than 0.6% had more than one case.

Dependent variables were created for all PTs, PT21/28 and PT8, further divided into three classes (all reported cases, those not reporting foreign travel and those reporting no travel either within the UK or abroad) giving a total of nine dependent variables.

The following explanatory variables were constructed for each LSOA:

Livestock density variables for cattle, sheep and pigs were calculated using the agricultural census of 2010. This census is performed every ten years by the Department for the Environment, Food and Rural Affairs and collects detailed information on land usage and livestock populations. Farm level data are aggregated to a 5x5 km grid and individual farms are not identified.

Estimates of deprivation were obtained from the Office of National Statistics (ONS). The index of Multiple Deprivation (IMD) was obtained for England for 2011 (162) which provides a set of relative measures of deprivation for LSOAs. This is based on seven domains of deprivation (income, employment, education, health, crime, housing and living environment). These domains are combined and ranked to produce the overall IMD score for each LSOA. For our analysis these data were divided into quintiles where quintile 1 is most deprived.

The degree of rurality for each LSOA was derived from the ONS rural urban classification used to distinguish rural and urban areas in England and Wales in 2011 (199). The

classification defines areas as rural if they are outside settlements with more than 10,000 resident population. For LSOAs, there are four urban classes (major conurbations, minor conurbations, cities and towns, cities, and towns in sparse settings) and four rural classes (town and fringe, town and fringe in a sparse setting, village and dispersed settlements, village and dispersed settlements in sparse settings). Due to the small numbers of cases resident in areas considered sparse, we grouped these eight classes into five by merging the three sparse categories with the corresponding non-sparse categories.

Outbreaks have been linked to beaches (51, 54), hence the straight-line distance from the centroid of each LSOA to the GB coastline was calculated in kilometres.

Inland water coverage was identified as a risk factor in a Finnish study (63). The shapes and areas of inland water features were extracted from the Ordnance Survey Master Map Topography Layer, summed and divided by the area of each LSOA to provide a proportional measure of freshwater coverage.

For each LSOA, the count of private water supplies was calculated using data submitted by local authorities to the Drinking Water Inspectorate during 2016. Local authorities are responsible for the enforcement and monitoring of the Private Water Supplies (PWS) Regulations (200) which require PWS to meet certain standards and for the location of each supply to be recorded. Three classes were created (0, 1-20 and >20 supplies).

Because the datasets used for inland water coverage and animal density differed from LSOAs in terms of geographical area or shape, we used a geographical information system (GIS) overlay function to create proportional measure in km^2 for each LSOA.

We used Jenks' natural breaks method to create four categorical variables for each animal species, distance from the coast and inland water coverage. This method is designed to

determine the best arrangement of values into different classes by seeking to minimise the variance within classes and maximise the variance between classes (201).

Statistical analysis

We considered three methods of regression analysis: Poisson, negative binomial and zero inflated Poisson. The results of a likelihood ratio test of alpha and goodness of fit test following Poisson regression indicated that the data were over dispersed. The same analysis was repeated using negative binomial and zero inflated negative binomial regression respectively. The Vuong test was not significant indicating that the standard negative binomial approach was best suited to the data. Proceeding with the negative binomial regression approach we conducted a multivariable analysis for each dependent and the independent variables. The first set of dependent variables were all sporadic cases, all sporadic cases minus those reporting foreign travel and all sporadic cases with no foreign or domestic travel. Two further models were then produced focussing upon PT21/28 and PT8 only. Person years (the total population of each LSOA multiplied by the years of observation) was included as an exposure variable. None of the multivariable analyses showed any associations with the distance from the coast and inland water coverage variable, hence these were removed from the analysis. The remaining independent variables were all included in the nine models to allow greater comparability between models. The dependent and independent variables were checked for correlation using Spearman's rank test. All coefficients showed low to moderate correlation except for cattle and sheep density with a coefficient of 0.7. An analysis for collinearity indicated that the addition of each independent variable in turn did not lead to significant changes in the coefficients or significance of any other variables in the model. Presenting each livestock density variable to the model as continuous variables did not affect the results.

Results are presented in terms of Incidence Rate Ratio (IRR) estimates and the 95% Confidence Interval (CI). The overall significance for a variable was estimated using the Wald test. All statistical analyses were performed using Stata version 13 (202).

3.4 Results

Rates of infection

Figure 3.2 describes the process followed to determine whether a case should be included in the study. A total of 3,559 (34% of 5,783) cases were eligible for inclusion in the statistical analysis.

The crude incidence of all sporadic confirmed STEC O157 cases (including those reporting foreign travel) reported during the study period was 9.1 per million person years. The rural rate (13.3 per million person years) was 1.6 times higher than that of the urban rate (8.1 per million person years). Rates varied across the country with the highest in the North of the country, the North West, Midlands and the South West Peninsula and this was seen each year during the study (Figure 3.2).

The crude incidence rate of PT21/28 was 2.5 per million person years and for PT8 it was 3.3 per million person years. There was a distinct seasonality both in rural and urban areas with rates comparable during the winter but higher in rural areas during the summer (Figure 3.4.). The rate of infection declined from 2012, particularly for PT21/28 infections in rural areas (Figure 3.3).

The spatial distribution of animals varied across the country (Figure 3.4 g-i). The mean cattle density ranged from 0-199 animals/km^2 with the highest densities observed in the South West Peninsula, areas of the North West (Cheshire) and Midlands (Staffordshire) and in the North. Sheep density ranged from 0-572 animals/km^2 with the highest densities observed along the Welsh Borders, Oxfordshire, the South West Peninsula and in the North. Pig density ranged from 0-499 animals/km^2 with the highest densities observed in East Anglia and the North East.

Figure 3.1 Flow diagram showing case selection process

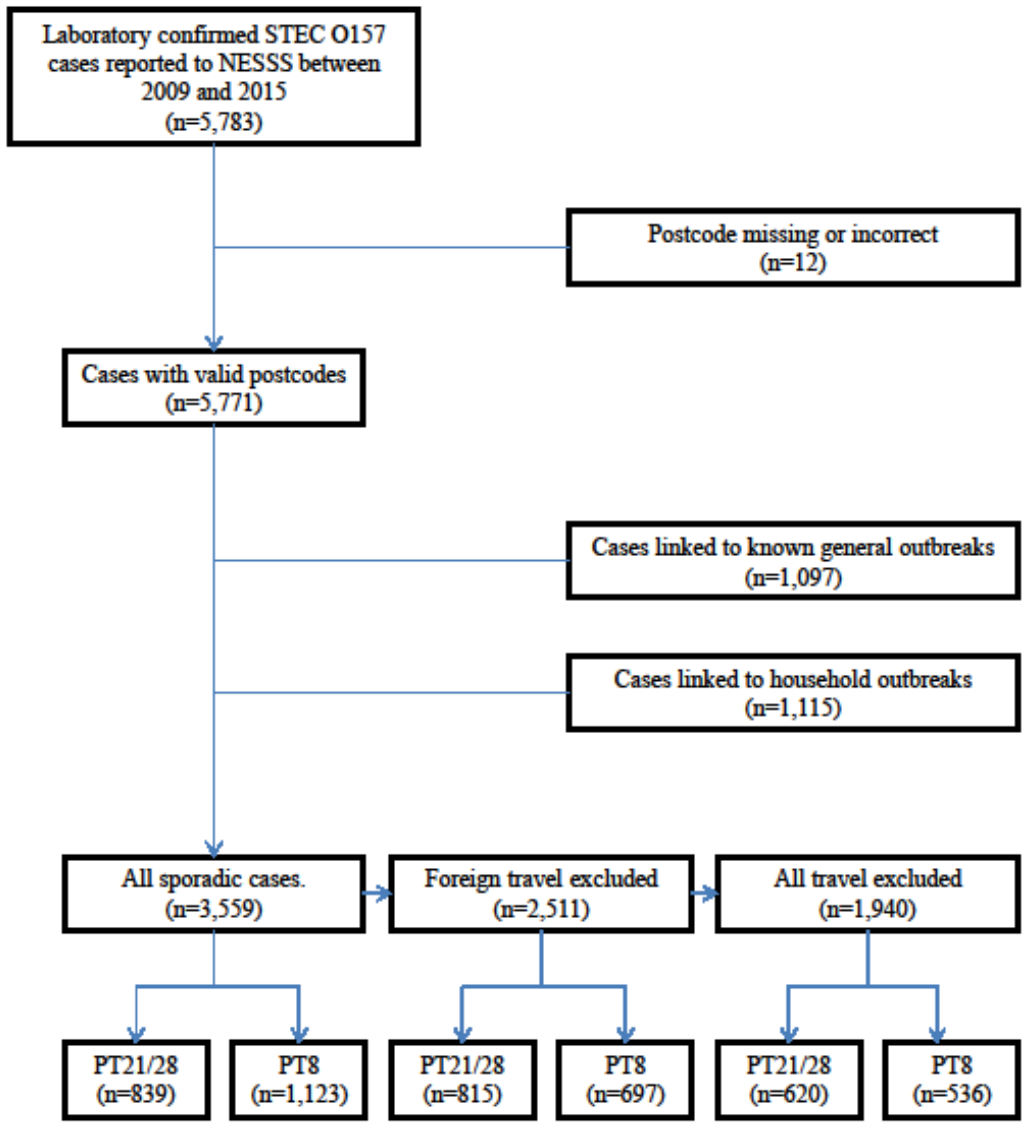


Figure 3.2 Annual incidence rate of STEC O157 per million population including cases reporting travel outside the UK in England between 2009-2015. (Unit of analysis is a local authority area)

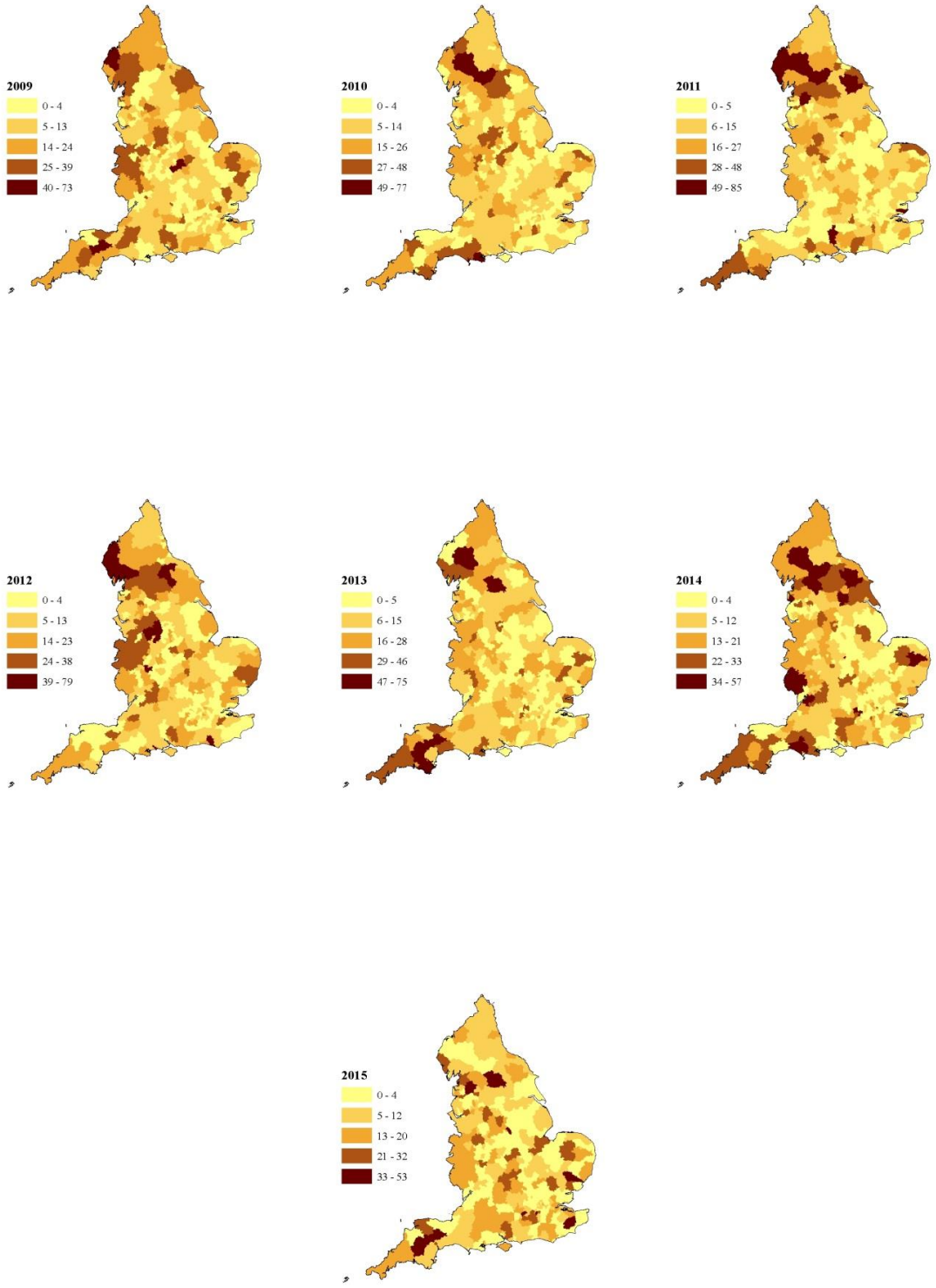


Figure 3.3 Monthly rate of sporadic STEC O157 infection per million population in urban and rural settings in England between 2009 and 2015

(Travel included)

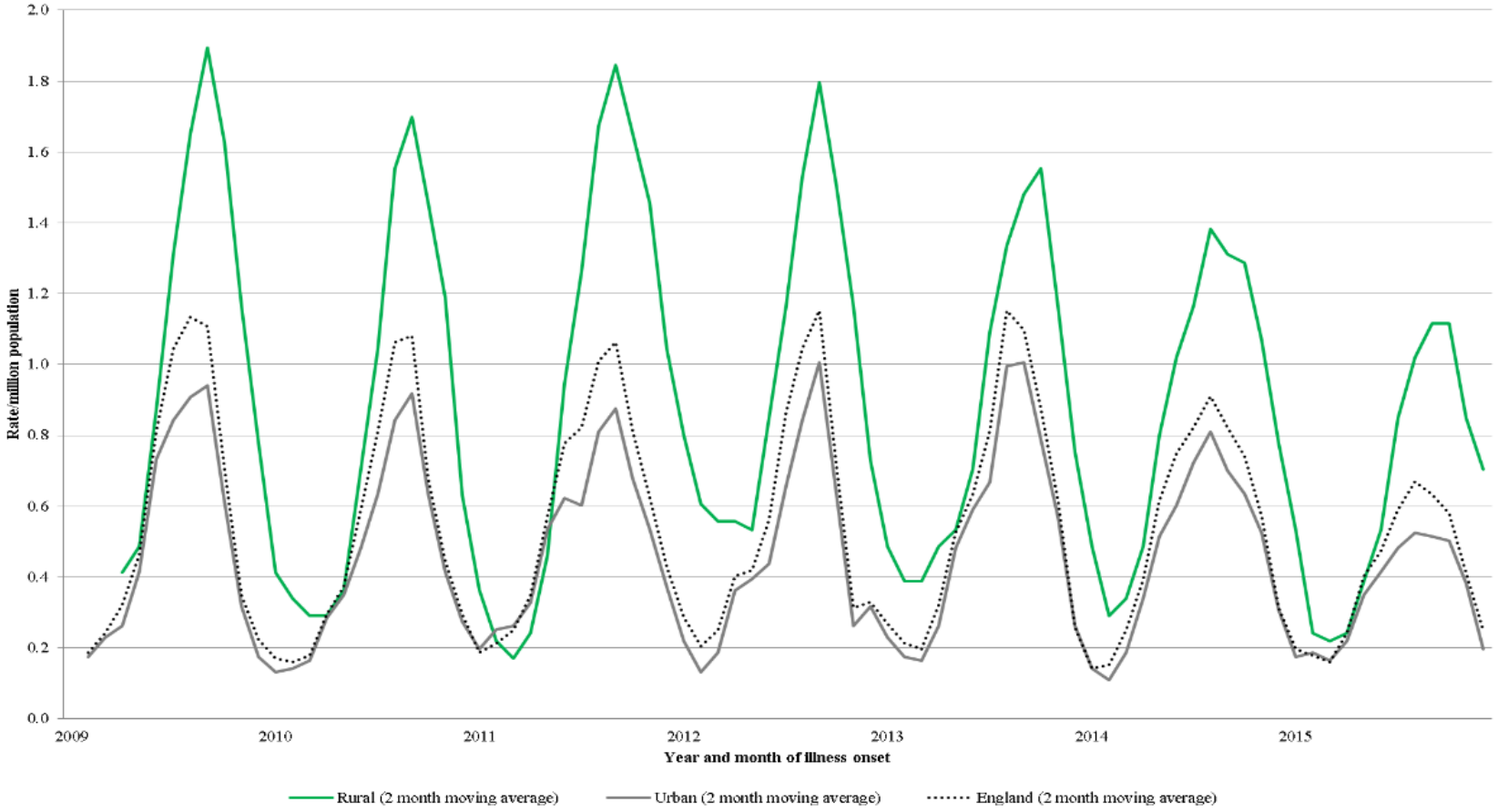


Figure 3.4 Monthly rate of sporadic STEC O157 PT21/28 infection per million population in urban and rural settings in England between 2009 and 2015 (Travel included)

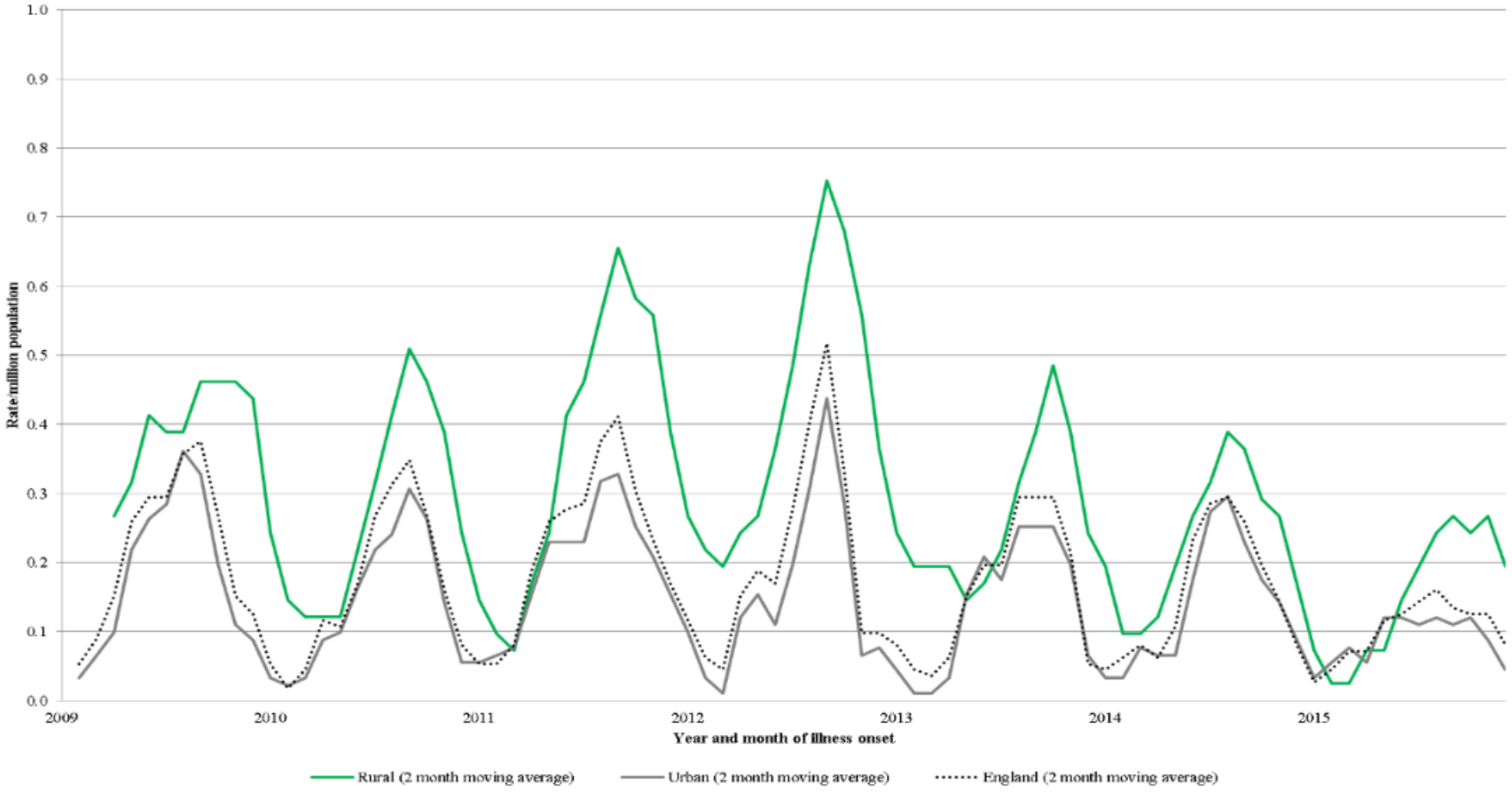


Figure 3.5 Monthly rate of sporadic STEC O157 PT8 infection per million population in urban and rural settings in England between 2009 and 2015 (Travel included)

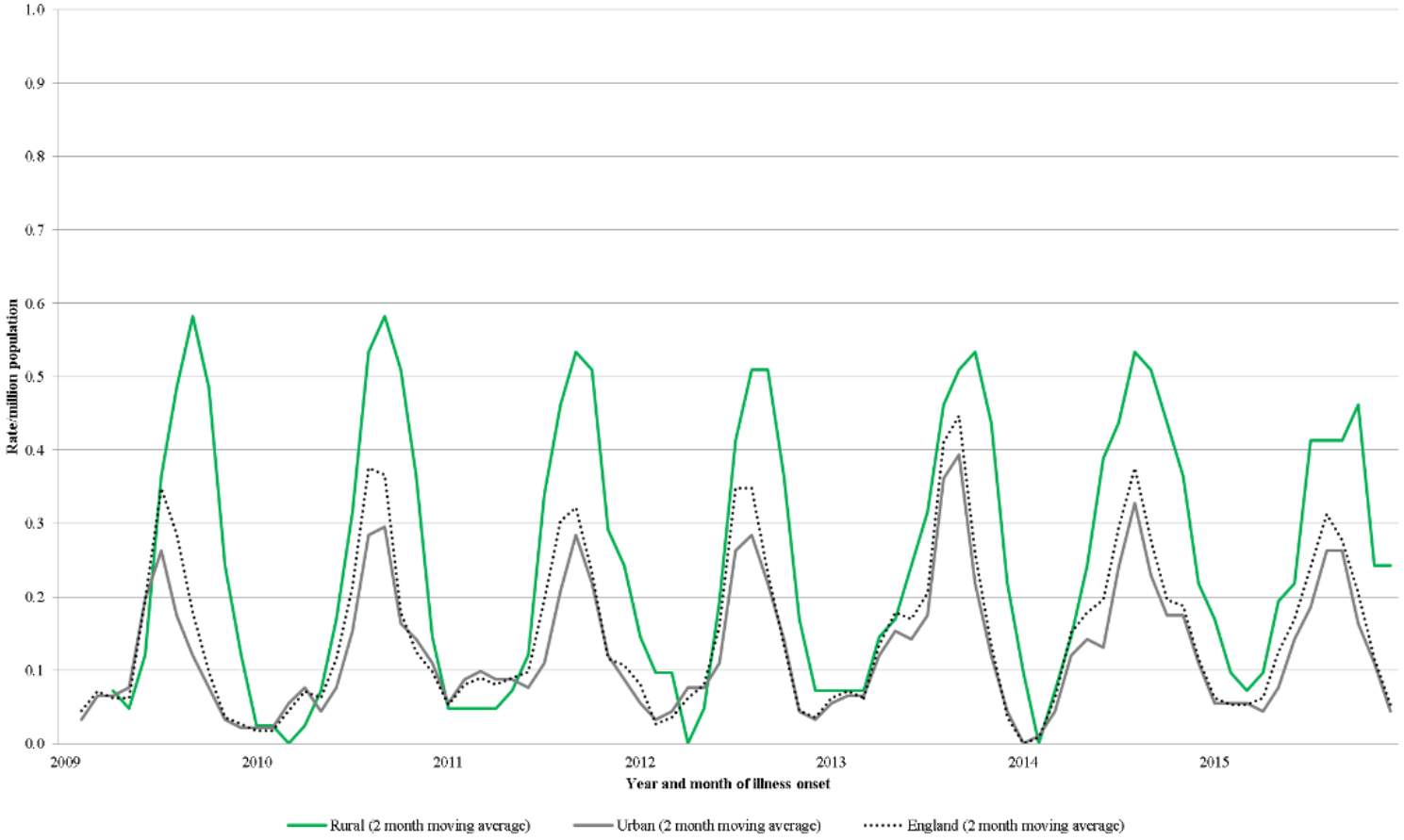
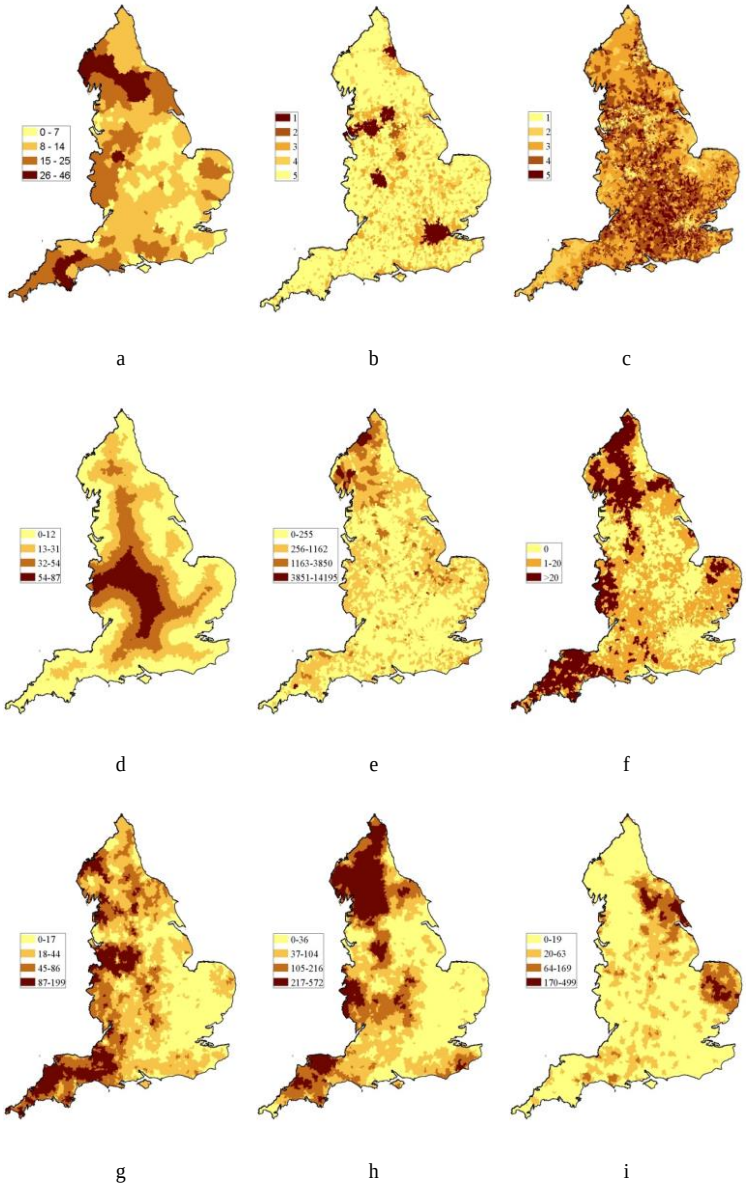


Figure 3.6 (a) Cumulative incidence rate (sporadic cases/million person years) and spatial distribution of the eight independent variables used in the analysis (b) Residence (1: urban–major conurbation, 2: urban–minor conurbation, 3: urban–city and town, 4: rural–town and fringe, 5: rural-village. (c) IMD (quintiles). (d) Distance from LSOA centroid to GB coast in kilometres. (e) Proportion of inland freshwater coverage of each LSOA (km^2). (f) Numbers of PWS in each LSOA. (g) Cattle density (animals/km^2). (h) Sheep density (animals/km^2). (i) Pig density (animals/km^2)



Multivariable analysis

The results of the multivariable analysis for all sporadic cases are presented in Table 3.1.

These indicate that living in a rural village, in an area with high densities of farmed animals (cattle, sheep or pigs), the presence of PWS and those areas considered least deprived were positively associated with risk for all sporadic cases. Removing cases reporting foreign travel removed the effect seen in the IMD variable.

The dataset was then split into PT21/28 and PT8. The results for PT21/28 are presented in Table 3.2 and indicate that living in an area with high densities of farmed animals and being served by PWS were positively associated with risk. Living in a rural village was a risk factor only for cases who reported no travel. Areas regarded as the most deprived were associated with increased risk of PT 21/28 infection.

The results of the multivariable analysis for PT8 are presented in Table 3.3 and indicate that living in a rural village and areas with high densities of cattle and/or pigs and areas considered least deprived (Quintiles 4&5) were positively associated with risk. Removing cases reporting foreign travel from the dependent variable increased the risk effect for cattle and pig density but removed the effect of deprivation. PWS and sheep density were not significant predictors of infection for PT8.

Table 3.1 Results of multivariable analysis of all STEC

		All (n = 3,559)			Excluding foreign travel (n = 2,511)			Excluding all travel (n = 1,940)		
Variable	**Detail**	**IRR**	**95% CI**	**P**	**IRR**	**95% CI**	**P**	**IRR**	**95% CI**	**P**
Cattle density (Animals/Km^2)	0–17	1	–	<0.001	1	–	<0.001	1	–	<0.001
	18–44	1.25	1.14–1.37		1.34	1.19–1.50		1.28	1.12–1.45	
	45–86	1.33	1.16–1.51		1.52	1.30–1.77		1.49	1.25–1.78	
	87–199	1.84	1.58–2.14		2.24	1.88–2.67		2.24	1.83–2.74	
Sheep density (Animals/Km^2)	0–36	1	–	<0.001	1	–	<0.001	1	–	<0.001
	37–104	1.09	0.98–1.20		1.08	0.96–1.21		1.13	0.98–1.29	
	105–216	1.20	1.05–1.38		1.25	1.07–1.46		1.26	1.05–1.51	
	217–572	1.56	1.30–1.88		1.63	1.32–2.01		1.71	1.34–2.17	
Pig density (Animals/Km^2)	0–19	1	–	0.005	1	–	<0.001	1	–	0.015
	20–63	1.03	0.93–1.14		1.10	0.98–1.23		1.08	0.94–1.23	
	64–169	1.29	1.11–1.51		1.43	1.20–1.71		1.40	1.15–1.71	
	170–499	1.35	0.94–1.93		1.59	1.07–2.35		1.46	0.91–2.32	
Number of PWS	0	1	–	0.003	1	–	0.002	1	–	0.019
	1–20	1.19	1.06–1.34		1.22	1.07–1.41		1.18	1.01–1.39	
	>20	1.35	1.09–1.68		1.48	1.16–1.88		1.44	1.09–1.89	
Residence	Urban: Major conurbation	1	–	<0.001	1	–	<0.001	1	–	<0.001
	Urban: Minor conurbation	1.23	1.02–1.49		1.18	0.93–1.49		1.33	1.04–1.71	
	Urban: City and town	1.06	0.97–1.16		1.04	0.94–1.17		1.01	0.89–1.14	
	Rural: Town and Fringe	1.12	0.98–1.28		1.16	0.99–1.37		1.17	0.98–1.41	
	Rural: Village	1.43	1.23–1.66		1.54	1.29–1.84		1.61	1.31–1.97	
IMD	1 (Most deprived)	1	–	<0.001	1	–	0.965	1	–	0.082
	2	1.04	0.92–1.17		1.01	0.88–1.15		0.90	0.77–1.04	
	3	1.12	1.00–1.26		1.03	0.90–1.18		0.88	0.75–1.02	
	4	1.13	1.01–1.27		0.99	0.86–1.13		0.80	0.69–0.94	
	5(Least deprived)	1.27	1.14–1.43		1.02	0.89–1.17		0.85	0.73–1.00	

IRR, incidence rate ratio; IMD, index of multiple deprivation; CI, confidence interval; PWS, private water supplies

Table 3.2 Results of multivariable analysis of STEC PT21/28

		All (n = 839)			Excluding foreign travel (n = 815)			Excluding all travel (n = 620)		
Variable	**Detail**	**IRR**	**95% CI**	***P***	**IRR**	**95% CI**	***P***	**IRR**	**95% CI**	***P***
Cattle density (animals/km^2)	0–17	1	–	<0.001	1	–	<0.001	1	–	0.002
	18–44	1.42	1.17–1.73		1.43	1.17–1.74		1.34	1.07–1.69	
	45–86	1.36	1.04–1.77		1.35	1.03–1.77		1.25	0.91–1.71	
	87–199	1.98	1.46–2.68		2.01	1.48–2.73		1.93	1.36–2.76	
Sheep density (animals/km^2)	0–36	1	–	<0.001	1	–	<0.001	1	–	<0.001
	37–104	1.19	0.97–1.46		1.21	0.98–1.50		1.27	1.00–1.62	
	105–216	1.83	1.42–2.38		1.85	1.43–2.41		2.04	1.50–2.77	
	217–572	2.48	1.77–3.46		2.49	1.78–3.50		2.75	1.85–4.09	
Pig density (animals/km^2)	0–19	1	–	0.029	1	–	0.021	1	–	0.034
	20–63	1.10	0.90–1.34		1.12	0.92–1.37		1.02	0.80–1.30	
	64–169	1.38	1.01–1.88		1.39	1.02–1.90		1.60	1.13–2.25	
	170–499	2.02	1.11–3.69		2.09	1.15–3.83		1.71	0.79–3.73	
Number of PWS	0	1	–	0.002	1	–	0.002	1	–	0.056
	1–20	1.35	1.07–1.70		1.34	1.06–1.69		1.10	0.83–1.46	
	>20	1.94	1.33–2.82		1.94	1.32–2.84		1.70	1.09–2.64	
Residence	Urban: major conurbation	1	–	0.212	1	–	0.193	1	–	0.018
	Urban: minor conurbation	1.23	0.83–1.83		1.23	0.82–1.84		1.34	0.87–2.05	
	Urban: city and town	1.00	0.83–1.21		1.00	0.83–1.22		0.92	0.74–1.15	
	Rural: town and fringe	1.09	0.83–1.44		1.11	0.84–1.47		1.09	0.79–1.51	
	Rural: village	1.35	0.99–1.83		1.37	1.01–1.87		1.52	1.07–2.18	
IMD	1 (most deprived)	1	–	0.256	1	–	0.167	1	–	0.002
	2	0.91	0.73–1.14		0.89	0.70–1.11		0.75	0.58–0.96	
	3	0.79	0.62–0.99		0.76	0.60–0.97		0.65	0.50–0.84	
	4	0.80	0.63–1.01		0.77	0.61–0.98		0.60	0.46–0.78	
	5 (least deprived)	0.85	0.68–1.08		0.83	0.66–1.06		0.67	0.51–0.87	

IRR, incidence rate ratio; IMD, index of multiple deprivation; CI, confidence interval

Table 3.3 Results of multivariable analysis of STEC PT8

Variable	Detail	All (n = 1123) IRR	95% CI	P	Excluding foreign travel (n = 697) IRR	95% CI	P	Excluding all travel (n = 536) IRR	95% CI	P
Cattle density (animals/km^2)	0–17	1	–	<0.001	1	–	<0.001	1	–	<0.001
	18–44	1.35	1.15–1.59		1.44	1.17–1.77		1.39	1.10–1.77	
	45–86	1.55	1.24–1.94		1.85	1.40–2.45		1.95	1.42–2.68	
	87–199	1.90	1.45–2.49		2.30	1.66–3.19		2.24	1.54–3.27	
Sheep density (animals/km^2)	0–36	1	–	0.790	1	–	0.420	1	–	0.076
	37–104	1.06	0.89–1.26		1.05	0.85–1.30		1.11	0.87–1.42	
	105–216	1.01	0.79–1.29		0.83	0.61–1.13		0.73	0.51–1.06	
	217–572	1.16	0.82–1.64		1.05	0.70–1.60		1.19	0.76–1.88	
Pig density (animals/km^2)	0–19	1	–	0.021	1	–	0.013	1	–	0.038
	20–63	1.12	0.95–1.33		1.20	0.98–1.48		1.25	0.99–1.58	
	64–169	1.46	1.13–1.88		1.58	1.16–2.14		1.54	1.09–2.19	
	170–499	1.42	0.77–2.62		1.54	0.75–3.16		1.54	0.67–3.53	
Number of PWS	0	1	–	0.932	1	–	0.467	1	–	0.372
	1–20	1.00	0.81–1.24		1.06	0.82–1.37		1.07	0.79–1.44	
	>20	1.08	0.72–1.61		1.32	0.85–2.07		1.42	0.87–2.34	
Residence	Urban: major conurbation	1	–	0.011	1	–	0.001	1	–	0.003
	Urban: minor conurbation	1.20	0.85–1.68		1.17	0.75–1.82		1.51	0.96–2.37	
	Urban: city and town	1.06	0.90–1.24		1.10	0.89–1.36		1.07	0.84–1.36	
	Rural: town and fringe	1.16	0.92–1.47		1.20	0.89–1.63		1.28	0.91–1.80	
	Rural: village	1.58	1.21–2.05		1.92	1.39–2.65		1.94	1.34–2.81	
IMD	1 (most deprived)	1	–	0.001	1	–	0.246	1	–	0.764
	2	1.17	0.94–1.45		1.09	0.83–1.43		1.04	0.77–1.41	
	3	1.32	1.07–1.64		1.31	1.01–1.71		1.19	0.89–1.60	
	4	1.38	1.12–1.70		1.23	0.94–1.61		1.07	0.79–1.44	
	5 (least deprived)	1.52	1.24–1.88		1.26	0.96–1.64		1.05	0.77–1.42	

IRR, incidence rate ratio; IMD, index of multiple deprivation; CI, confidence interval

3.5 Discussion

Risk was positively associated with cattle density across all models. The risk of a case occurring in areas with 87 animals/km^2 or more was more than twice that of area with fewer than 18 animals/km^2. This finding was somewhat expected as cattle are regarded as the main reservoir of STEC O157 (194) and the most common subtypes shed in cattle faeces are PT21/28 followed by PT8 (196).

Sheep density was positively associated with risk for all STEC O157 cases and PT21/28 cases but not for PT8. The greatest effect of sheep density was seen in the PT21/28 model where the risk was increased almost threefold in areas with high densities of sheep. There is increasing evidence that sheep and other small ruminants are an important reservoir of STEC, and that sheep density is associated with non O157 STEC infections (64). The association with PT21/28 is interesting because the carriage of PT21/28 in sheep is low (14%) compared to cattle (37%) (203) yet they appear strongly significant in our model. This exposure to sheep and lambs has been linked to at least two PT21/28 outbreaks at petting farms/lambing events in England and an extended outbreak of closely related strains was linked to an ovine source (204). A recent study in the Republic of Ireland demonstrated a geographical association between sheep density and STEC O26 but not STEC O157 (101). Disentangling the relative contribution of ruminant species to the overall burden of infection is difficult due to scant contemporary information on shedding by sheep compared to cattle, a lack of genetic difference between cattle and sheep strains (109) and the fact that sheep and cattle are farmed in the same geographical areas in the UK. However, our results suggest that the role of sheep as a reservoir and potential source of infection in humans should not be overlooked.

Pig density was positively associated with risk across all models. However, for the PT8 and PT21/28 models, the observed effect was not linear. Pigs can shed STEC (205), and

pork products have been implicated as the source or vehicle of infection in outbreaks worldwide, but they are not generally considered to be an important reservoir for STEC O157 (206) or source for human infection (58). Studies of intestinal carriage in England showed a low carriage rate in pigs and that the characteristics of pig strains differed from those seen in humans during the same period (196, 203, 207). Pig density was not identified as a risk factor for STEC infection in the Netherlands (59). Despite this, associations with pig density appear in this study. This dichotomy may be partly explained by the few areas of the country with high pig densities or may be due to differences in pig husbandry practices between countries.

We found an increased risk associated with the presence of PWS for all STEC O157 cases, PT21/28 cases but not PT8. Private water supplies that do not meet the requirements of the EC directive present a high risk of infection with STEC(57), however, the reason for the difference between PTs is unclear and may relate to factors not considered by this study.

Living in a rural village was a risk factor for all STEC O157 cases and for PT8. For the PT21/28 model, living outside a major urban conurbation was a significant risk factor but only for cases reporting no travel. Residents of rural areas are more likely to come into contact with contaminated environments either through work or leisure (185) and residential proximity to the ruminant reservoir also increases the possibility of exposure from wildlife and insect vectors (208, 209).

Living in less deprived areas was strongly associated with all STEC O157 cases and PT8. What is intriguing was that when foreign travel cases were removed, this effect was removed for all STEC O157 cases and PT8. This indicates that the deprivation effect is strongly driven by foreign travel and that risk factors for these groups differ from indigenously acquired infection. The strong association with foreign travel in the PT8 model was anticipated as a greater proportion of cases of PT8 report travel abroad

compared to other PTs (1). Lower deprivation was protective for cases of PT21/28 reporting no travel but the reasons for this are unclear. One explanation could be related to deprivation and/or likelihood of exposure to PT21/28 compared to other strains. PT21/28 is a strain indigenous to the UK and rates of infection are highest in the north of England where there are also more areas considered to be the most deprived. However, crude rates of infection with STEC are lower in the most deprived areas, cases are less likely to travel abroad or within the UK (110) and levels of social interaction differ from residents of less deprived areas (111). Socioeconomic status has been shown to be associated with risk for other gastrointestinal infections (210-212) introducing the possibility that whilst risk factors may differ broadly across the country, within regions, socio-economic status has a greater influence on risk factors and transmission dynamics. This is an area that would benefit from further research.

In developed countries, disadvantaged children, but not adults, appear to be at greater risk of gastrointestinal infection and that living in deprived areas is protective for infectious intestinal disease (IID) overall but is associated with more severe symptoms in those who do become infected (213). Our findings indicate that once the effect of foreign travel is removed, deprivation has little effect on sporadic infection with STEC O157. This suggests that infection is a result of a localised and stochastic process driven by exposure to the local environment and that exposures related to affluence, such as diet, occupation or leisure pursuits are likely to be less important.

Residential distance from the coast and living in an area with a high proportion of freshwater coverage were not significant and removed at an early stage of the modelling process. These environments may act as reservoirs for STEC (214) and recreational exposure to fresh water has been suggested as a risk factor for STEC infection in epidemiological studies from other countries. The lack of an association in the UK may

relate to the general unpopularity of freshwater swimming in the UK in comparison to other countries (67, 153, 215-217). These variables were proxies for exposure and so do not capture details of individual interactions with these environments.

There are several potential limitations to our study. First, molecular typing methods, used routinely in England since 2015, are superior to the phenotypic methods we used to discriminate between cases and have been shown to reduce the number of cases considered to be sporadic (16). It is therefore possible that a small number of cases included in our study were microbiologically linked and therefore may not be considered truly sporadic. Secondly, we excluded cases linked to household outbreaks. We made this decision based on the difficulty in identifying the primary case amongst co-primary and asymptomatic (microbiologically confirmed) cases generated by microbiological screening of household contacts. In addition, we noted epidemiological links between foodborne outbreaks and secondary transmission within households which may have introduced a bias away from the exposures of interest. Thirdly, for every STEC O157 infection reported to national surveillance systems in England, there are an estimated 7.4 in the community (29). The reasons for this are likely to be related to severity of disease, health seeking behaviours and whether or not a clinician takes a sample and requests a microbiological examination from a laboratory. Notwithstanding that the ratio of STEC O157 reports is considerably smaller than other pathogens, the cases reported to national surveillance represent a biased sample of the true community burden of STEC O157 in England. Transmission routes are varied, and infection is a result of complex interactions between people and their local environment. Our approach meant that individually reported exposures could not be considered in our analysis. Finally, this is an ecological study and association does not equal causation which could only be inferred from other study designs involving an intervention. We are confident that the association with animal density is most likely due

to environmental exposure, however, other factors not included in our study (for example, locally sourced food) cannot be ruled out as a potential route of infection.

In conclusion, using arguably one of most comprehensive enhanced surveillance of STEC datasets in the world, we found that two thirds of infections were sporadic, and that the spatial and temporal distribution of these cases showed distinct variation within England. We provide evidence that living in a rural area with high densities of farmed animals and served by private water supplies partly explain this variation. Our results indicate that travel abroad may expose individuals to risks not present in their local residential environment and that this risk is influenced by socio-economic status i.e., the ability to afford foreign travel. Further analysis is required to elucidate the relative importance of exposures reported by individual cases including travel, contact with animals and the agricultural environment and consumption of food and water.

To reduce the overall burden of infection in England, interventions designed to reduce the number of sporadic infections with STEC O157 should focus on the residents of rural areas with high densities of livestock and the effective management of non-municipal water supplies.

4 Using spatial relative risk to identify modifiable risk factors for STEC O157 infection in England

4.1 Abstract

Evidence from outbreak investigations suggests a shift in importance from food borne to animal contact transmission since STEC was first identified in the UK and increased risk is associated with areas that have high densities of ruminant animals. Most control measures designed to prevent transmission of STEC O157 in England are focused on the food chain, do not account for geographical differences in incidence and are not always tailored to specific groups.

We used the geographical location of likely exposure to known risk factors for STEC infection to explore whether reports of 'risky' behaviours differed between cases identified in England between 2009 and 2018. Using kernel density estimation, we identified areas of the country where the risk of infection was significantly higher or lower. Using the residential postcode, we assigned each case to one of these areas to create a binary residential risk variable (either high or low risk). We also identified cases who had travelled within the UK (including travel to high-risk areas) during their incubation period.

Using logistic regression, we then explored the relationship between reported exposure to the agricultural environment, place of residence and place of exposure for those who travelled.

Adjusting for age sex and socio-economic status, increased odds of a case occurring in a high risk area were associated with direct exposure to the agricultural environment (aOR 1.52, 95% CI 1.27-1.82) contact with dogs (aOR 1.25 , 95% CI 1.07-1.45) or drinking water from a PWS (aOR 2.14 , 95% CI 1.52-3.02) although the proportion of cases exposed to PWS was less than 10%. Decreased odds were associated with reported UK travel (aOR 0.56, 95% CI 0.47-0.67).

Increased odds of a case occurring in the UK travel group were associated with going on a daytrip (aOR 4.94, 95% CI 4.00-6.11, direct environmental contact (aOR 1.40, 95% CI 1.12-1.76), indirect environmental contact (aOR 2.10, 95% CI 1.70-2.59) contact with domestic animals (aOR 1.22 95% CI 1.0-1.48) and drinking from a PWS (aOR 2.47 95% CI 1.49-4.09).

The population attributable risk (PAR%) for residents of high-risk areas were similar (direct contact; 14%, indirect contact;12%, contact with dogs;12%) whereas for travellers, indirect contact was greatest (44%) followed by daytrips (37%) and direct contact (26%).

We recommend that, where appropriate, residential risk (as opposed to residential location) should be considered when developing new public health policy or revising existing guidance related to STEC O157 infection. The greatest risk reductions would be gained by targeting people travelling in the UK with focused communications highlighting the risks of STEC infection and, most importantly, providing simple guidance designed to reduce risk.

4.2 Introduction

Infection with STEC O157 occurs via multiple routes including contaminated food, person to person or via direct or indirect contact with farm animals. The risk of infection with Shiga- toxin producing *E. coli* (STEC) O157 varies significantly across England. The highest risk areas are in the North and Southwest (65) and significantly lower rates of infection are found in the South and South-East. Remaining cases are widely distributed across areas where the spatial distribution of cases does not differ significantly from the underlying population. This pattern is geographically consistent from year to year suggesting that infection rates are driven by the presence of risk factors in those areas (14, 65).

Most control measures designed to prevent transmission of STEC O157 in England are focused on the food chain and are universal, i.e., they do not account for geographical differences in incidence. These control measures have largely been introduced as a result of large foodborne outbreaks (32, 45, 218) but since these were introduced, STEC case numbers have remained fairly stable (11, 27) suggesting either that they are having little effect or other routes of transmission may be of importance.

Evidence from outbreak investigations suggests a shift in importance from food borne to animal contact transmission since STEC was first identified in the UK. Increased risk is associated with areas that have high densities of ruminant animals (14) and contact with the agricultural environment has been identified as a major risk factor for infection in England (44) and Scotland (181), but not Wales (219). However, direct contact with farm animals is reported infrequently and many cases report no contact with the farming or rural environment at all (1). Travel outside the UK is also risk factor, particularly for Lineage II strains (1) and this appears to be associated with cases living in more affluent areas of England (65). Cases living in areas considered high risk are less likely to report travel outside the UK (65) and although travel within the UK is a potential risk factor (15, 44), this has not been explored in detail.

Since 2009, all laboratory confirmed cases of STEC O157 in England have been interviewed using a standardised questionnaire (1) to collect epidemiological (age, sex, occupation) and clinical (duration of illness, hospitalisation) information, as well as details of travel, foods eaten and exposure to animals and/or the environment in the seven-day period preceding illness onset. This information is matched with laboratory typing information to create a rich source of information on this important pathogen.

We used this information to examine the relationship between exposures reported by cases living in areas of England deemed high risk in the 7 days preceding their illness onset in

comparison to those living in areas of low risk. It is based upon the location of the case during the median incubation period of 4 days, (inter-quartile range 3-7 days) (5).

The anticipated impact of this study is to identify modifiable risk factors that can be used to develop or refine geographically specific public messaging or interventions.

4.3 Methods

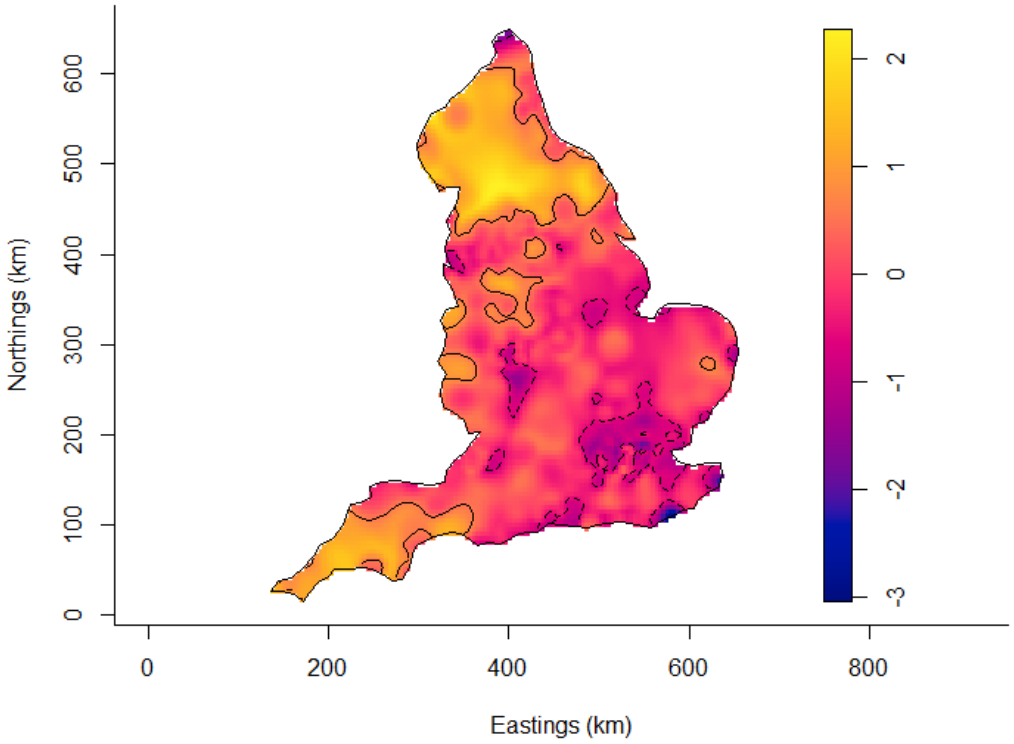
We selected primary cases of STEC O157 reported to the National Enhanced STEC Surveillance System (NESSS) between the 1st January 2009 and the 31st December 2018. All cases were provided with a spatial location based upon their residential postcode.

Questionnaires were administered by staff in local Health Protection Teams (HPT) who take the case through the questionnaire recording responses throughout. The questionnaire is divided into sections designed to capture the following details: case classification (primary, co-primary, secondary), personal details (age, sex, occupation), symptoms, travel history, food (including water) history and exposure to water, animals, and the environment. Questions about food, water and environmental exposures are binary. If the case answers "yes" to a question, then further details are requested. For example, if a case states that they travelled within the UK, details about destinations, accommodation and daytrips are requested and recorded. A copy of the questionnaire is provided in the supplementary material.

Dependent variables

To create dependent variables, we first identified the geographical areas of England where risk was significantly higher or lower (Figure 4.1).

Figure 4.1 Log relative risk surface for primary cases of STEC O157 in England between 2009 and 2018. Tolerance contours are superimposed at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk higher risk and dashed lines indicate areas of lower risk.

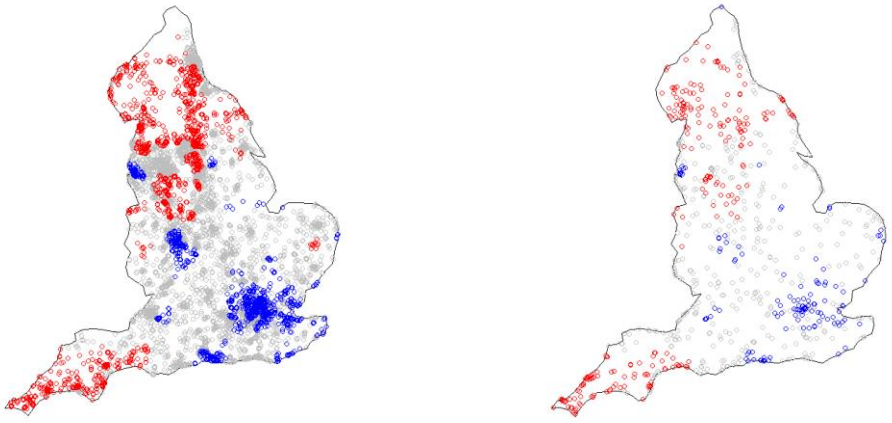


To do this we used the kernel smoothing method described in (165) and (65) to estimate the relative risk of STEC O157 between 2009 and 2018. To provide the best estimate of indigenous risk, we excluded cases who had travelled abroad on the basis that their place of exposure may not reflect their place of residence. Cases from known outbreaks were also excluded to avoid introducing bias from large numbers of cases associated with national and regional outbreaks. Once the relative risk surfaces were produced, we then used the tolerance contours from this indigenous risk estimate to categorise the residential risk for all cases (including those that reported travel and were part of an outbreak). Cases who fell into areas of significantly increased risk were categorised as high and those who fell into areas of significantly decreased risk (or where risk did not differ significantly) were categorised as low (Figure 4.2). Using this classification, we created a dichotomous dependent variable indicating whether or not a case lived in a high-risk area (1) or a low-risk area (0).

Independent variables

Independent variables were created by collapsing reported exposures from the standardised questionnaire into five categorical variables: direct exposure to the farming environment, indirect exposure to the farming environment, contact with domestic animals, contact with dogs and consumption of water from private water supplies or unpasteurised milk. Full details of the exposures and how they were grouped are provided in Table 4.1. We included exposure to unpasteurised milk or water from private water supplies *a priori* because these are known risk factors for STEC infection and PWS have a similar geographical distribution to cases (14). Contact with dogs was considered separately from other domestic animals because unlike other pets, they often require regular outdoor exercise which may directly or indirectly expose their owners to the agricultural environment.

Figure 4.2 Maps of England showing risk categorisation of all cases for place of residence (left panel) and UK travel destination (right panel). Red = significantly increased risk, Blue = significantly decreased risk, Grey = risk not significantly different to underlying population



Details of individual socio-economic status are not collected as part of the NESSS. To provide a proxy for socio-economic status, each case was assigned a deprivation measure, the Index of Multiple Deprivation 2010 (IMD), based upon the lower super output area (LSOA) that contained their residential postcode. LSOAs are the smallest geographical area at which IMD data is available, each with a population of around 1,500 people. The IMD was divided into quintiles, with the first quintile being the most deprived and the fifth quintile the least deprived LSOA.

Within the NESSS, UK travel is defined as at least one night away from home. Cases reporting travel within the UK were given an additional spatial reference: their UK travel location. The level of detail provided by cases about their movement within the UK was variable. The best spatial reference was the postcode of the accommodation. Where this

was unavailable, we used the description of the destination (village, town or city) to retrieve a spatial location from Streetmap (available from www.openstreetmap.org).

If only a region (county) was reported, the polygon centroid was used. Where the destination was ambiguous (e.g., where the same town name occurred in two or more regions), we checked the free text detail to confirm the location. If this was unavailable, the location was not used but the case was included as having travelled in the UK.

Table 4.1 Construction of independent variables

Variable	Exposure Detail
Direct contact with environment	Case visited a farm (including petting farms) Case reported contact with cattle, sheep, goats, pigs, poultry or horses. Case handled or bottle-fed animals at a farm.
Indirect contact with environment	Case walked across a field grazed by animals. Case reported recreational contact with fresh or sea water. Case handled soil or animal manure.
Private water consumption	Case drank from a private water supply.
Unpasteurised milk consumption	Case drank unpasteurised milk.
Contact with domestic animals	Case reported contact with cats, rabbits, reptiles, rodents etc.
Contact with dogs	Case reported contact with dogs.

We explored univariate relationships between the dependent and independent variables. Exposures reported by at least 20% of cases or controls and with a p value of 0.1 or less were included in multivariate logistic regression models. Raw drinking milk and PWS consumption were included as *a priori* risk factors. Each model was adjusted for age, sex, and socio-economic status. Each model was run including age in years as a continuous covariate and then stratified by age group (adult/child) with age in years included as a continuous covariate.

To check for collinearity, we created a correlation matrix before fitting each model using a coefficient value of 0.7 or above to indicate strong correlation. A postestimation check for collinearity was also performed after each model run using a variance inflation factor (VIF) value of 10 or more to indicate the presence of multicollinearity.

To investigate whether the strength of any relationship was moderated by the inclusion of another variable, we tested the effect of adding interaction terms. These terms were selected where there were two or more significant main effects in the multivariable model and included only if the presence of an interaction was biologically plausible.

Using adjusted ORs and the fraction of cases exposed, the population attributable risk percentage (PAR%) was determined for each factor that increased risk in the final models. The concept of population attributable risk or fraction is to estimate how much disease burden might be reduced if an exposure were eliminated.

All geocoding and spatial joins were performed using Arc GIS version 10.5.1(176). Spatially varying risk was estimated using the sparr package (166) in R (179) and all remaining analyses were conducted in STATA v.15 (202).

4.4 Results

A total of 6,792 cases were eligible for analysis. Of these, UK travel destinations were mapped to 1,276 cases using postcode for 58% (n= 746), destination description for 35% (n= 446). For 7% (n=84).

For the remaining analyses, cases reporting travel outside the UK (22%, n=1,525) or known to be part of an outbreak (23%, n= 1,290) were excluded. Of the remaining 4,157 cases, just over half (51%, n=2,125) were adults (aged 18 years or over) and 53% of these (n=2,189) were female. The majority (69%, n=2878) described themselves as white British and the proportion of cases living in each IMD quintile varied from 16% (n= 671) in the most deprived to 21% (n= 873) in the least deprived quintile.

Just over one fifth (22%) of cases (n= 897) reported travelling within the UK. Of these, 30% (266) travelled to an area considered 'high risk'.

Less than a third (28%, 1147) of these cases lived in an area considered high risk. The majority (61%, 2529) lived outside these areas in areas that were not significantly different and 12% (481) lived in areas considered to present a significantly lower risk.

Risk factors associated with residence in high-risk areas compared to low-risk areas.

The results for this analysis are presented in Table 4.2. Adjusting for age sex and socio-economic status, increased odds of a case occurring in a high risk area were associated with direct exposure to the agricultural environment (aOR 1.52, 95% CI 1.27-1.82) contact with dogs (aOR 1.25 , 95% CI 1.07-1.45) or drinking water from a PWS (aOR 2.14 , 95% CI 1.52-3.02). The proportion of cases exposed to PWS was less than 10%. Decreased odds were associated with reported UK travel (aOR 0.56, 95% CI 0.47-0.67).

Table 4.2 Adjusted and unadjusted multivariable logistic regression analysis comparing STEC O157 cases living in high-risk areas with those living in medium and low risk areas

	Exposure	Cases (n=1147)		Controls (n=3010)		OR*	95% CI		p	aOR**	95% CI		p
		Exposed	%	Exposed	%		lower	upper			lower	upper	
All	Travelled in UK	196	17.09	701	23.29	0.57	0.47	0.68	0.000	0.56	0.47	0.67	0.000
	Direct environmental contact	388	33.83	701	23.29	1.53	1.28	1.83	0.000	1.52	1.27	1.82	0.000
	Indirect environmental contact	516	44.99	1126	37.41	1.16	0.98	1.36	0.081	1.13	0.96	1.34	0.139
	Contact with domestic animals	433	37.75	1054	35.02	0.95	0.81	1.10	0.490	0.95	0.81	1.10	0.496
	Contact with dogs	469	40.89	985	32.72	1.25	1.07	1.45	0.004	1.25	1.07	1.45	0.004
	Drank unpasteurised milk	32	2.79	43	1.43	1.40	0.87	2.28	0.167	1.39	0.86	2.25	0.181
	Drank from a private water supply	74	6.45	76	2.52	2.14	1.52	3.02	0.000	2.14	1.52	3.02	0.000
	Exposure	Cases (n=558)		Controls (n=1567)		OR	95% CI		p	aOR**	95% CI		p
		Exposed	%	Exposed	%		lower	upper			lower	upper	
Adults	Travelled in UK	96	17.2	359	22.91	0.62	0.48	0.80	0.000	0.62	0.48	0.81	0.000
	Direct environmental contact	140	25.09	281	17.93	1.37	1.06	1.79	0.017	1.39	1.07	1.80	0.015
	Indirect environmental contact	217	38.89	497	31.72	1.24	0.99	1.56	0.060	1.21	0.96	1.52	0.104
	Contact with dogs	226	40.5	540	34.46	1.16	0.95	1.43	0.153	1.18	0.96	1.45	0.125
	Drank unpasteurised milk	9	1.61	23	1.47	0.91	0.41	2.01	0.819	0.88	0.40	1.94	0.747
	Drank from a private water supply	28	5.02	46	2.94	1.50	0.91	2.48	0.113	1.46	0.88	2.41	0.143
	Exposure	Cases (n=589)		Controls (n=1443)		OR	95% CI		p	aOR**	95% CI		
		Exposed	%	Exposed	%		lower	upper			lower	upper	p
Children	Travelled in UK	100	16.98	342	23.7	0.53	0.40	0.69	0.000	0.53	0.41	0.70	0.000
	Direct environmental contact	248	42.11	420	29.11	1.62	1.27	2.08	0.000	1.56	1.22	2.00	0.000
	Indirect environmental contact	299	50.76	629	43.59	1.03	0.81	1.30	0.838	1.01	0.79	1.28	0.967
	Contact with domestic animals	255	43.29	529	36.66	1.05	0.85	1.31	0.643	1.07	0.86	1.34	0.529
	Contact with dogs	243	41.26	445	30.84	1.33	1.07	1.65	0.010	1.39	1.11	1.73	0.004
	Drank unpasteurised milk	23	3.9	20	1.39	1.82	0.95	3.47	0.069	1.72	0.90	3.29	0.100
	Drank from a private water supply	46	7.81	30	2.08	2.97	1.81	4.87	0.000	2.99	1.82	4.92	0.000

* Unadjusted ** Adjusted for age, sex and IMD quintile

This effect was seen across all strata for direct environmental contact (Adults aOR 1.39, 95% CI 1.07-1.80; Children aOR 1.56, 95% CI 1.22-2.0) and UK travel (Adults aOR 0.62, 95% CI 0.48-0.81 ; Children aOR 0.53, 95% CI 0.41-0.70). Contact with dogs and consumption of water from a PWS remained only for children (Dogs aOR 1.39, 95% CI 1.11-1.73: PWS aOR 2.99, 95% CI 1.82-4.92).

Adding an interaction term between direct environmental contact and contact with dogs was significant for all cases and children (All 1.39 95% CI 1.01-1.90: Children 1.84 95% CI 1.18-2.86)). The main effect of direct environmental exposure alone remained for all cases but not children. The main effect for dogs alone was lost for both groups.

Risk factors for residents of low-risk areas travelling in the UK.

The results for this analysis are presented in Table 4.3. Accounting for age, sex, and socio-economic status, increased odds were associated with going on a daytrip (aOR 4.94, 95% CI 4.00-6.11 , direct environmental contact (aOR 1.40, 95% CI 1.12-1.76), indirect environmental contact (aOR 2.10, 95% CI 1.70-2.59) contact with domestic animals (aOR 1.22 95% CI 1.0-1.48) and drinking from a PWS (aOR 2.47 95% CI 1.49-4.09).

Daytrips, indirect environmental contact and drinking from a PWS remained significant for both age strata but direct contact remained only for children. (Adults: daytrip aOR 5.56, 95% CI 4.08-7.56), indirect contact aOR 1.73 95% CI 1.29-2.31, PWS aOR 2.40, 95% CI 1.25-4.63. Children: daytrip aOR 3.70 95% CI 2.85-4.80, direct contact aOR 1.79, 95% CI 1.37-2.3, indirect contact aOR 2.95 95% CI 2.24-3.88, PWS aOR 2.30, 95% CI 1.10-4.81).Interaction terms between daytrips and indirect contact with the environment or contact with domestic animals were significant for all cases (Indirect 0.62 95% CI 0.40-0.95 Domestic animals 0.66 95% CI 0.44-1.0). The main effects of direct and indirect environmental exposure remained. The inclusion of the interaction terms increased the

main effect of daytrips from an aOR of 4.9 to an aOR of 8. No interaction terms were significant for adults (Daytrip indirect env 0.55 95%CI 0 .3-1.01) or children (daytrip direct 1.15 .62-2.12, daytrip indirect 0.55 95% CI 0.28-1.09.

Travel to high-risk areas amongst residents of low-risk areas

The results for this analysis are presented in Table 4.4. Accounting for age, sex, and socio-economic status, the odds of indirect exposure to the environment was increased amongst UK travellers to high-risk areas (aOR 1.71, 95% CI 1.21-2.40). A summary table of these results is provided in Table 4.5

Attributable risk.

Population attributable risk percentages (PAR%) are shown in Table 4.6. For residents of high-risk areas, direct exposure had the highest attributable risk (13.7%), not dissimilar to indirect exposure or dogs (both 12.1%). However, for children the PAR for direct environmental exposure was twice that of adults (18.3% vs 8.7%) and higher for contact with dogs (15.1% vs. 9.2%).

For those who travelled in the UK, the highest PAR% was for indirect environmental exposure (43.5%), followed by going on a daytrip (37%) and direct environmental exposure (26%). The PAR for indirect exposure amongst travellers was threefold that of residents (43.5% vs. 12.1%) and almost double for direct environmental exposure (25.5% vs 13.7%). For children, the highest PAR% was for indirect contact which was more than double that of adults (61% vs. 21.3)

For those travelling to high-risk areas, the PAR% for indirect environmental exposure was 52.5%. For children, the PAR % was 71.5% compared to 37.1% for adults.

Table 4.3 Adjusted and unadjusted multivariable regression analysis for cases living outside high-risk areas and UK travel

	Exposure	Cases (n=701)		Controls (n=2309)		OR	95% CI		p	aOR	95% CI		p
		Exposed	%	Exposed	%		lower	upper			lower	upper	
All	Went on a daytrip	307	43.79	248	10.74	4.83	3.92	5.97	0.000	4.94	4.00	6.11	0.000
	Direct environmental contact	275	39.23	426	18.45	1.41	1.13	1.76	0.003	1.40	1.12	1.76	0.003
	Indirect environmental contact	425	60.63	701	30.36	2.17	1.76	2.67	0.000	2.10	1.70	2.59	0.000
	Contact with domestic animals	319	45.51	735	31.83	1.21	0.99	1.47	0.062	1.22	1.00	1.48	0.048
	Contact with dogs	262	37.38	723	31.31	0.86	0.70	1.05	0.145	0.85	0.70	1.04	0.117
	Drank unpasteurised milk	13	1.85	30	1.3	0.97	0.46	2.04	0.941	0.94	0.44	1.97	0.864
	Drank from a private water supply	39	5.56	37	1.6	2.55	1.54	4.23	0.000	2.47	1.49	4.09	0.000
	Exposure	**Cases (n=359)**		**Controls (n=1208)**		**OR**	**95% CI**		**p**	**aOR**	**95% CI**		**p**
		Exposed	**%**	**Exposed**	**%**		**lower**	**upper**			**lower**	**upper**	
Adults	Went on a daytrip	129	35.93	100	8.28	5.52	4.07	7.50	0.000	5.56	4.08	7.56	0.000
	Direct environmental contact	98	27.3	183	15.15	1.27	0.91	1.77	0.167	1.22	0.87	1.71	0.253
	Indirect environmental contact	171	47.63	326	26.99	1.81	1.36	2.41	0.000	1.73	1.29	2.31	0.000
	Contact with domestic animals	148	41.23	377	31.21	1.21	0.93	1.59	0.156	1.21	0.92	1.59	0.164
	Contact with dogs	137	38.16	403	33.36	0.87	0.66	1.14	0.306	0.88	0.67	1.16	0.357
	Drank unpasteurised milk	5	1.39	18	1.49	0.73	0.24	2.18	0.571	0.74	0.25	2.26	0.602
	Drank from a private water supply	22	6.13	24	1.99	2.38	1.23	4.59	0.010	2.40	1.25	4.63	0.009
	Exposure	**Cases (n=342)**		**Controls (n=1101)**		**OR**	**95% CI**		**p**	**aOR**	**95% CI**		**p**
		Exposed	**%**	**Exposed**	**%**		**lower**	**upper**			**lower**	**upper**	
Children	Went on a daytrip	178	52.05	148	13.44	3.62	2.80	4.68	0.000	3.70	2.85	4.80	0.000
	Direct environmental contact	177	51.75	243	22.07	1.73	1.33	2.25	0.000	1.79	1.37	2.33	0.000
	Indirect environmental contact	254	74.27	375	34.06	2.94	2.24	3.85	0.000	2.95	2.24	3.88	0.000
	Contact with domestic animals	171	50	358	32.52	1.07	0.84	1.38	0.571	1.05	0.82	1.35	0.708
	Contact with dogs	125	36.55	320	29.06	0.86	0.66	1.11	0.232	0.78	0.60	1.01	0.061
	Drank unpasteurised milk	8	2.34	12	1.09	1.65	0.74	3.69	0.220	1.56	0.70	3.51	0.280
	Drank from a private water supply	17	4.97	13	1.18	2.52	1.21	5.24	0.013	2.30	1.10	4.81	0.026

* - Unadjusted ** Adjusted for age, sex and IMD quintile

Table 4.4 Adjusted and unadjusted multivariable regression analysis for cases travelling to high-risk areas

	Exposure	Cases (n=266)		Controls (n=631)		OR	95% CI		p	aOR	95% CI		p
		Exposed	%	Exposed	%		lower	upper			lower	upper	
All	Went on a daytrip	127	47.74	265	42	1.11	0.82	1.50	0.494	1.12	0.83	1.51	0.475
	Direct environmental contact	132	49.62	246	38.99	1.22	0.89	1.68	0.218	1.25	0.91	1.73	0.170
	Indirect environmental contact	191	71.8	365	57.84	1.64	1.17	2.29	0.004	1.70	1.21	2.40	0.002
	Drank unpasteurised milk	9	3.38	10	1.58	1.81	0.71	4.59	0.211	1.86	0.73	4.72	0.192
	Drank from a private water supply	21	7.89	33	5.23	1.31	0.73	2.34	0.363	1.28	0.71	2.30	0.405
	Exposure	**Cases (n=130)**		**Controls (n=325)**		**OR**	**95% CI**		**p**	**aOR**	**95% CI**		**p**
		Exposed	**%**	**Exposed**	**%**		**lower**	**upper**			**lower**	**upper**	
Adults	Indirect environmental contact	76	58.46	150	46.15	1.71	1.12	2.61	0.013	1.78	1.16	2.73	0.009
	Contact with domestic animals	50	38.46	137	42.15	0.83	0.55	1.27	0.397	0.84	0.54	1.28	0.414
	Contact with dogs	42	32.31	135	41.54	0.61	0.39	0.94	0.026	0.61	0.39	0.95	0.028
	Drank unpasteurised milk	1	0.77	5	1.54	0.51	0.06	4.40	0.537	0.53	0.06	4.66	0.569
	Drank from a private water supply	12	9.23	20	6.15	1.51	0.70	3.26	0.299	1.51	0.69	3.29	0.298
	Exposure	**Cases (n=136)**		**Controls (n=306)**		**OR**	**95% CI**		**p**	**aOR**	**95% CI**		**p**
		Exposed	**%**	**Exposed**	**%**		**lower**	**upper**			**lower**	**upper**	
Children	Went on a daytrip	82	60.29	149	48.69	1.35	0.88	2.08	0.174	1.35	0.87	2.07	0.177
	Direct environmental contact	90	66.18	149	48.69	1.55	0.98	2.45	0.059	1.55	0.98	2.47	0.062
	Indirect environmental contact	115	84.56	215	70.26	1.74	0.99	3.06	0.053	1.76	1.00	3.09	0.051
	Drank unpasteurised milk	8	5.88	5	1.63	2.92	0.91	9.36	0.072	3.00	0.93	9.69	0.066
	Drank from a private water supply	9	6.62	13	4.25	1.26	0.51	3.11	0.615	1.25	0.51	3.09	0.626

* - Unadjusted ** Adjusted for age, sex and IMD quintile

Table 4.5 Adjusted odds ratios and significance levels for all models

Exposure	High risk residence			Travelled in UK			Travelled to high-risk area		
	All	Adults	Children	All	Adults	Children	All	Adults	Children
Travelled in UK	0.56***	0.62***	0.53***	–	–	–	–	–	–
Went on a daytrip	–	–	–	4.94***	5.56***	3.7***	1.12	–	1.35
Direct environmental contact	1.52***	1.39*	1.56***	1.4**	1.22	1.79***	1.25	–	1.55
Indirect environmental contact	1.13	1.21	1.01	2.1**	1.73***	2.95***	1.7**	1.78**	1.76*
Contact with domestic animals	0.95	–	1.07	1.22*	1.21	1.05	–	0.84	–
Contact with dogs	1.25**	1.18	1.39***	0.85	0.88	0.78	–	0.61*	–
Drank unpasteurised milk†	1.39	0.88	1.72	0.94	0.74	1.56	1.86	0.53	3.00
Drank from a private water supply†	2.14***	1.46	2.99***	2.47***	2.4**	2.3*	1.28	1.51	1.25

Significance levels: * <0.05, ** <0.01, *** <0.001 †: <10% cases or controls exposed

Table 4.6 Population attributable risk (PAR%) of exposures significantly associated with high-risk residence and travel within the UK

Exposure	High risk residence			Travelled in UK			Travelled to high-risk area in UK		
	All	Adults	Children	All	Adults	Children	All	Adults	Children
Went on a daytrip	-	-	-	37.0	30.2	44.6	-	-	-
Direct environmental contact	13.7	8.7	18.3	25.5	14.3	38.1	-	-	-
Indirect environmental contact	-	-	-	43.5	28.3	61.0	52.5	37.1	71.5
Contact with dogs	12.1	9.2	15.1	-	-	-	-	-	-
Drank from a private water supply	4.0	2.1	5.9	4.0	4.2	3.8	-	-	-

4.5 Discussion

Our approach of combining behaviours and geographical risk allows us to describe the risk gradient between individuals living in different areas of the country. We were also able to examine the effect of travel within the UK in the days preceding illness onset.

Previous studies comparing risk factors between those living in rural and urban areas (1, 110) do not consider that risk is associated with rural areas with high densities of farmed animals and other risk factors, not just rurality alone. Studies incorporating geography and ecological risk factors (14, 210, 220) provide information on the likelihood of exposure to known risk factors and the nature of the population at risk but don't account for variation in people's behaviour within that environment.

The importance of environmental exposures overall, and indirect exposure in particular, were greater for those cases who travelled in the UK than for those living in high-risk areas or non-travellers. Jones et al (221) investigated awareness of STEC O157 amongst residents of, and visitors to, areas of high disease incidence in Scotland and Wales. Awareness and understanding of STEC O157 amongst visitors to high incidence areas was poor; most had never heard of STEC and if they had, thought it more likely to be foodborne and not picked up from animals or the wider environment. Visitors are likely to be immunologically naive to pathogens more common in rural areas [20] and this, combined with a lack of awareness means that less likely to adjust their behaviour in relation to the risk posed by livestock or the environment. This has obvious implications for the probability and outcome of infection.

This is important when considering that around a quarter of holidays in the UK are described as countryside breaks (Statista) and about six million people visit rural Wales and five million visit Scotland's countryside each year [21]. These destinations provide opportunities to explore rural and coastal environments. Animal based attractions are

popular destinations for daytrips for families with children. Walking and swimming are also reasons that people visit these areas. STEC and other GI pathogens can be recovered from the rural environment even in the absence of large numbers of animals (187, 222) and environmental contamination can occur over wide geographical areas via marine and freshwater means, animal movements and manure application. The observed differences in exposures reported by adults and children may reflect differences in the types of activities undertaken by those travelling with children (e.g., direct exposure from petting farms versus indirect exposure through walking).

Direct contact with the agricultural environment and PWS consumption were risk factors both for residents of high-risk areas and travellers. However, the PAR% for environmental exposures for residents in high-risk areas were all less than 15% and much lower than cases who had travelled.

Drinking from a PWS increased risk but accounted for only a small proportion of cases and a relatively low PAR%. According to the UK Drinking Water Inspectorate, around 1% of the population of England and Wales are served by a PWS, the majority of which supply single dwellings. Poorly managed PWS are a known risk factor for a range of infections including STEC O157 (14). These supplies tend to be located in very rural areas and the risk of contamination is likely to be higher in areas with high densities of animals.

The odds of a case occurring in a high-risk area were increased for children when the joint effect of contact with dogs and direct contact with the farming environment was considered. Dogs require regular exercise accompanied by a human and the probability of both encountering farm animals and/or their faeces is greater in areas high risk areas. Dogs also demonstrate a propensity to roll in mud, faeces and animal carcasses resulting in contamination of cars, domestic environments and direct exposure to other family members who come into contact with the dog.

Drinking unpasteurised milk was not a risk factor in any of the models. Although only a small fraction of the UK population drink RDM, consumption has increased from 3% in 2012 to 10% in 2018 and it is now readily available via internet sales and on-farm vending machines. Current advice in England is that RDM should not be consumed by vulnerable groups including children, so it is of concern that children formed the highest proportion of consumers in our study, even though numbers exposed were low.

However, living in a high-risk area or travelling within the UK cannot explain the majority of sporadic cases that occurred during the study period. The spatial distribution of these cases did not differ significantly from the underlying population at risk suggesting that their infections resulted from eating contaminated food, being exposed to an infected individual or possibly as a result of changes in localised environmental contamination due to transient events such as flooding or seasonal animal movements.

Data quality and potential limitations

One potential limitation to our study is that for every STEC O157 infection reported to national surveillance systems in England, there are an estimated 7.4 in the community (29). The reasons for this are likely to be related to severity of disease, health seeking behaviours and whether a clinician takes a sample and requests a microbiological examination from a laboratory. It is unknown whether these reporting biases vary geographically and hence would affect the spatial patterns presented in this paper.

Another potential limitation is that spatial risk does vary from year to year and our analysis was purely spatial. However, the areas described as high risk in our study are consistently associated with increased risk from year to year (65). The reasons for this variation are likely due to multiple factors, including outbreak activity. Our approach controlled for this by excluding cases linked to known outbreaks.

The quality of information collected as part of NESSS can be variable and subject to recall and confirmation bias, particularly for food related exposures. Because we focused on non-foodborne exposures, the potential impact of recall bias could be considered to be lower than that of foods consumed. There is little evidence in the literature that can be used to quantify the true effect of this on our results. One study suggests differences in recall over time regarding consumption and experience (223) but we suggest that the lack of awareness of potential risks associated with animals or the environment described by Jones et al (221) indicates that the effects of confirmation bias are likely to be minimal.

Public messaging and interventions

Infection with STEC is the result of a complex and active interaction between the pathogen, its reservoir, the environment and the host. Modification of any of these factors has the potential to reduce the risk of infection. At a population level, controls for STEC are focused on the food chain. Food producers are legally required to produce food that is safe for consumption and this process is subject to regulation and enforcement. Most STEC controls focus on the safety of nationally distributed food, specific demographic groups (children) or settings (nurseries and petting farms). These controls build upon sector specific legislation and are subject to regulation and enforcement.

Reducing the risk of infection from environmental exposure is much more challenging due to the relatively unpredictable and uncontrolled ways in which individuals may be exposed. For example, public access to farmland and other areas grazed by animals is largely unrestricted in England so walking through cattle faeces may result in exposure to STEC when cleaning shoes afterwards; swimming in the sea or freshwater following heavy rainfall may increase the risk of being exposed to STEC due to agricultural runoff or sewage discharges.

However, there is evidence that awareness-based interventions (including distribution of signage, fliers, and presentations, implementable in different geographical scales) are both cost-effective and have the highest impact on lowering Lyme disease risk compared to attempts to control the reservoir and/or vector (224) and further advice on how and when to raise awareness utilising social media is available (PHE Guidance https://www.gov.uk/government/publications/tick-bite-risks-and-prevention-of-lyme-disease).

Adopting similar strategies to raise awareness of the risks of infection with STEC could also work. Our results help to provide focus for the groups most at risk and inform when and where these would be most effective. For those living in high-risk areas, raising awareness of the risks of direct and indirect contact with the farming environment throughout the year, aimed particularly at families with children, would have the most impact.

However, at a personal level, people need to be aware of the hazard and understand what action they need to take to reduce the risk of infection. Recent assessment of compliance with COVID guidelines (225-228) emphasise the importance of communicating potential risks and providing simple, consistent guidance on how to reduce the spread of the virus in improving compliance with preventive behaviours.

For STEC, awareness of risk varies depending on place of residence (221) and is skewed by the focus on large outbreaks that tend to be food borne and the link to ground beef in the USA. Awareness of zoonotic transmission is low and actions taken to prevent risk tend to be focused on cleanliness, food and the home, not animals, farms or cattle (221).

We recommend that, where appropriate, residential risk (as opposed to residential location) should be considered when developing new public health policy or revising existing

guidance related to STEC O157 infection. The greatest risk reductions would be gained by targeting people travelling in the UK with focused communications highlighting the risks of STEC infection and, most importantly, providing simple guidance designed to reduce risk. For example, advice around environmental exposure could be targeted at families travelling with children and made available when results are searched for on web browsers or bookings are made for accommodation in or close to high-risk areas, or those served by PWS.

5 The spatio-temporal distribution of COVID-19 in England between January and June 2020

5.1 Abstract

The spatio-temporal dynamics of an outbreak provide important insights to help direct public health resources intended to control transmission. They also provide a focus for detailed epidemiological studies and allow the timing and impact of interventions to be assessed.

A common approach is to aggregate case data to administrative regions. Whilst providing a good visual impression of change over space, this method masks spatial variation and assumes that disease risk is constant across space. Risk factors for COVID-19 (e.g. population density, deprivation and ethnicity) vary from place to place across England so it follows that risk will also vary spatially. Kernel density estimation compares the spatial distribution of cases relative to the underlying population, unfettered by arbitrary geographical boundaries, to produce a continuous estimate of spatially varying risk.

Using test results from healthcare settings in England (Pillar 1 of the UK Government testing strategy) and freely available methods and software, we estimated the spatial and spatio-temporal risk of COVID-19 infection across England for the first six months of 2020. Widespread transmission was underway when partial lockdown measures were introduced on the 23rd March 2020 and the greatest risk erred towards large urban areas. The rapid growth phase of the outbreak coincided with multiple introductions to England from the European mainland. The spatio-temporal risk was highly labile throughout.

In terms of controlling transmission, the most important practical application of our results is the accurate identification of areas within regions that may require tailored intervention strategies. We recommend that this approach is absorbed into routine surveillance outputs

in England. Further risk characterisation using widespread community testing (Pillar 2) data is needed as is the increased use of predictive spatial models at fine spatial scales.

5.2 Introduction

On the 31 December 2019, the World Health Organization (WHO) was informed of a cluster of cases of pneumonia of unknown cause detected in Wuhan City, Hubei Province, China. Since the initial identification of SARS-CoV-2 as the cause of COVID-19, over 32 million cases have been diagnosed globally, with more than 900,000 fatalities, as of 27th September 2020 (229). The first laboratory confirmed case in England was reported on the 31st January 2020. A series of interventions designed to slow rates of infection followed, culminating in a partial lockdown announced by the UK Government on the 23rd March 2020.

Understanding the spatiotemporal dynamics of COVID-19 helps to clarify the extent and impact of the pandemic and can aid decision making, planning and community action intended to control transmission (230). It also provides an opportunity to assess the impact of interventions over space and time.

One approach is to describe changes in infection rates within administrative boundaries in England have been published widely. This approach expresses the disease risk per head of population and assumes that risk is constant across space i.e., the risk of disease does not depend upon spatial location. This is rarely the case and the distribution of risk factors for COVID-19 (for example population density, deprivation and ethnicity) are known to vary across England so it follows that absolute risk will also vary spatially.

Another approach is to plot points to produce a spatial point pattern. This is useful for small data sets but as the number of points increases, over plotting makes it difficult to discriminate between the relative densities of points.

Kernel density estimation (KDE), also known as kernel smoothing, is a flexible, non-parametric method by which spatially varying risk may be estimated without the need to aggregate data. Smoothing a spatial point pattern (using an appropriate bandwidth) overcomes the over plotting problem by expressing the number of points as an intensity function. Comparing the intensities of two groups, for example those with an infectious disease and those without, across a defined geographical area results in an intensity (or risk) ratio. If the ratio is ~1, this suggests that the risk of infection is unrelated to spatial location. Evidence of spatial variation in risk occurs where the intensities differ. Ratio values >1 indicate an increased risk and values <1 indicate lower risk.

As the COVID-19 outbreak progresses in England, KDE provides a scalable means to identify areas of significantly higher or lower risk to inform national policy and local action.

Using established methods (165, 167, 168, 180) and freely accessible software (166, 179), we conducted a spatio-temporal point pattern analysis of COVID-19 risk in England between January and June 2020. Our aims were to describe the spatio-temporal dynamics of the first six months of the COVID-19 outbreak and assess the potential use of this method to inform and support public health policy decisions as the outbreak progresses.

5.3 Methods

The method we followed is described in detail by Davies et al. [3] and Elson et al. (65). For the spatial estimates, each set of points (case and control) were smoothed using an adaptive (165, 167) bandwidth to determine the spread of smoothing kernels centred on each point to produce a density surface. Adaptive bandwidths account for greater uncertainty in areas with fewer points (e.g. rural areas) so the bandwidth is large resulting in greater smoothing. In urban areas, more data points mean the bandwidth is smaller resulting in a surface with less smoothing. Calculating the ratio of case and control

densities provides a continuous estimate of relative risk which can be plotted on a map (128, 129, 167).

Case locations

We selected confirmed cases of COVID-19 reported to the PHE Second Generation Surveillance System (SGSS) under Pillar 1 of the UK Government testing strategy between the 31st January 2020 and the 30th June 2020. Pillar 1 includes tests only for those with a medical need (symptomatic and seen by a clinician) but may also include some healthcare workers and samples taken as part of outbreak investigations (231). The data was checked for duplicates and presence of a valid residential postcode. Postcodes are like US Zip codes and represent a single residential street or group of houses.

The statistical methodology for spatial point processes are very sensitive to duplicate data points (178). We used unique control locations but included multiple cases with the same postcode to account for sporadic and outbreak cases.

Population at risk ('control') locations

The underlying population at risk ('controls') was represented by points randomly sampled from the National Population Database (NPD). The NPD is a Geographical Information System (GIS) dataset that combines multiple layers of data (including population) in a 100-metre by 100-metre grid (161, 163). Based on the centroid coordinates of each grid square, 'control' locations were randomly drawn without replacement. The probability of a location being drawn was weighted by the summed population of each grid square to reflect the spatially varying nature of the underlying population at risk. The number of controls was chosen to match the number of cases.

For the spatial estimates, we attempted four bandwidth initialisation methods to set the 'global' and 'pilot' smoothing parameters needed to calculate the adaptive bandwidths

themselves: maximal smoothing (175), bootstrapping (165, 232), least-squares cross validation (LSCV) (233) and likelihood cross-validation (234) (171).

The resulting bandwidths were used to produce density estimates at all locations of a fine grid of co-ordinates laid within a simplified polygon of the mainland boundary of England and the Isle of Wight.

To explore the temporal variation in the spatial risk, we marked each case with the date that their specimen was taken. For cases with multiple test results, the specimen date that gave the most recent positive result was used. We then calculated the number of days that had elapsed from the specimen date of the first confirmed case (31st January 2020) as the temporal event. The spatiotemporal relative risk surface was then calculated using the fixed estimator of Fernando & Hazelton (174).

All estimates are edge-corrected to account for kernel weight lost over the boundary of the study region. This correction reweights each observation-specific kernel and the correction term is the reciprocal of the kernel mass inside the study region (172, 173). For the spatial analyses only, the estimates are calculated as symmetric adaptive risk functions using the pooled case/control data and equal global and pilot bandwidths (168). Unless stated otherwise, results are reported as log-relative risk surfaces.

Contours identifying areas of significantly higher risk were superimposed at the 1% significance level for the spatial estimates and 1% and 0.01% levels for the spatio-temporal estimates. Wherever temporal results are referred to in terms of weeks, this refers to the corresponding International Organization for Standardization (ISO) week.

All analyses were performed using the contributed packages sparr (166) and spatstat (177, 178) in the R language (179).

5.4 Results

Between the 31st January and the 30th June 2020, 160,976 cases of COVID-19 were reported to PHE under Pillar 1 of the UK Government testing strategy (231). Of these, residential postcodes were available for 154,210 (96%). Of these, multiple cases were recorded at 44,989 (30%) postcode locations.

Bandwidth selection

The oversmoothed and bootstrap methods produced usable spatial and space-time bandwidths. The LSCV approach did not provide a result and the likelihood-based approach produced a very small bandwidth that resulted in an under smoothed, ‘spiky’ surface.

Spatial risk

Figure 5.1 shows the areas in England classified as urban by the Office for National Statistics (ONS) (199). The relative risk across England during the study period is presented in Figure 5.2. With some exceptions, the areas with the highest risk tended to be large urban areas.

Spatio-temporal risk

An animation of the spatio-temporal analysis combining an epidemic curve with the risk surfaces using the oversmoothed and bootstrap estimators is provided in the supplementary materials (SM 4).

The 14-day space-time slices are presented in Figures 5.3 and 5.4 for the oversmoothed and bootstrap estimators respectively.

Fewer than twenty cases were recorded between the confirmation of the first case (Week 5 commencing January 27th) and the end of Week 8 (February 23rd). Weeks 9 -10 were characterised by a greater geographical spread of small areas of elevated risk and an increase in case numbers to ~400 by the end of Week 10 (March 8th). Week 11 (commencing 9th March) saw a rapid increase in case numbers and significantly elevated risk, particularly in the cities of London and Birmingham. By the end of Week 14 (5th April), 47,668 cases had been reported. From Week 15 (commencing 6th April) onwards, areas of significantly elevated risk became more dispersed with some areas in the North and far South East of the country experiencing sustained periods of elevated risk, even as case numbers declined towards the end of the study period.

Of note is the generalised increase in risk across the country between Weeks 13-19 and the abrupt change in risk seen in London between Weeks 13 and 15.

Figure 5.1 Geographical distribution of urban areas in England

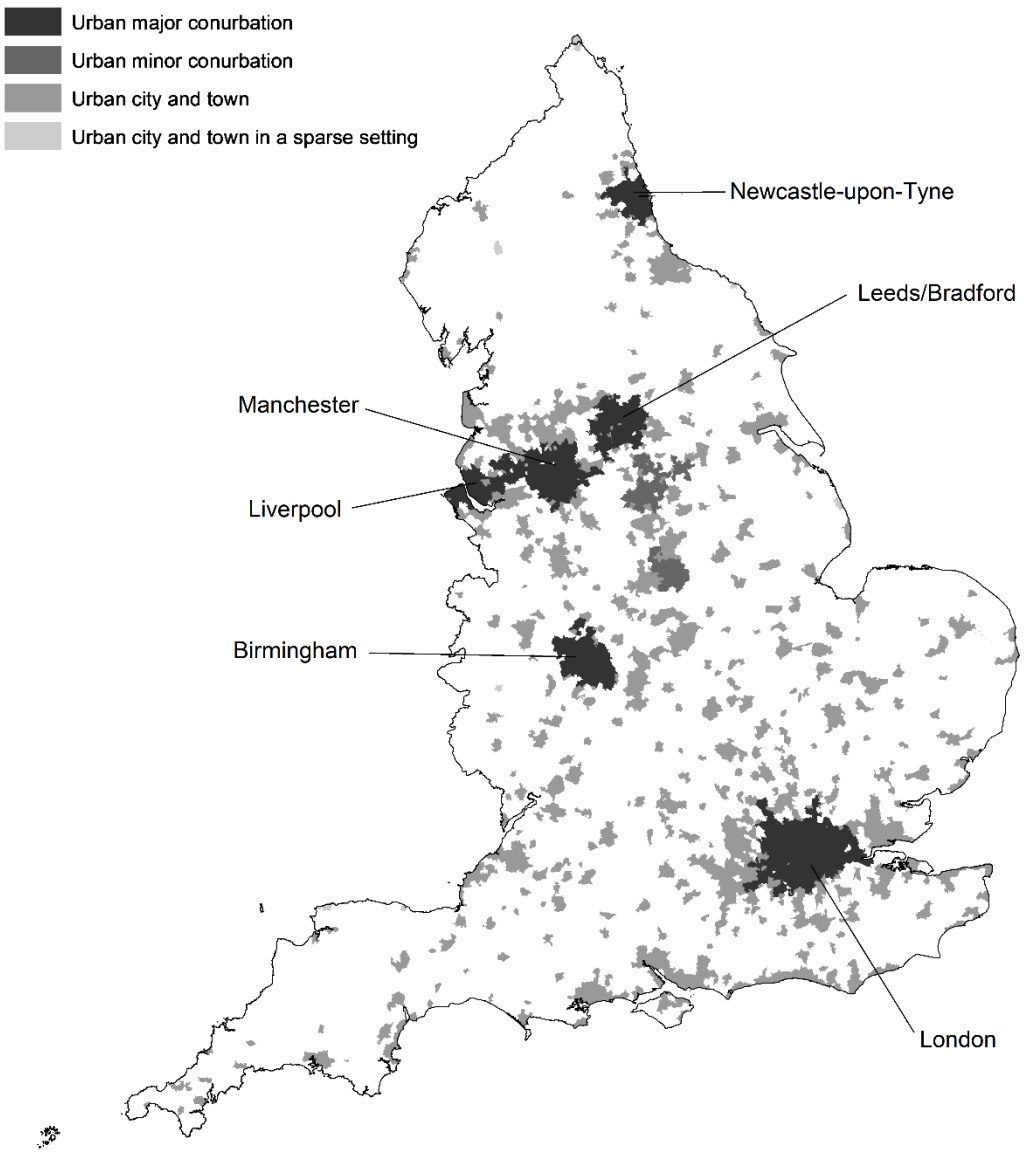


Figure 5.2 Log relative risk estimates for COVID-19 in England between January and June 2020 using different bandwidths: oversmoothed (left), likelihood cross validation (centre) and bootstrapping (right). Tolerance contours indicating areas of significantly higher risk are superimposed as solid lines at the 1% confidence level

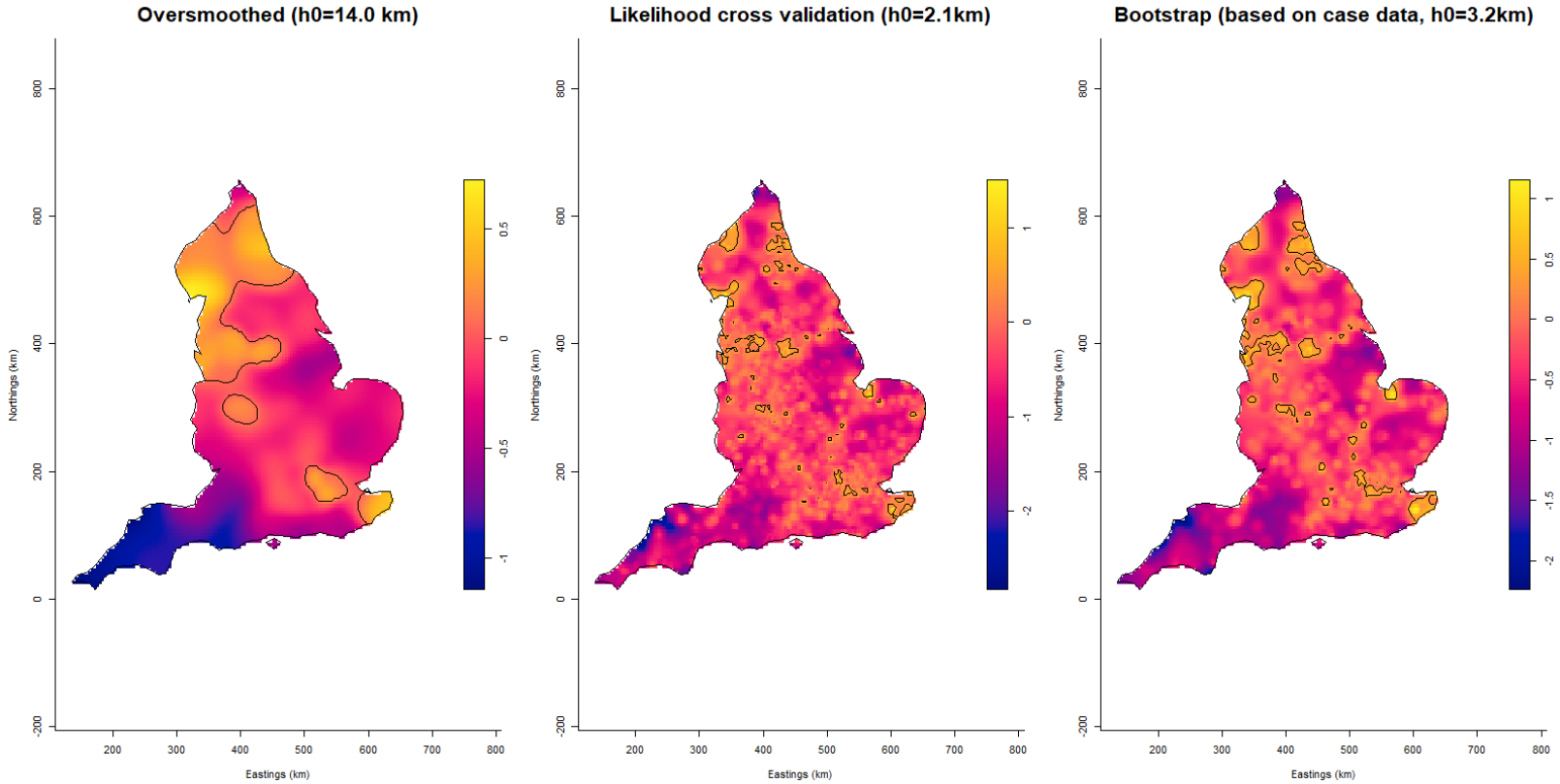


Figure 5.3 Log relative risk spacetime slices using an oversmoothed bandwidth (*h*=15.4km, *lambda (λ)*=2.04) in 14-day periods from the date of the first case confirmed in ISO Week 5. Solid lines outline areas of significantly higher risk at the 1% confidence level

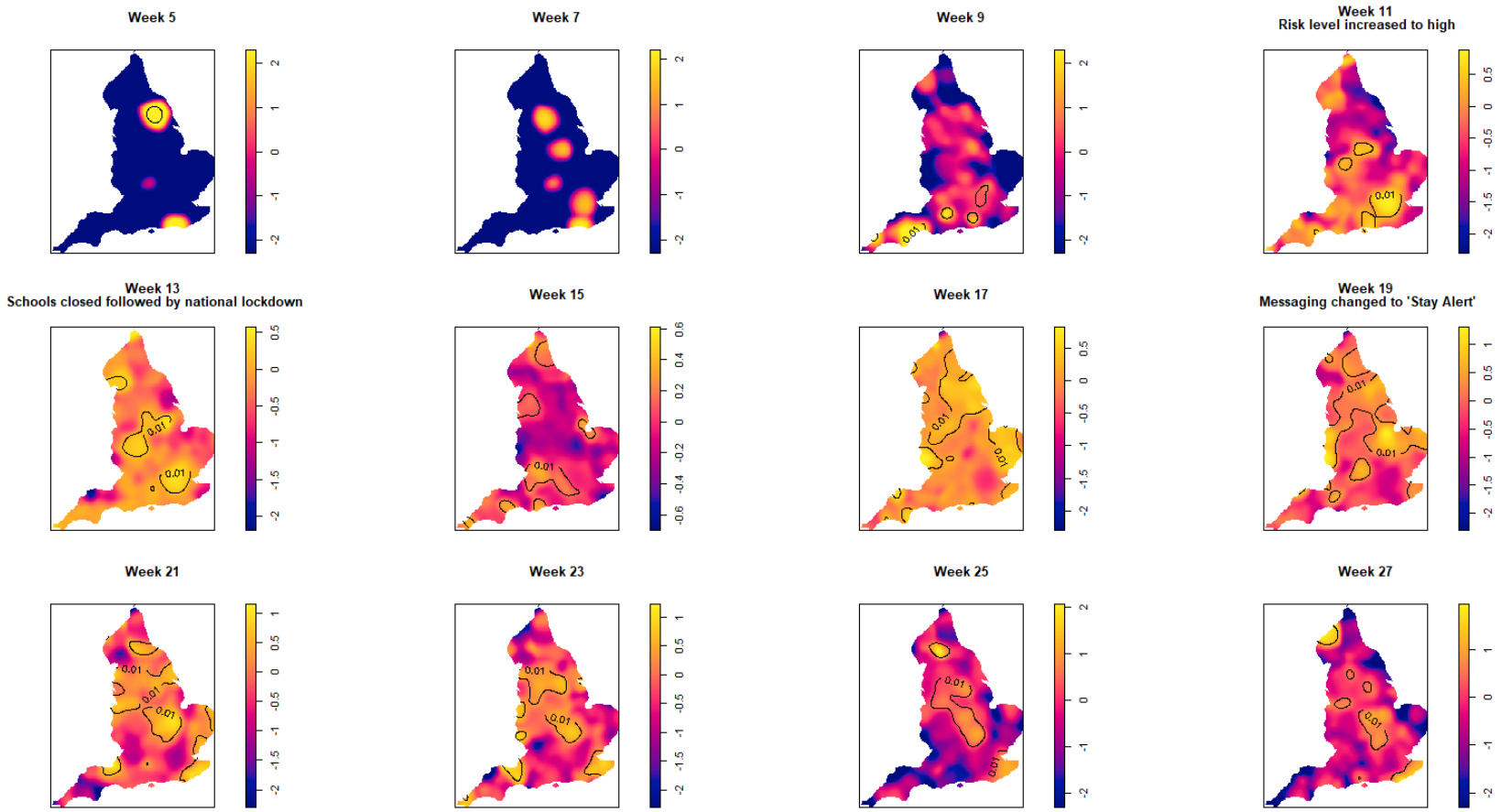
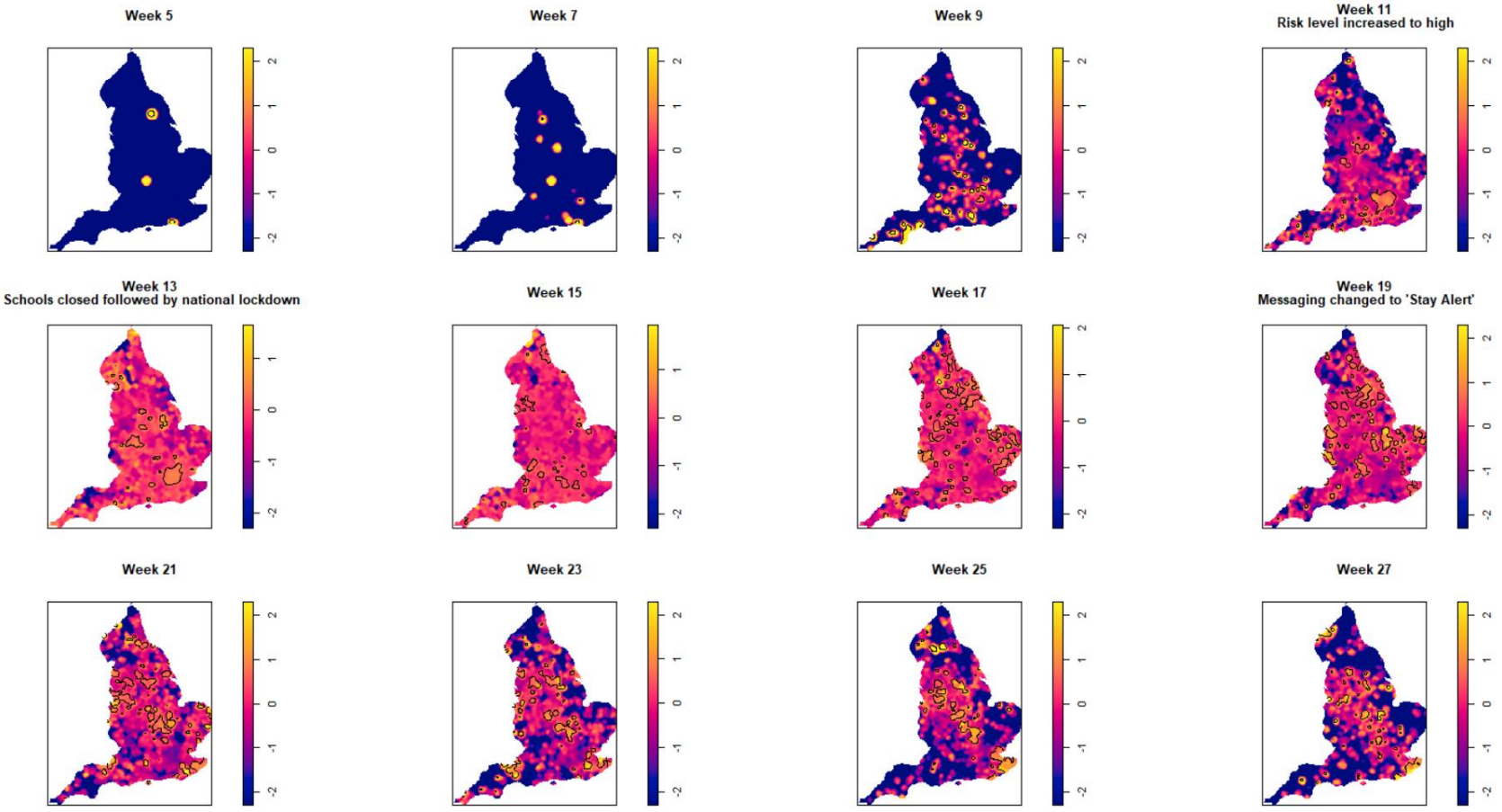


Figure 5.4 Log relative risk spacetime slices using bootstrap bandwidth (based on cases only: *h*=4.8km, *lambda (λ)* =3.2) in 14-day periods from the date of the first case confirmed in ISO Week 5. Solid lines outline areas of significantly higher risk at the 1% confidence level



5.5 Discussion

To the best of our knowledge, this is the first description of the spatio-temporal distribution of COVID-19 in England using unaggregated data. As such, it defines areas of statistically significant high and low risk at a very fine spatial scale, unhampered by administrative boundaries.

Taking into account a seven-day lag for the incubation period (235) prior to sample collection, our results show that geographically widespread transmission was underway at least one week prior to the partial lockdown announced on the 23rd March 2020.

The rapid increase in cases and geographical spread in risk coincided with the roll out of PCR assays to hospitals during March resulting in greater ascertainment. However, intensive sequencing of SARS-CoV-2 genomes revealed that there were multiple introductions from European countries. The frequency of these imports (introduced via multiple entry points by travellers returning to the UK predominately from Spain, Italy and France) reached a peak in mid-March 2020 (Week 12) and led to widespread onward transmission within the UK (236).

The risk was greatest in some, but not all, large urban areas. At the beginning of the outbreak, the risk in London was significantly elevated for a prolonged period but changed abruptly within the period of a single week (Week 15). The reasons for this are unclear but may be related to the impact of non-pharmaceutical interventions (social-distancing, reduced use of public transport etc.) or factors related to immunity. Seroprevalence of antibodies to SARS- CoV-2 in samples from healthy adult blood donors in England showed that the prevalence in London, adjusted for assay accuracy, age and sex, increased from 1.5% in Week 13 to 12.3% in Weeks 15 to 16 and 17.5% in Week 18. Given that the antibody response takes at least two weeks to become detectable, those displaying a

positive result in Week 18 are likely to have become infected before mid-April. By the end of our study period (Week 27), prevalence had dropped to 10% in London (237).

Large urban areas in England have higher population densities and tend to have higher numbers of black, Asian and minority ethnic residents. They are also the areas with the highest deprivation and air pollution scores: all factors associated with an increased risk of infection and/or poorer outcomes following infection with SARS-CoV 2 (238-240).

Selecting the 'right' bandwidth is crucial for this approach. Calculation of the over smoothing bandwidth is extremely quick, and the results provide a good overview of elevated risk. However, this somewhat rudimentary approach is unlikely to identify focused hotspots. The bootstrap method, whilst more computationally intensive, produced a usable bandwidth in less than thirty minutes for the spatial analysis and around ten hours for the spatio-temporal bandwidth. The resulting output provides superior geographical detail allowing resources to be targeted more efficiently. Notwithstanding this, too small a bandwidth results in an under smoothed surface which can erroneously identify 'significant' peaks in risk as a result of increased variability of the kernel estimator. The numeric stability of LSCV and likelihood based methods is known to be questionable in practice, with resulting estimates often being under-smoothed (128).

There are some limitations to this analysis. First, our approach was exploratory and does not account for groups more likely to experience poorer outcomes following infection due to socio-demographic, occupational and environmental factors. Also, the data we used represents those who were symptomatic and sought healthcare. In common with all surveillance systems, this is biased towards the severe end of the disease spectrum. One way of overcoming this bias would be to include results from the wider community testing performed under Pillar 2 of the UK testing regime. We decided not to include this data because Pillar 2 testing was introduced part way through the study and was also subject to

data quality issues until May 2020 (231). The decline in cases described here is likely to be an underestimate of the true community incidence and may not reflect the spatial locations of cases identified under Pillar 2. This requires further investigation, however, considering the way that SARS-CoV-2 is transmitted, we anticipate that this will not differ considerably, and our analysis represents the spatio-temporal 'tip of the iceberg' for COVID-19 in England during the study period. Finally, the PCR assay used by hospitals was rolled out nationally during March 2020 resulting in greatly improved case ascertainment. This coincided with the rapid increase seen during March 2020 so may be an artefact of improved surveillance. However, this does not explain the spatial variation noted beyond the end of March (when the assay was in widespread use) nor the sudden decline in cases in Birmingham and London.

Our analysis demonstrates how KDE can identify areas of England where the risk of COVID-19 infection differs significantly. In terms of controlling transmission, the most important practical application is the accurate identification of areas *within* regions that require improved public health messaging or tailored intervention strategies. Spatial modelling can be used to the predict the spread of infection (133, 241) and the methodology to do this has been available for some time (133, 242). Such approaches have already been applied to COVID-19 case data (243)and self-reported symptoms (242). It is hoped that this will form part of the UK response strategy in the coming months and will be most informative at very fine spatial scales (242). To harness the benefits of such modelling approaches, public health organisations and academic centres must find ways to share information and promote collaboration without compromising patient confidentiality.

To conclude, we present a spatio-temporal analysis of COVID-19 in England covering the first six months of 2020. We recommend that this approach is absorbed into routine surveillance outputs and that ways to confidentially share patient data with academic

collaborators are explored. Further work using Pillar 2 test data and the development of predictive spatial models at fine spatial scales is needed.

6 Discussion

6.1 Chapter findings

This thesis builds upon previous research on STEC O157 and explores hypothesised links from previous research conducted within the UK and further afield.

In Chapter 2, we investigated the spatio-temporal variation in risk of STEC O157 infections in England using over 3,000 records recorded by the NESSS between 2009 and 2015. Our analysis provides evidence that the distribution of STEC O157 infection in England is non-uniform with respect to the distribution of the at-risk population; that the spatial distribution of the three main genetic lineages infecting humans differs significantly and that the spatio-temporal risk is highly dynamic. We also provide evidence that cases of STEC O157 reporting travel within or outside the UK are more likely to live in the south/south-east of the country, meaning that their residential location may not reflect the location of exposure that led to their infection. For non-travel cases we propose that the observed variation in risk is likely to reflect a differential exposure to a source of STEC O157 that is geographically prescribed.

We further investigated the spatial relationship between infections and the presence of hypothesised environmental risk factors in Chapter 3. We observed that risk was elevated for those living in rural areas with high densities of farmed animals and served by private water supplies. Risk was positively associated with cattle density and, for strains associated with severe disease, with sheep density. Our results also indicate that travel abroad may expose individuals to risks not present in their local residential environment. This risk appears to disproportionately affect more affluent people (110, 213) , presumably due to the cost of foreign travel but socio-economic status has also been shown to influence transmission networks amongst those who did not travel outside the UK (111).

In Chapter 4, we also explored spatial risk, but additionally incorporated risk at an individual level by comparing exposures reported by those living in high and low risk areas. We also investigated whether travel within the UK had a modifiable effect on reported exposure. We found that direct contact with the agricultural environment and contact with dogs were important risk factors for residents of high-risk areas, particularly children. For those who travelled in the UK, indirect environmental contact and going on daytrips were the factors associated with increased risk.

Building on the finding in Chapter 3, in Chapter 5, we applied KDE to COVID-19 disease in England during the first six-months of the pandemic to demonstrate the wider applicability of this method. In this chapter we estimated the spatial and spatiotemporal risk of COVID-19 infection across England for the first six months of 2020. Our results demonstrated that widespread transmission was underway when partial lockdown measures were introduced on the 23rd March 2020 and the greatest risk erred towards large urban areas. The rapid growth phase of the outbreak coincided with multiple introductions to England from the European mainland. The spatio-temporal risk was highly variable throughout highlighting that control measures needed to be developed at a smaller spatial scale than the national, regional or local authority level. Our results and experiences led us to recommend that KDE and other spatio-temporal statistical approaches should be incorporated into routine surveillance.

6.2 Recurring themes

6.2.1 Spatial Location

Geography and health are intrinsically linked. Spatial location plays a major role in shaping exposure to environmental risks and many other health effects. These spatial locations may change throughout people's lives, and this is influenced by socio-economic

and labour market conditions (244). At a population level, exposure to environmental risk factors may vary in space and time. At an individual level, exposure to risk factors (and health outcomes) is influenced by socio-demographics, place of residence, behaviour or occupation.

For example, household exposure to radon (a ubiquitous, naturally occurring radioactive gas produced by the decay of uranium present in all rocks and soil) is associated with an increased risk of lung cancer. At a population level, this risk differs markedly because some rocks and soils produce greater amounts of radon. In England, these rocks occur more frequently in the Southwest and Midlands. At an individual level, the risk is around 25 times higher amongst smokers compared with non-smokers but can also be reduced by modifying built environments by, for example, increasing ventilation rates in homes.

In this example, the individual has no control over the geographical presence of the hazard but their probability of exposure can be modified and individual choices may further reduce the risk of developing cancer. However, the degree to which exposure can be modified will differ depending upon an individual's socio-economic status; those who are able to move to an area with lower radon levels or pay for household modifications are more likely to be affluent.

To study the effect of geography on health is therefore challenging, requiring data not only on the outcome of interest but also on the presence of environmental hazards linked to information on exposures, behaviours, movement, and the socio-demographic/economic details of the individuals affected. However, except for specific epidemiological studies, this data is rarely available at the individual level and ecological measures of socio-economic and other demographic information and risk factors are commonly used as surrogates for individual level data.

In this thesis there are numerous examples of how spatial location affects disease risk. For example, Chapters 3 and 6 the risk of STEC and Covid-19 are both shown to vary spatially and also over time. In Chapter 4 there was a lower emphasis on mapping, but risk was demonstrated to vary spatially associated with environmental and socioeconomic factors which themselves vary spatially. In Chapter 5 we explore some of the geographical as well as individual level factors behind low and high risk STEC areas.

6.2.2 Residential risk

One consistent theme throughout is the importance of residential risk; that is, the association between defined geographical areas and increased occurrence of infections seen over time. This suggests that the risk of becoming an STEC case is increased simply by living in a particular area of the country. Although these areas of high risk tend to be sparsely populated, rural areas, not all rural areas are high risk suggesting that it is the presence of the zoonotic reservoir, perhaps in combination with other features, that drives infection and not simply location.

In the UK, residential location is defined in a number of ways from an individual address to various aggregations of these. One commonly used residential location which is frequently used as it perseveres a degree of anonymity is the postcode. Postcodes are part of a coding system designed to help mail delivery and are an abbreviated form of address allowing a set of households to be grouped together. On average, there are 15 households per postcode, but this varies depending on population density meaning that postcode areas in rural areas are larger and contain fewer households. The postcode combines letters and numbers that are based on four geographical levels known as area, districts, sectors and the postcode itself. There are 124 areas, 2,827 districts, 9,487 sectors and around 1.8 million postcodes in the UK (245). In urban areas, postcodes may represent part of a street or a

large organisation. In rural areas, individual postcodes may cover a single building, or a group of properties dispersed over a large area.

One key issue around postcodes is that because they are designed for the operational requirements of postal deliveries the addresses within may not be homogeneous in terms of social or environmental conditions. In England residential postcode data is collected as part of routine STEC surveillance and is well completed (99% of laboratory confirmed cases for the data we used). Cases are also asked to provide the postcode of their workplace and, if the case had travelled within the UK during their incubation period, the postcode(s) of places they had visited.

In this thesis residential postcode of cases was used in Chapter 2 and Chapter 5, to describe the distribution of disease in relation to the population at risk. The analysis simply focused on whether the distribution of STEC O157 was dependent upon spatial location. Our results showed significant spatial variation in risk defined in detail by contour lines that straddled administrative boundaries. This effect was seen at Lineage and sub-lineage level. The spatio-temporal analysis also showed that the areas of increased risk were stable from year to year, suggesting that a greater proportion of infection is driven by the presence of risk factors within these areas as opposed to introductions from elsewhere (for example from food or infected individuals).

In Chapter 4 we also used postcode to accurately identify destinations for cases who reported travelling within the UK. This enabled us to determine whether cases had travelled to an area considered high risk during their incubation period.

However, postcodes are still an imprecise measure of residential location and an increasing number of studies are using individual address or unique property reference number for a

more precise measure of residential location (246, 247), better suited to measure the effects of highly proximal risk factors (248) .

However, the potential benefits of fine spatial resolution of human case data is of lower value unless it can be mapped to similarly fine exposure data. An ideal situation, although there are issues around mobility discussed in the next section, would be fine resolution (e.g., postcode) case and exposure data. However, we found that the spatial unit at which census and other data were available to us differed and were always coarser than the case data. For example, information on rurality was available at Output Area (OA) level. OAs are created specifically for the output of census estimates (age, sex, population density etc.) in England and Wales (ref). Built from clusters of adjacent postcodes, OAs are the smallest unit, containing ~125 households and with a population of ~300. They vary greatly in size and shape as a result of the population density gradient between urban and rural regions. For example, a single tower block in a large city may consist of more than one OA, whereas a large area of remote moorland may be covered by a single OA.

Detailed socio-economic status of an area as well as PWS counts were only at Lower Super Output Area (LSOA). LSOAs are built of OAs, typically 5, and so contain ~625 households or a mean population of ~1500, with a minimum population of 1000. There are 32,482 LSOAs in England. SOAs are built from groups of OAs and are more homogenous in terms of population and size and are not subject to boundary changes meaning that datasets can be reliably compared across time. Importantly the aggregation of OAs to LSOAs is designed in such as way as to generate LSOAs which are relatively similar from a socioeconomic perspective (249).

Information on animal density was derived from a 5x5 km grid based upon aggregated returns from the UK agricultural census. More accurate agricultural data at the farm level is available but unavailable to confidential reasons. Other researchers have looked to

overcome this coarse resolution using dasymetric mapping to redistribution animals within these relatively coarse grid cells using information on landcover within (210, 250-254). Freshwater coverage was extracted from a detailed topographical map and in its raw form was not spatially aggregated.

Given this complexity of different geographies the finest spatial scale at which we could explore the effect of hypothesised risk factors in Chapter 3 was at LSOA level. This aggregation implicitly assumes that the risk of disease or distribution of risk factors is constant within and between the areas of aggregation. This is rarely the case for infectious diseases and hampers detailed exploration of spatially varying risk factors in relation to disease. However, LSOA is a much finer scale than that used in the majority of comparable work performed at state (75, 185, 255) or municipality level (140, 141, 144) . However, despite efforts to make these areas as socially homogenous as possible, the characteristics of LSOAs vary (Area range: 0.02-684 km2, Population range: 985-8,300 persons), and being a fairly recent development, rarely capture the complexities of the underlying populations, often dissecting single streets or neighbourhoods rather than aligning with them. Features within their boundaries will also vary spatially and this will be amplified in larger LSOAs that tend to be in rural areas with sparse populations. The other data were also at LSOA level again making the implicit assumption that the data is homogenous throughout.

In the future there would be options to perform the analysis at a finer spatial resolution. One relatively straightforward extension would be to perform the analysis at the level of OA. LSOA was the unit used in Chapter 4 due to the spatial resolution of the Index of Material Deprivation. However, other measures of deprivation exist (e.g., Carstairs or Jarman Index) which only require census information. Hence although they may contain

less information on deprivation (e.g., income and crime data) their benefit is that they are available at a finer spatial resolution.

6.2.3 Importance of mobility

The residential location of a person is often the best available surrogate for the micro-environment to which they are principally exposed as they go about their daily business (256)(246, 248). However, this assumes that the individual did not move far from home during their incubation period when in fact many people commute to work and also go on daytrips or move further away from home for short breaks or holidays. In our own study (Chapter 4), 44% reported large-scale travelling (abroad, within the UK or on daytrips) during their incubation period. Clearly, if a case was exposed away from home, using residential location may be misleading when considering risk in space and time.

One major advance in this study over others is that we considered the impact of large-scale travel behaviour of cases (i.e., day trips, travel within the UK and overseas). In Chapter 2, we used travel status to stratify our analyses to focus on those who were most likely infected close to home by comparing risk factors against cases who had travelled. There was significant impact on results, particularly the interaction between SES and travel, suggesting that more affluent cases were exposed to risk factors outside the UK more often than those living in less affluent areas. In Chapter 2, we compared the spatial location of cases reporting travel with those not reporting travel and found significant differences. Cases reporting travel were more likely to reside in the more affluent South-East of England and less likely to live in the North and Southwest of the country, areas that are considered to present a higher risk of indigenous infection.

Whilst travel outside the UK is accepted to be a major risk factor for a wide range of infectious intestinal disease (IID) (257-259) and many studies remove cases that have

travelled overseas (1, 14) , we believe this is the first study of any IID to use details of travel within the UK and link this to ecological and behavioural risk factors as a reliable proxy for place of exposure. Not only did this enable us to explore the importance of these risk factors between groups who lived in high and low risk areas but also amongst those who lived in low-risk areas but travelled to other parts of the UK where their probability of exposure to risk factors may have been increased. Overall, those living in high-risk areas were less likely to travel away from home.

We went further by linking travel location to reported behaviours at those places. In Chapter 4, we used data at a fine spatial to show where people had travelled to in the days preceding the onset of their symptoms. This allowed us to investigate how behaviour differed between cases living in high-risk areas, travellers and those who stayed at home. The effect of travel on reported 'risky' behaviour was significant with those travelling reporting more exposure to the agricultural environment and this was particularly significant for those travelling to high-risk areas. The significance of these analyses taking national travel into account suggests that IID surveillance should record as much travel information as possible to aid future IID studies.

Yet we acknowledge the limitations of questionnaire approaches to record many of the movements away from home that individuals undertake. Longer and more detailed questionnaires may be infeasible and lead to low response rates or poor-quality information (260, 261). Real time recording of individual movement using GPS trackers or mobile phone technology are starting to be used for epidemiological studies but are currently clearly fraught with ethical and practical complexities (262). Despite these challenges, highly detailed mobility information could provide insights on where an individual had been during their incubation period e.g. supermarkets, restaurants, petting farms,

agricultural areas, which would be less subject to the recall and cognitive bias often associated with traditional data collection methods..

In addition to cases, risk factors may not be spatially static either. In this study the data we used is a static estimate of livestock/animal density at a predefined geographical scale (X km2). Yet livestock move throughout their lives over relatively small distances such as within and between field. Larger scale movement also occur on weekly and seasonal scales according to the production cycle with peaks in late spring and early autumn for cattle in the UK. This is coincident with the start and finish of the 'human' STEC season. The geographical distribution of these larger scale movements is fairly stable (263) and reflects regional networks of animal movements (263). Movement of animals increases stress and has been shown to increase STEC shedding levels (264, 265). In addition, a study of hides at slaughter showed that these were always contaminated with strains not present on the farm of origin indicating that cross contamination occurs readily when animals are transported (266).

Both these small and large-scale movements mean that animal density (and therefore risk) changes throughout the year. In England generating space – time surfaces of animal density should be relatively straightforward as data on every movement of every cow, sheep, goat, and deer in the UK is collected at postcode level. These data could be used Unfortunately access to these data is restricted making this impossible. Had we had detailed animal movement data, we could have explored seasonal changes in human STEC rates with corresponding spatio-temporal changes in animal density at species level by creating a dynamic movement surface, ideally combined with WGS results from veterinary surveillance systems.

Variations in animal density over the year are also likely to have an impact on the environmental presence and load of STEC. Previous research has estimated this through incorporation of shedding rates in different aged livestock(210). Variations in animal density may also result in changes to strains colonising the animal reservoir with subsequent effects on human health. It will also affect the spatial distribution of STEC in the environment.

6.2.4 Typing

During the period of this research, typing methods for STEC O157 underwent enormous change form phage typing, through MLVA and since 2015, WGS. We embraced these changes by using comparable data sets. In Chapter 2, we used PT to stratify our temporal analysis to describe how the decline in STEC O157 cases since 2014 was largely due to a reduction in PT21/28/ Lineage II only, particularly in rural areas. We identified significant differences between PT 21/28 and PT8 cases with particular reference to socio-economic status and foreign travel as well as a previously unidentified association with PT21/28 and sheep density that requires further investigation. In Chapter 3, we used the highly specific results of WGS to explore spatial variation at lineage and sub-lineage level and also link to PT results using recently described phylogenetic relationships allocating PT to one of three common lineages circulating in the UK (12). This an emerging area of research but there are currently few examples in the literature that describe the spatio-temporal characteristics of IID at such a fine geographical and microbiological scale (118). Our results suggest this is a positive development and indicates that surveillance should consider typing as routine.

6.3 Final thoughts

This thesis aimed to build upon previous research on STEC O157 and explore hypothesised links from previous research conducted within the UK and further afield with a specific focus on applied spatio-temporal methods. Previously, these aspects had not

been studied or had been performed at coarse spatial scales such as UTLA level. Throughout, we focused on sporadic infections (the major fraction of cases) and the role of residential location and the environment at the finest spatial scales possible.

To do this, we produced novel datasets with spatial links to ecological and molecular typing information from a wide range of sources. For the first time in England, we investigated the effects of hypothesised risk factors (including animal density, SES, PWS) on case incidence, described geographical differences in the distribution of STEC cases using phenotypic and genetic information, and explored the relationship between residential location and the risk of infection. Previously, the role of small ruminants and other animals as a reservoir or maintenance species had received much less attention than cattle.

Our approach also considered the role of travel and its interaction with the socio-demographic and behavioural characteristics of cases in relation to the risk of infection. We did this by exploiting the information collected on UK travel destinations and developing a robust method to deal with any inconsistencies.

The outputs from this thesis have contributed to and expanded the scientific literature by showing that the spatio-temporal risk of STEC differs significantly across England, that this risk is associated with residence in rural areas with high densities of farmed animals and that infection with more virulent strains may be associated with sheep density. We also show that cases living in high-risk areas were less likely to travel in the UK, suggesting that their infections were acquired close to home with direct contact with the farming environment and contact with dogs emerging as important risk factors. Risk factors for cases living in low-risk areas who travelled within the UK, during their incubation period, daytrips and direct/indirect contact with the environment was more important.

Finally, although the COVID pandemic interrupted the progress of this research, we were able to rapidly apply the methods to describe the spatio-temporal distribution of COVID-19 in England over the first six- months of 2020 and incorporated these outputs into routine surveillance systems for new variants of SARS CoV-2 during 2021 (267). This highlighted the challenges of using very large (>50,000 daily cases) datasets and the limitations of computationally intensive adaptive bandwidth selectors in this scenario. It also prompted fruitful collaborations between UKHSA and academic partners on the use of cutting-edge spatial statistics during the COVID response but highlighted the ongoing and urgent need for public health organisations and academic centres to find ways to share information and promote collaboration without compromising patient confidentiality.

6.4 Reflections on the thesis

A more sophisticated approach for Chapters 2 and 3 would have been to use point process models. Spatial statistics are embedded in this approach and these models can also be used to predict the spatio-temporal spread of infection as well as exploring interactions between cases. The use of point process models requires not only data on the distribution of cases, but also covariate data at a similar spatial scale but this was not available, or the time scale and/or resource required to access the data was prohibitive.

6.5 Future research

In Chapter 3, we identified a spatial association between sheep density and cases falling into Lineage II associated with more severe disease. Although sheep meat is not considered a significant source of foodborne STEC infection, prevalence is fairly high in faeces and on fleeces (268). Our analysis explored the spatial relationship between animal density and cases as a proxy for direct zoonotic or indirect environmental transmission. This would benefit from further exploration, particularly as colonisation in sheep is comparatively

high, varies throughout the year and may be characterised by strains that are of public health significance (269) yet underestimated by surveillance systems (114, 270).

7 Supplementary material

SM 1 Copy of STEC enhanced surveillance questionnaire

SM 2 Spatio-temporal animation of STEC infections 2009-2015 using oversmoothed and likelihood cross validation bandwidths

SM 3 Estimated log relative risk surfaces for STEC cases and controls in rural areas only and controls stratified by rural urban status

SM 4 Spatio-temporal animation of COVID-19 in England between January and June 2020 using oversmoothed and bootstrapped bandwidths

8 Glossary of terms

CI	Confidence Interval
GBRU	Gastrointestinal Bacterial Reference Unit
GIS	Geographical Information System
HUS	Haemolytic Uraemic Syndrome
IMD	Indices of Multiple Deprivation
KDE	Kernel Density Estimation
km	Kilometre
LSOA	Lower Super Output Area
MLVA	Multi-Locus Variable Tandem Repeat Analysis
NESSS	National Enhanced STEC Surveillance System
NPD	National Population Database
ONS	Office for National Statistics
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PHE	Public Health England
PT	Phage Type
PWS	Private water supply
SECs	Socio Economic Circumstances
STEC	Shiga-toxin producing *E. coli*
TTP	Thrombotic Thrombocytopenic Purpura
UKHSA	UK Health Security Agency
WGS	Whole Genome Sequencing

9 References

(1) **Byrne L, et al.** The epidemiology, microbiology and clinical impact of Shiga toxin-producing Escherichia coli in England, 2009-2012. *Epidemiol Infect* 2015; **143**(16): 3475-3487.

(2) **Melton-Celsa AR.** Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol Spectr* 2014; **2**(4): Ehec-0024-2013.

(3) **Riley LW, et al.** Hemorrhagic colitis associated with a rare Escherichia coli serotype. *N Engl J Med* 1983; **308**(12): 681-685.

(4) **Chase-Topping M, et al.** Super-shedding and the link between human infection and livestock carriage of Escherichia coli O157. *Nat Rev Microbiol* 2008; **6**(12): 904-912.

(5) **Awofisayo-Okuyelu A, et al.** Analysis of individual patient data to describe the incubation period distribution of Shiga-toxin producing Escherichia coli. *Epidemiol Infect* 2019; **147**: e162.

(6) **Ludwig S, Zarbock A.** Coronaviruses and SARS-CoV-2: A Brief Overview. *Anesth Analg* 2020; **131**(1): 93-96.

(7) **Lai MM.** RNA recombination in animal and plant viruses. *Microbiol Rev* 1992; **56**(1): 61-79.

(8) **Drexler JF, Corman VM, Drosten C.** Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. *Antiviral Res* 2014; **101**: 45-56.

(9) Anonymous. The Health Protection (Notification) Regulations 2010. In: Limited TSO, ed. UK: The Stationery Office Limited, 2010.

(10) **Food Standards Agency.** The second study of infectious intestinal disease in the community (IID2 Study). 2012.

(11) **Adams NL, et al.** Shiga Toxin-Producing Escherichia coli O157, England and Wales, 1983-2012. *Emerg Infect Dis* 2016; **22**(4): 590-597.

(12) **Dallman TJ, et al.** Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. *Microb Genom* 2015; **1**(3): e000029.

(13) **Dallman TJ, et al.** Whole-genome sequencing for national surveillance of Shiga toxin-producing Escherichia coli O157. *Clin Infect Dis* 2015; **61**(3): 305-312.

(14) **Elson R, et al.** A spatial and temporal analysis of risk factors associated with sporadic Shiga toxin-producing Escherichia coli O157 infection in England between 2009 and 2015. *Epidemiol Infect* 2018; **146**(15): 1928-1939.

(15) **Butcher H, et al.** Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing Escherichia coli O157 associated with raw drinking milk. *Epidemiol Infect* 2016; **144**(13): 2812-2823.

(16) **Byrne L, et al.** Evaluating the use of multilocus variable number tandem repeat analysis of Shiga toxin-producing Escherichia coli O157 as a routine public health tool in England. *PLoS One* 2014; **9**(1): e85901.

(17) **Schutz K, et al.** Evolutionary Context of Non-Sorbitol-Fermenting Shiga Toxin-Producing Escherichia coli O55:H7. *Emerg Infect Dis* 2017; **23**(12): 1966-1973.

(18) **Cowley LA, et al.** Analysis of whole genome sequencing for the Escherichia coli O157:H7 typing phages. *BMC Genomics* 2015; **16**(1): 271.

(19) **Khakhria R, Duck D, Lior H.** Extended phage-typing scheme for Escherichia coli O157:H7. *Epidemiol Infect* 1990; **105**(3): 511-520.

(20) **Launders N, et al.** Disease severity of Shiga toxin-producing E. coli O157 and factors influencing the development of typical haemolytic uraemic syndrome: a retrospective cohort study, 2009-2012. *BMJ Open* 2016; **6**(1): e009933.

(21) **Lynn RM, et al.** Childhood hemolytic uremic syndrome, United Kingdom and Ireland. *Emerg Infect Dis* 2005; **11**(4): 590-596.

(22) **Allerberger F, et al.** [Enterohemorrhagic Escherichia coli and hemolytic-uremic syndrome]. *Wien Klin Wochenschr* 1997; **109**(17): 669-677.

(23) **Jokiranta TS.** HUS and atypical HUS. *Blood* 2017; **129**(21): 2847-2856.

(24) **Karmali MA.** Factors in the emergence of serious human infections associated with highly pathogenic strains of shiga toxin-producing Escherichia coli. *Int J Med Microbiol* 2018; **308**(8): 1067-1072.

(25) **Karmali MA.** Host and pathogen determinants of verocytotoxin-producing Escherichia coli-associated hemolytic uremic syndrome. *Kidney Int Suppl* 2009(112): S4-7.

(26) **Majowicz SE, et al.** Global incidence of human Shiga toxin-producing Escherichia coli infections and deaths: a systematic review and knowledge synthesis. *Foodborne Pathog Dis* 2014; **11**(6): 447-455.

(27) **European Centre for Disease Prevention and Control.** Shiga toxin-producing Escherichia coli (STEC) infection - Annual Epidemiological Report for 2019. 2021.

(28) **Buchholz U, et al.** German outbreak of Escherichia coli O104:H4 associated with sprouts. *N Engl J Med* 2011; **365**(19): 1763-1770.

(29) **Tam CC, et al.** Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut* 2012; **61**(1): 69-77.

(30) **Lal A, et al.** Seasonality in human zoonotic enteric diseases: a systematic review. *PLoS One* 2012; **7**(4): e31883.

(31) **Roberts JA, Upton PA, Azene G.** Escherichia coli O157:H7; an economic assessment of an outbreak. *J Public Health Med* 2000; **22**(1): 99-107.

(32) **Cowden JM, et al.** Epidemiological investigation of the central Scotland outbreak of Escherichia coli O157 infection, November to December 1996. *Epidemiol Infect* 2001; **126**(3): 335-341.

(33) **Food Standards Agency.** The Burden of Foodborne Disease in the UK 2018. 2020.

(34) **Davis MA, et al.** Correlation between geographic distance and genetic similarity in an international collection of bovine faecal Escherichia coli O157:H7 isolates. *Epidemiol Infect* 2003; **131**(2): 923-930.

(35) **Innocent GT, et al.** Spatial and temporal epidemiology of sporadic human cases of Escherichia coli O157 in Scotland, 1996-1999. *Epidemiol Infect* 2005; **133**(6): 1033-1041.

(36) **Widgren S, et al.** Data-driven network modelling of disease transmission using complete population movement data: spread of VTEC O157 in Swedish cattle. *Vet Res* 2016; **47**(1): 81.

(37) **Scotland FS**. E. coli O157 Super-shedding in Cattle and Mitigation of Human Risk; 2018.

(38) **Franz E, et al.** Genetic features differentiating bovine, food, and human isolates of shiga toxin-producing Escherichia coli O157 in The Netherlands. *J Clin Microbiol* 2012; **50**(3): 772-780.

(39) **Mikhail AFW, et al.** An outbreak of Shiga Toxin-producing Escherichia coli O157:H7 associated with contaminated salad leaves: epidemiological, genomic and food trace back investigations - CORRIGENDUM. *Epidemiol Infect* 2018; **146**(14): 1879.

(40) **Farrington CP, et al.** A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1996; **159**(3): 547-563.

(41) **Noufaily A, et al.** An improved algorithm for outbreak detection in multiple surveillance systems. *Stat Med* 2013; **32**(7): 1206-1222.

(42) **Pearce MC, et al.** Temporal and spatial patterns of bovine Escherichia coli O157 prevalence and comparison of temporal changes in the patterns of phage types associated with bovine shedding and human E. coli O157 cases in Scotland between 1998-2000 and 2002-2004. *BMC Microbiol* 2009; **9**: 276.

(43) **Health Protection Scotland.** VTEC/E. coli O157 Action Plan for Scotland, 2013-2017. 2013.

(44) **O'Brien SJ, Adak GK, Gilham C.** Contact with farming environment as a major risk factor for Shiga toxin (Vero cytotoxin)-producing Escherichia coli O157 infection in humans. *Emerg Infect Dis* 2001; **7**(6): 1049-1051.

(45) **Ihekweazu C, et al.** Large outbreak of verocytotoxin-producing Escherichia coli O157 infection in visitors to a petting farm in South East England, 2009. *Epidemiol Infect* 2012; **140**(8): 1400-1413.

(46) **Koyun OY, et al.** Disease Occurrence in- and the Transferal of Zoonotic Agents by North American Feedlot Cattle. *Foods* 2023; **12**(4).

(47) **Abney SE, et al.** Toilet hygiene-review and research needs. *J Appl Microbiol* 2021; **131**(6): 2705-2714.

(48) **Angelakis E, et al.** Paper money and coins as potential vectors of transmissible disease. *Future Microbiol* 2014; **9**(2): 249-261.

(49) **Cliver DO.** Cutting boards in Salmonella cross-contamination. *J AOAC Int* 2006; **89**(2): 538-542.

(50) **Williams AP, et al.** Leaching of bioluminescent Escherichia coli O157:H7 from sheep and cattle faeces during simulated rainstorm events. *J Appl Microbiol* 2008; **105**(5): 1452-1460.

(51) **Anon.** VTEC O157 outbreak linked to beach holidays. *Commun Dis Rep CDR Wkly* 1999; **9**(37): 327, 330.

(52) **Bruneau A, et al.** Outbreak of E. coli O157:H7 associated with bathing at a public beach in the Montreal-Centre region. *Can Commun Dis Rep* 2004; **30**(15): 133-136.

(53) **Harrison S, Kinra S.** Outbreak of Escherichia coli O157 associated with a busy bathing beach. *Commun Dis Public Health* 2004; **7**(1): 47-50.

(54) **Ihekweazu C, et al.** Outbreak of E. coli O157 infection in the south west of the UK: risks from streams crossing seaside beaches. *Euro Surveill* 2006; **11**(4): 128-130.

(55) **Paunio M, et al.** Swimming-associated outbreak of Escherichia coli O157:H7. *Epidemiol Infect* 1999; **122**(1): 1-5.

(56) **Richardson HY, et al.** Microbiological surveillance of private water supplies in England: the impact of environmental and climate factors on water quality. *Water Res* 2009; **43**(8): 2159-2168.

(57) **Inspectorate DW**. RISK ASSESSMENT OF VTEC INFECTIONS IN ENGLISH AND WELSH DRINKING WATER report for contract DWI/ 70/2/256; 2015.

(58) **Mughini-Gras L, et al.** Attribution of human infections with Shiga toxin-producing Escherichia coli (STEC) to livestock sources and identification of source-specific risk factors, The Netherlands (2010-2014). *Zoonoses Public Health* 2018; **65**(1): e8-e22.

(59) **Friesema IH, et al.** Geographical association between livestock density and human Shiga toxin-producing Escherichia coli O157 infections. *Epidemiol Infect* 2011; **139**(7): 1081-1087.

(60) **Haus-Cheymol R, et al.** Association between indicators of cattle density and incidence of paediatric haemolytic-uraemic syndrome (HUS) in children under 15 years of age in France between 1996 and 2001: an ecological study. *Epidemiol Infect* 2006; **134**(4): 712-718.

(61) **Frank C, et al.** Cattle density and Shiga toxin-producing Escherichia coli infection in Germany: increased risk for most but not all serogroups. *Vector Borne Zoonotic Dis* 2008; **8**(5): 635-643.

(62) **Kistemann T, et al.** GIS-supported investigation of human EHEC and cattle VTEC O157 infections in Sweden: geographical distribution, spatial variation and possible risk factors. *Epidemiol Infect* 2004; **132**(3): 495-505.

(63) **Jalava K, et al.** Agricultural, socioeconomic and environmental variables as risks for human verotoxigenic Escherichia coli (VTEC) infection in Finland. *BMC Infect Dis* 2011; **11**: 275.

(64) **Ohaiseadha C, et al.** A geostatistical investigation of agricultural and infrastructural risk factors associated with primary verotoxigenic E. coli (VTEC) infection in the Republic of Ireland, 2008-2013. *Epidemiol Infect* 2017; **145**(1): 95-105.

(65) **Elson R, et al.** Application of kernel smoothing to estimate the spatio-temporal variation in risk of STEC O157 in England. *Spatial and Spatio-temporal Epidemiology* 2020; **32**: 100305.

(66) **Friesema IH, et al.** Risk factors for sporadic Shiga toxin-producing Escherichia coli O157 and non-O157 illness in The Netherlands, 2008-2012, using periodically surveyed controls. *Epidemiol Infect* 2015; **143**(7): 1360-1367.

(67) **Marion JW, et al.** Association of gastrointestinal illness and recreational water exposure at an inland U.S. beach. *Water Res* 2010; **44**(16): 4796-4804.

(68) **Bentancor AB, et al.** Risk factors for Shiga toxin-producing Escherichia coli infections in preadolescent schoolchildren in Buenos Aires, Argentina. *J Infect Dev Ctries* 2012; **6**(5): 378-386.

(69) **Bettelheim KA, Goldwater PN.** An Escherichia coli O157:H7 outbreak? *Clin Infect Dis* 2000; **30**(6): 984.

(70) **Nace ME, et al.** Notes from the Field: Escherichia coli O157:H7 Outbreak in Children with Clostridioides difficile Colonization Associated with an Improperly Treated Swimming Pool - Pennsylvania, June 2021. *MMWR Morb Mortal Wkly Rep* 2022; **71**(20): 690-691.

(71) **Pereboom M, et al.** Shiga toxin-producing Escherichia coli (STEC) O157 outbreak associated with likely transmission in an inflatable home paddling pool in England, June 2017. *Perspect Public Health* 2018; **138**(5): 279-281.

(72) **Verma A, et al.** An outbreak of E. coli O157 associated with a swimming pool: an unusual vehicle of transmission. *Epidemiol Infect* 2007; **135**(6): 989-992.

(73) **Butt S, et al.** Evidence of on-going transmission of Shiga toxin-producing Escherichia coli O157:H7 following a foodborne outbreak. *Epidemiol Infect* 2021; **149**: e147.

(74) **Conrad CC, et al.** Farm Fairs and Petting Zoos: A Review of Animal Contact as a Source of Zoonotic Enteric Disease. *Foodborne Pathog Dis* 2017; **14**(2): 59-73.

(75) **Daly RF, Hill NT.** Characterizing the Role of Animal Exposures in Cryptosporidiosis and Shiga Toxin-producing Escherichia coli Infections: South Dakota, 2012. *Zoonoses Public Health* 2016; **63**(6): 467-476.

(76) **Isler M, et al.** Animal petting zoos as sources of Shiga toxin-producing Escherichia coli, Salmonella and extended-spectrum β-lactamase (ESBL)-producing Enterobacteriaceae. *Zoonoses Public Health* 2021; **68**(2): 79-87.

(77) **Rowell S, et al.** An outbreak of Shiga toxin-producing Escherichia coli serogroup O157 linked to a lamb-feeding event. *Epidemiol Infect* 2016; **144**(12): 2494-2500.

(78) **Pennington TH.** E. coli O157 outbreaks in the United Kingdom: past, present, and future. *Infect Drug Resist* 2014; **7**: 211-222.

(79) **McLauchlin J, et al.** Microbiological quality of raw drinking milk and unpasteurised dairy products: results from England 2013-2019. *Epidemiol Infect* 2020; **148**: e135.

(80) **Anon.** VTEC O157 infection and unpasteurised cream--update. *Commun Dis Rep CDR Wkly* 1998; **8**(44): 389, 392.

(81) **Anon.** Cases of Escherichia coli O157 infection associated with unpasteurised cream. *Commun Dis Rep CDR Wkly* 1998; **8**(43): 377.

(82) **Anon.** Escherichia coli O157 associated with eating unpasteurised cheese--update. *Commun Dis Rep CDR Wkly* 1999; **9**(15): 131, 134.

(83) **Anon.** Escherichia coli O157 associated with eating unpasteurised cheese. *Commun Dis Rep CDR Wkly* 1999; **9**(13): 113, 116.

(84) **Allerberger F, et al.** Escherichia coli O157 infections and unpasteurised milk. *Euro Surveill* 2001; **6**(10): 147-151.

(85) **Clark A, et al.** A community outbreak of Vero cytotoxin producing Escherichia coli O157 infection linked to a small farm dairy. *Commun Dis Rep CDR Rev* 1997; **7**(13): R206-211.

(86) **Guzman-Hernandez R, et al.** Mexican unpasteurised fresh cheeses are contaminated with Salmonella spp., non-O157 Shiga toxin producing Escherichia coli and potential uropathogenic E. coli strains: A public health risk. *Int J Food Microbiol* 2016; **237**: 10-16.

(87) **Jenkins C, et al.** Outbreak of STEC O157:H7 linked to a milk pasteurisation failure at a dairy farm in England, 2019. *Epidemiol Infect* 2022; **150**: e114.

(88) **Liptakova A, et al.** A family outbreak of haemolytic uraemic syndrome and haemorrhagic colitis caused by verocytotoxigenic Escherichia coli O157 from unpasteurised cow's milk in Slovakia. *Clin Microbiol Infect* 2004; **10**(6): 576-578.

(89) **Marozzi S, et al.** Prevalence and Molecular Characterisation of Shiga Toxin-Producing Escherichia Coli in Raw Milk Cheeses from Lazio Region, Italy. *Ital J Food Saf* 2016; **5**(1): 4566.

(90) **Kintz E, et al.** Transmission pathways for sporadic Shiga-toxin producing E. coli infections: A systematic review and meta-analysis. *Int J Hyg Environ Health* 2017; **220**(1): 57-67.

(91) **Kintz E, et al.** Outbreaks of Shiga Toxin-Producing Escherichia coli Linked to Sprouted Seeds, Salad, and Leafy Greens: A Systematic Review. *J Food Prot* 2019; **82**(11): 1950-1958.

(92) **Dallman TJ, et al.** Identification of domestic reservoirs and common exposures in an emerging lineage of Shiga toxin-producing Escherichia coli O157:H7 in England: a genomic epidemiological analysis. *Lancet Microbe* 2022; **3**(8): e606-e615.

(93) **Holme R.** Drinking water contamination in Walkerton, Ontario: positive resolutions from a tragic event. *Water Sci Technol* 2003; **47**(3): 1-6.

(94) **Chalmers RM, Aird H, Bolton FJ.** Waterborne Escherichia coli O157. *Symp Ser Soc Appl Microbiol* 2000(29): 124s-132s.

(95) **Exner M, et al.** Consensus report: E. coli O104:H4 (HUSEC041) and the potential threat to European water supplies. *Int J Hyg Environ Health* 2011; **214**(6): 500-501.

(96) **Garvey P, et al.** Verotoxigenic Escherichia coli transmission in Ireland: a review of notified outbreaks, 2004-2012. *Epidemiol Infect* 2016; **144**(5): 917-926.

(97) **Jaros P, et al.** A prospective case-control and molecular epidemiological study of human cases of Shiga toxin-producing Escherichia coli in New Zealand. *BMC Infect Dis* 2013; **13**: 450.

(98) **Khabo-Mmekoa CM, Genthe B, Momba MNB.** Enteric Pathogens Risk Factors Associated with Household Drinking Water: A Case Study in Ugu District Kwa-Zulu Natal Province, South Africa. *Int J Environ Res Public Health* 2022; **19**(8).

(99) **Licence K, et al.** An outbreak of E. coli O157 infection with evidence of spread from animals to man through contamination of a private water supply. *Epidemiol Infect* 2001; **126**(1): 135-138.

(100) **Money P, et al.** Cattle, weather and water: mapping Escherichia coli O157:H7 infections in humans in England and Scotland. *Environ Microbiol* 2010; **12**(10): 2633-2644.

(101) **Óhaiseadha C, et al.** A geostatistical investigation of agricultural and infrastructural risk factors associated with primary verotoxigenic E. coli (VTEC) infection in the Republic of Ireland, 2008-2013. *Epidemiol Infect* 2017; **145**(1): 95-105.

(102) **O'Sullivan MB, et al.** Increase in VTEC cases in the south of Ireland: link to private wells? *Euro Surveill* 2008; **13**(39).

(103) **Pebody RG, et al.** An international outbreak of Vero cytotoxin-producing Escherichia coli O157 infection amongst tourists; a challenge for the European infectious disease surveillance network. *Epidemiol Infect* 1999; **123**(2): 217-223.

(104) **Quilliam RS, et al.** Seroprevalence and risk factors associated with Escherichia coli O157 in a farming population. *Zoonoses Public Health* 2012; **59**(2): 83-88.

(105) **Raseala CM, Ekwanzala MD, Momba MNB.** Multilocus-based phylogenetic analysis of extended-spectrum beta-lactamase Escherichia coli O157:H7 uncovers related strains between agriculture and nearby water sources. *J Infect Public Health* 2020; **13**(12): 1899-1906.

(106) **Reynolds C, et al.** Evaluating the risks associated with Shiga-toxin-producing Escherichia coli (STEC) in private well waters in Canada. *Can J Microbiol* 2020; **66**(5): 337-350.

(107) **Schack RM, et al.** A fatal case of Shiga toxin-producing Escherichia coli linked to a private drinking water supply. *Commun Dis Intell (2018)* 2020; **44**.

(108) **Schets FM, et al.** Escherichia coli O157:H7 in drinking water from private water supplies in the Netherlands. *Water Res* 2005; **39**(18): 4485-4493.

(109) **Strachan NJ, et al.** Escherichia coli O157: burger bug or environmental pathogen? *Int J Food Microbiol* 2006; **112**(2): 129-137.

(110) **Adams NL, et al.** Influence of socio-economic status on Shiga toxin-producing Escherichia coli (STEC) infection incidence, risk factors and clinical features. *Epidemiol Infect* 2019; **147**: e215.

(111) **Rotheram S, et al.** Linking pathogens, people and places: Using geo-ethnography to understand place-based, socio-economic inequalities in gastrointestinal infections in the UK. *Health Place* 2022; **74**: 102741.

(112) **Salmon RL, et al.** Bloody diarrhoea: Test early for verotoxin producing Escherichia coli. *Bmj* 2008; **336**(7654): 1147-1148.

(113) **Murphy MS.** Management of bloody diarrhoea in children in primary care. *Bmj* 2008; **336**(7651): 1010-1015.

(114) **Byrne L, et al.** Epidemiology and microbiology of Shiga toxin-producing Escherichia coli other than serogroup O157 in England, 2009-2013. *J Med Microbiol* 2014; **63**(Pt 9): 1181-1188.

(115) **Vally H, et al.** Epidemiology of Shiga toxin producing Escherichia coli in Australia, 2000-2010. *BMC Public Health* 2012; **12**: 63.

(116) **Locking ME, et al.** Escherichia coli O157 infection and secondary spread, Scotland, 1999-2008. *Emerg Infect Dis* 2011; **17**(3): 524-527.

(117) **Rivas M, et al.** [The epidemiology of hemolytic uremic syndrome in Argentina. Diagnosis of the etiologic agent, reservoirs and routes of transmission]. *Medicina (B Aires)* 2006; **66 Suppl 3**: 27-32.

(118) **Tarr GAM, et al.** Geogenomic Segregation and Temporal Trends of Human Pathogenic Escherichia coli O157:H7, Washington, USA, 2005-2014(1). *Emerg Infect Dis* 2018; **24**(1): 32-39.

(119) **Jaros P, et al.** Geographic divergence of bovine and human Shiga toxin-producing Escherichia coli O157:H7 genotypes, New Zealand. *Emerg Infect Dis* 2014; **20**(12): 1980-1989.

(120) **Mellor GE, et al.** Geographically distinct Escherichia coli O157 isolates differ by lineage, Shiga toxin genotype, and total shiga toxin production. *J Clin Microbiol* 2015; **53**(2): 579-586.

(121) **Cha W, et al.** Factors associated with increasing campylobacteriosis incidence in Michigan, 2004-2013. *Epidemiol Infect* 2016; **144**(15): 3316-3325.

(122) **Anselin L.** Local Indicators of Spatial Association—LISA. *Geographical Analysis* 1995; **27**(2): 93-115.

(123) **Diggle PJ, Chetwynd AG.** Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics* 1991; **47**(3): 1155-1163.

(124) **Kulldorff M.** A spatial scan statistic. *Communications in Statistics - Theory and Methods* 1997; **26**(6): 1481-1496.

(125) **Jacquez GM, et al.** Global, local and focused geographic clustering for case-control data with residential histories. *Environ Health* 2005; **4**(1): 4.

(126) **Bithell JF.** Estimation of relative risk functions. *Stat Med* 1991; **10**(11): 1745-1751.

(127) **Bithell JF.** An application of density estimation to geographical epidemiology. *Stat Med* 1990; **9**(6): 691-701.

(128) **Kelsall JE, Diggle PJ.** Non-parametric estimation of spatial variation in relative risk. *Stat Med* 1995; **14**(21-22): 2335-2342.

(129) **Hazelton ML, Davies TM.** Inference based on kernel estimates of the relative risk function in geographical epidemiology. *Biom J* 2009; **51**(1): 98-109.

(130) Shepard D. A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conference: ACM, 1968: 517-524.

(131) **Krige DG.** A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy* 1951; **52**(6): 119-139.

(132) **Diggle PJ.** Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Stat Methods Med Res* 2006; **15**(4): 325-336.

(133) **Diggle P, Rowlingson B, Su T-l.** Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* 2005; **16**(5): 423-434.

(134) **Sanderson RA, et al.** Spatio-temporal models to determine association between Campylobacter cases and environment. *Int J Epidemiol* 2018; **47**(1): 202-216.

(135) **Becker AD, et al.** Development and dissemination of infectious disease dynamic transmission models during the COVID-19 pandemic: what can we learn from other pathogens and how can we move forward? *The Lancet Digital Health* 2021; **3**(1): e41-e50.

(136) **Smith CM, et al.** Spatial methods for infectious disease outbreak investigations: systematic literature review. *Euro Surveill* 2015; **20**(39).

(137) **Valcour JE, et al.** A descriptive analysis of the spatio-temporal distribution of enteric diseases in New Brunswick, Canada. *BMC Public Health* 2016; **16**: 204.

(138) **Tang F, et al.** Spatio-temporal trends and risk factors for Shigella from 2001 to 2011 in Jiangsu Province, People's Republic of China. *PLoS One* 2014; **9**(1): e83487.

(139) **Gabriel E, et al.** Spatio-temporal epidemiology of Campylobacter jejuni enteritis, in an area of Northwest England, 2000-2002. *Epidemiol Infect* 2010; **138**(10): 1384-1390.

(140) **Varga C, et al.** Area-level global and local clustering of human Salmonella Enteritidis infection rates in the city of Toronto, Canada, 2007-2009. *BMC Infect Dis* 2015; **15**: 359.

(141) **Varga C, et al.** Evaluating area-level spatial clustering of Salmonella Enteritidis infections and their socioeconomic determinants in the greater Toronto area, Ontario, Canada (2007 - 2009): a retrospective population-based ecological study. *BMC Public Health* 2013; **13**: 1078.

(142) **Zhang P, et al.** [Temporal-spatial analysis of bacillary dysentery in the Three Gorges Area of China, 2005-2016]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2018; **39**(1): 47-53.

(143) **Xiao G, et al.** Spatial-temporal pattern and risk factor analysis of bacillary dysentery in the Beijing-Tianjin-Tangshan urban region of China. *BMC Public Health* 2014; **14**: 998.

(144) **Varga C, et al.** Spatial-temporal epidemiology of human Salmonella Enteritidis infections with major phage types (PTs 1, 4, 5b, 8, 13, and 13a) in Ontario, Canada, 2008-2009. *BMC Public Health* 2015; **15**: 1247.

(145) **Seixas R, et al.** Demographic characterization and spatial cluster analysis of human Salmonella 1,4,[5],12:i:- infections in Portugal: A 10year study. *J Infect Public Health* 2018; **11**(2): 178-182.

(146) **Ma Y, et al.** Spatio-Temporal Pattern and Socio-Economic Factors of Bacillary Dysentery at County Level in Sichuan Province, China. *Sci Rep* 2015; **5**: 15264.

(147) **Bowie C, et al.** Social and spatial inequalities in Rotaviral enteritis: a case for universally funded vaccination in New Zealand. *N Z Med J* 2016; **129**(1431): 59-66.

(148) **Inaida S, et al.** The spatial diffusion of norovirus epidemics over three seasons in Tokyo. *Epidemiol Infect* 2015; **143**(3): 522-528.

(149) **Lal A, et al.** A Bayesian spatio-temporal framework to identify outbreaks and examine environmental and social risk factors for infectious diseases monitored by routine surveillance. *Spatial and Spatio-temporal Epidemiology* 2018; **25**: 39-48.

(150) **Xu C, et al.** Spatial-temporal detection of risk factors for bacillary dysentery in Beijing, Tianjin and Hebei, China. *BMC Public Health* 2017; **17**(1): 743.

(151) **Xu C, et al.** Spatiotemporal Risk of Bacillary Dysentery and Sensitivity to Meteorological Factors in Hunan Province, China. *Int J Environ Res Public Health* 2017; **15**(1).

(152) **Zhang T, et al.** Bayesian spatio-temporal random coefficient time series (BaST-RCTS) model of infectious disease. *Math Biosci* 2014; **258**: 93-100.

(153) **Anon.** <Book1.pdf>.

(154) **Rivas M, et al.** Risk Factors for Shiga Toxin-Producing Escherichia coli-Associated Human Diseases. *Microbiol Spectr* 2014; **2**(5).

(155) **Strachan NJ, et al.** Whole Genome Sequencing demonstrates that Geographic Variation of Escherichia coli O157 Genotypes Dominates Host Association. *Sci Rep* 2015; **5**: 14145.

(156) **Tarr PI, Gordon CA, Chandler WL.** Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. *Lancet* 2005; **365**(9464): 1073-1086.

(157) **Mikhail AFW, et al.** An outbreak of Shiga toxin-producing Escherichia coli O157:H7 associated with contaminated salad leaves: epidemiological, genomic and food trace back investigations. *Epidemiol Infect* 2018; **146**(2): 187-196.

(158) **Moran PA.** Notes on continuous stochastic phenomena. *Biometrika* 1950; **37**(1-2): 17-23.

(159) **Varga C, et al.** Incidence, distribution, seasonality, and demographic risk factors of Salmonella Enteritidis human infections in Ontario, Canada, 2007-2009. *BMC Infect Dis* 2013; **13**: 212.

(160) **Dallman TJ, et al.** Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. *Microb Genom* 2015; **1**(3): e000029.

(161) **Health and Safety Executive**. RR678: Updating and improving the National Population Database to National Population Database 2. United Kingdom: Health and Safety Executive; 2008 2008.

(162) Office for National Statistics. 2011 Census aggregate data. . In. June 2016 ed: Office for National Statistics ; National Records of Scotland ; Northern Ireland Statistics and Research Agency., 2016.

(163) **Health and Safety Executive**. RR 297: A National Population Data Base for Major Accident Hazard Modelling. United Kingdom: Health and Safety Executive; 2005 2005.

(164) **Prince MI, et al.** The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology* 2001; **34**(6): 1083-1088.

(165) **Davies TM, Baddeley A.** Fast computation of spatially adaptive kernel estimates. *Statistics and Computing* 2018; **28**(4): 937-956.

(166) **Davies TM, Marshall JC, Hazelton ML.** Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Stat Med* 2018; **37**(7): 1191-1221.

(167) **Davies TM, Hazelton ML.** Adaptive kernel estimation of spatial relative risk. *Stat Med* 2010; **29**(23): 2423-2437.

(168) **Davies TM, Jones K, Hazelton ML.** Symmetric adaptive smoothing regimens for estimation of the spatial relative risk function. *Computational Statistics & Data Analysis* 2016; **101**: 12-28.

(169) **Abramson IS.** On Bandwidth Variation in Kernel Estimates-A Square Root Law. *Ann Statist* 1982; **10**(4): 1217-1223.

(170) **Davies TM.** Jointly optimal bandwidth selection for the planar kernel-smoothed density-ratio. *Spat Spatiotemporal Epidemiol* 2013; **5**: 51-65.

(171) **Davies TM, Lawson AB.** An evaluation of likelihood-based bandwidth selectors for spatial and spatiotemporal kernel estimates. *Journal of Statistical Computation and Simulation* 2019; **89**(7): 1131-1152.

(172) **Jones MC.** Simple boundary correction for kernel density estimation. *Statistics and Computing* 1993; **3**(3): 135-146.

(173) **Gelfand AE, et al.** Nonparametric methods. In: Guttorp P, ed. *Handbook of spatial statistics*. Boca Raton, Fla.: CRC Press, 2010.

(174) **Sarojinie Fernando WT, Hazelton ML.** Generalizing the spatial relative risk function. *Spat Spatiotemporal Epidemiol* 2014; **8**: 1-10.

(175) **Terrell GR.** The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association* 1990; **85**(410): 470-477.

(176) (ESRI) ESRI. ArcGIS Desktop. In. ArcGIS Release 10.2 ed. Redlands, CA:

, (2012).

(177) **Baddeley A, Turner R.** spatstat: An R Package for Analyzing Spatial Point Patterns. *2005* 2005; **12**(6): 42.

(178) **Baddeley A, Rubak E, Turner R**. *Spatial Point Patterns: Methodology and Applications with R*: Chapman and Hall/CRC Press, 2015.

(179) R Core Team. R: A language and environment for statistical computing. . In. Vienna, Austria: R Foundation for Statistical Computing, 2017.

(180) **Davies TM, Flynn CR, Hazelton ML.** On the utility of asymptotic bandwidth selectors for spatially adaptive kernel density estimation. *Statistics & Probability Letters* 2018; **138**: 75-81.

(181) **Locking ME, et al.** Risk factors for sporadic cases of Escherichia coli O157 infection: the importance of contact with animal excreta. *Epidemiol Infect* 2001; **127**(2): 215-220.

(182) **Brehony C, et al.** Shiga toxigenic Escherichia coli incidence is related to small area variation in cattle density in a region in Ireland. *Sci Total Environ* 2018; **637-638**: 865-870.

(183) **Innocent GT, et al.** Spatial and temporal epidemiology of sporadic human cases of Escherichia coli O157 in Scotland, 1996-1999. *Epidemiol Infect* 2005; **133**(6): 1033-1041.

(184) **Valcour JE, et al.** Associations between indicators of livestock farming intensity and incidence of human Shiga toxin-producing Escherichia coli infection. *Emerg Infect Dis* 2002; **8**(3): 252-257.

(185) **Pearl DL, et al.** A multi-level approach for investigating socio-economic and agricultural risk factors associated with rates of reported cases of Escherichia coli O157 in humans in Alberta, Canada. *Zoonoses Public Health* 2009; **56**(8): 455-464.

(186) Anonymous. National Parks - Tourism. In, 2019.

(187) **Jones NR, et al.** Novel Sampling Method for Assessing Human-Pathogen Interactions in the Natural Environment Using Boot Socks and Citizen Scientists, with Application to Campylobacter Seasonality. *Appl Environ Microbiol* 2017; **83**(14).

(188) Kintz E. Regional Differences in Presence of Shiga toxin-producing E. coli in the Environment in England. In: NIHR HPRU in Gastrointestinal infections. Liverpool.

(189) **Zhang ZJ, et al.** Identification of high-risk regions for schistosomiasis in the Guichi region of China: an adaptive kernel density estimation-based approach. *Parasitology* 2013; **140**(7): 868-875.

(190) **Byrne L, et al.** Highly Pathogenic Clone of Shiga Toxin-Producing Escherichia coli O157:H7, England and Wales. *Emerg Infect Dis* 2018; **24**(12): 2303-2308.

(191) **Kelsall JE, Diggle PJ.** Spatial Variation in Risk of Disease: A Nonparametric Binary Regression Approach. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1998; **47**(4): 559-573.

(192) **Diggle PJ, Tawn JA, Moyeed RA.** Model-Based Geostatistics. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1998; **47**(3): 299-350.

(193) **European Centre for Disease Prevention and Control.** Shigatoxin/verocytotoxin-producing Escherichia coli (STEC/VTEC) infection - Annual Epidemiological Report 2016.

(194) **Persad AK, LeJeune JT.** Animal Reservoirs of Shiga Toxin-Producing Escherichia coli. *Microbiol Spectr* 2014; **2**(4): EHEC-0027-2014.

(195) **Chapman PA, et al.** Escherichia coli O157 in cattle and sheep at slaughter, on beef and lamb carcasses and in raw beef and lamb products in South Yorkshire, UK. *Int J Food Microbiol* 2001; **64**(1-2): 139-150.

(196) **Milnes AS, et al.** Intestinal carriage of verocytotoxigenic Escherichia coli O157, Salmonella, thermophilic Campylobacter and Yersinia enterocolitica, in cattle, sheep and pigs at slaughter in Great Britain during 2003. *Epidemiol Infect* 2008; **136**(6): 739-751.

(197) **Avery SM, Moore A, Hutchison ML.** Fate of Escherichia coli originating from livestock faeces deposited directly onto pasture. *Lett Appl Microbiol* 2004; **38**(5): 355-359.

(198) **Williams AP, et al.** Persistence of Escherichia coli O157 on farm surfaces under different environmental conditions. *J Appl Microbiol* 2005; **98**(5): 1075-1083.

(199) **Peter Bibby DoTaRP, University of Sheffield and Paul Brindley, School of Computer Science, University of Nottingham.** Urban and Rural Area Definitions for Policy Purposes in England and Wales: Methodology (v1.0). 2013; **2017**(03/04/2017).

(200) Anonymous. The Private Water Supplies (England) Regulations 2016. In, 2016.

(201) **Jenks GF, Caspall FCJAotAoAG.** Error on choroplethic maps: definition, measurement, reduction. 1971; **61**(2): 217-244.

(202) **StataCorp.** Stata Statistical Software: Release 13. College Station, TX: StataCorp LP. 2013.

(203) **Milnes AS, et al.** Factors related to the carriage of Verocytotoxigenic E. coli, Salmonella, thermophilic Campylobacter and Yersinia enterocolitica in cattle, sheep and pigs at slaughter. *Epidemiol Infect* 2009; **137**(8): 1135-1148.

(204) **Mikhail AFW, et al.** An outbreak of Shiga toxin-producing Escherichia coli O157:H7 associated with contaminated salad leaves: epidemiological, genomic and food trace back investigations. *Epidemiol Infect* 2018; **146**(2): 187-196.

(205) **Cha W, et al.** Prevalence and characteristics of Shiga toxin-producing Escherichia coli in finishing pigs: Implications on public health. *Int J Food Microbiol* 2018; **264**: 8-15.

(206) **Tseng M, et al.** Shiga toxin-producing Escherichia coli in swine: the public health perspective. *Anim Health Res Rev* 2014; **15**(1): 63-75.

(207) **Chapman PA, et al.** A 1-year study of Escherichia coli O157 in cattle, sheep, pigs and poultry. *Epidemiol Infect* 1997; **119**(2): 245-250.

(208) **Nielsen EM, et al.** Verocytotoxin-producing Escherichia coli in wild birds and rodents in close proximity to farms. *Appl Environ Microbiol* 2004; **70**(11): 6944-6947.

(209) **Nichols GL.** Fly transmission of Campylobacter. *Emerg Infect Dis* 2005; **11**(3): 361-364.

(210) **Lake IR, et al.** Case-control study of environmental and social factors influencing cryptosporidiosis. *Eur J Epidemiol* 2007; **22**(11): 805-811.

(211) **Simonsen J, Frisch M, Ethelberg S.** Socioeconomic risk factors for bacterial gastrointestinal infections. *Epidemiology* 2008; **19**(2): 282-290.

(212) **Newman KL, et al.** The impact of socioeconomic status on foodborne illness in high-income countries: a systematic review. *Epidemiol Infect* 2015; **143**(12): 2473-2485.

(213) **Adams NL, et al.** Socioeconomic status and infectious intestinal disease in the community: a longitudinal study (IID2 study). *Eur J Public Health* 2018; **28**(1): 134-138.

(214) **Williams AP, et al.** Persistence, dissipation, and activity of Escherichia coli O157:H7 within sand and seawater environments. *FEMS Microbiol Ecol* 2007; **60**(1): 24-32.

(215) **Joosten R, et al.** Risk factors for gastroenteritis associated with canal swimming in two cities in the Netherlands during the summer of 2015: A prospective study. *PLoS One* 2017; **12**(4): e0174732.

(216) **Sartorius B, et al.** Outbreak of norovirus in Västra Götaland associated with recreational activities at two lakes during August 2004. *Scand J Infect Dis* 2007; **39**(4): 323-331.

(217) **Schets FM, De Roda Husman AM, Havelaar AH.** Disease outbreaks associated with untreated recreational water use. *Epidemiol Infect* 2011; **139**(7): 1114-1125.

(218) **Salmon R, Outbreak Control T.** Outbreak of verotoxin producing E.coli O157 infections involving over forty schools in south Wales, September 2005. *Euro Surveill* 2005; **10**(10): E051006 051001.

(219) **Parry SM, et al.** Risk factors for and prevention of sporadic infections with vero cytotoxin (shiga toxin) producing Escherichia coli O157. *Lancet* 1998; **351**(9108): 1019-1022.

(220) **Lake IR, et al.** Using infectious intestinal disease surveillance data to explore illness aetiology; a cryptosporidiosis case study. *Health Place* 2009; **15**(1): 333-339.

(221) **Jones CD, et al.** Escherichia coli O157: comparing awareness of rural residents and visitors in livestock farming areas. *Epidemiol Infect* 2011; **139**(10): 1522-1530.

(222) **Kintz E, et al.** Regional differences in presence of Shiga toxin-producing Escherichia coli virulence-associated genes in the environment in the North West and East Anglian regions of England. *Lett Appl Microbiol* 2020; **71**(2): 179-186.

(223) **Bell A, et al.** Assessing recall bias and measurement error in high-frequency social data collection for human-environment research. *Popul Environ* 2019; **40**: 325-345.

(224) **Behler RP, et al.** Analyzing the cost-effectiveness of Lyme disease risk reduction approaches. *J Public Health Policy* 2020; **41**(2): 155-169.

(225) **Wright L, et al.** Facilitators and barriers to compliance with COVID-19 guidelines: a structural topic modelling analysis of free-text data from 17,500 UK adults. *BMC Public Health* 2022; **22**(1): 34.

(226) **Wright L, Steptoe A, Fancourt D.** Patterns of compliance with COVID-19 preventive behaviours: a latent class analysis of 20 000 UK adults. *J Epidemiol Community Health* 2022; **76**(3): 247-253.

(227) **Wright L, Steptoe A, Fancourt D.** Trajectories of Compliance With COVID-19 Related Guidelines: Longitudinal Analyses of 50,000 UK Adults. *Ann Behav Med* 2022; **56**(8): 781-790.

(228) **Wu JS, Font X, McCamley C.** COVID-19 social distancing compliance mechanisms: UK evidence. *Environ Res* 2022; **205**: 112528.

(229) **World Health Organization.** Coronavirus disease (COVID-19) Weekly Epidemiological Update 2020.

(230) **Franch-Pardo I, et al.** Spatial analysis and GIS in the study of COVID-19. A review. *Sci Total Environ* 2020; **739**: 140033.

(231) **Anonymous.** Covid-19 Testing data: Methodology note. Accessed 22 October 2020. 2020.

(232) **Taylor C.** Bootstrap choice of the smoothing parameter in kernel density estimation. . *Biometrika* 1989(76): 705–712.

(233) **Adrian W B, Adelchi Azzalini**. *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations*: Oxford University Press Inc., New York., 1997.

(234) **Silverman BW**. Density Estimation for Statistics and Data Analysis. In. New York: Chapman & Hall, 1986.

(235) **Biggerstaff M, et al.** Early Insights from Statistical and Mathematical Modeling of Key Epidemiologic Parameters of COVID-19. *Emerg Infect Dis* 2020; **26**(11).

(236) Oliver Pybus AR, Louis du Plessis, Alexander E Zarebski, Moritz U G Kraemer, Jayna Raghwani, Bernardo Gutiérrez, Verity Hill, John McCrone, Rachel Colquhoun, Ben Jackson, Áine O'Toole, Jordan Ashworth on behalf of the COG-UK consortium. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages. In, 2020.

(237) **Public Health England**. Weekly Coronavirus Disease 2019 (COVID-19) surveillance report: Summary of COVID-19 surveillance systems.; 2020.

(238) **de Lusignan S, et al.** Disparities in the excess risk of mortality in the first wave of COVID-19: Cross sectional study of the English sentinel network. *J Infect* 2020: 4817.

(239) **Setti L, et al.** Potential role of particulate matter in the spreading of COVID-19 in Northern Italy: first observational study based on initial epidemic diffusion. *BMJ Open* 2020; **10**(9): e039338.

(240) **Konstantinoudis G, et al.** Long-term exposure to air-pollution and COVID-19 mortality in England: a hierarchical spatial analysis. *medRxiv* 2020.

(241) **Brix A, Diggle PJ.** Spatiotemporal Prediction for Log-Gaussian Cox Processes. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2001; **63**(4): 823-841.

(242) **Fry R, et al.** Real-time spatial health surveillance: mapping the UK COVID-19 epidemic. 2020: 2020.2008.2017.20175117.

(243) **Fronterre C, et al.** COVID-19 in England: spatial patterns and regional outbreaks. 2020: 2020.2005.2015.20102715.

(244) **Dummer TJ.** Health geography: supporting public health policy and planning. *Cmaj* 2008; **178**(9): 1177-1180.

(245) Anonymous. Royal Mail UK Postcodes Explained. In, 2022.

(246) **Cordery R, et al.** Transmission of SARS-CoV-2 by children to contacts in schools and households: a prospective cohort and environmental sampling study in London. *Lancet Microbe* 2022; **3**(11): e814-e823.

(247) **Merrick R, et al.** Differential impact of quarantine policies for recovered COVID-19 cases in England: a case cohort study of surveillance data, June to December 2020. *BMC Public Health* 2022; **22**(1): 1915.

(248) **Swanson J, Vincent TJ, Bunch KJ.** Relative accuracy of grid references derived from postcode and address in UK epidemiological studies of overhead power lines. *J Radiol Prot* 2014; **34**(4): N81-86.

(249) **Office for National Statistics.** Changes to Output

Areas and Super

Output Areas in

England and Wales,

2001 to 2011. 2012.

(250) **Baynes J, Neale A, Hultgren T.** Improving intelligent dasymetric mapping population density estimates at 30 m resolution for the conterminous United States by excluding uninhabited areas. *Earth Syst Sci Data* 2022; **14**(6): 2833-2849.

(251) **Bergroth C, et al.** A 24-hour population distribution dataset based on mobile phone data from Helsinki Metropolitan Area, Finland. *Sci Data* 2022; **9**(1): 39.

(252) **Depsky NJ, Cushing L, Morello-Frosch R.** High-resolution gridded estimates of population sociodemographics from the 2020 census in California. *PLoS One* 2022; **17**(7): e0270746.

(253) **Schug F, et al.** Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates. *PLoS One* 2021; **16**(3): e0249044.

(254) **Swanwick RH, et al.** Dasymetric population mapping based on US census data and 30-m gridded estimates of impervious surface. *Sci Data* 2022; **9**(1): 523.

(255) **Vachon MS, et al.** Farm animal contact is associated with progression to Hemolytic uremic syndrome in patients with Shiga toxin-producing Escherichia coli - Indiana, 2012-2018. *One Health* 2020; **11**: 100175.

(256) **Diggle PJ**. *Statistical analysi of spatial and sptio-temporal point patterns. Third Edition*. Third ed: CRC Press, 2013.

(257) **Hughes GJ, Gorton R.** Inequalities in the incidence of infectious disease in the North East of England: a population-based study. *Epidemiol Infect* 2015; **143**(1): 189-201.

(258) **Tam CC, Rodrigues LC, O'Brien SJ.** The study of infectious intestinal disease in England: what risk factors for presentation to general practice tell us about potential for selection bias in case-control studies of reported cases of diarrhoea. *Int J Epidemiol* 2003; **32**(1): 99-105.

(259) **Wilson A, et al.** Characterisation of strains of enteroaggregative Escherichia coli isolated during the infectious intestinal disease study in England. *Eur J Epidemiol* 2001; **17**(12): 1125-1130.

(260) **Hsia J, et al.** Estimating bias from internet non-use for a hybrid web vaccination survey - 2013-2017 Behavioral Risk Factor Surveillance System. *Vaccine* 2021; **39**(30): 4160-4165.

(261) **Marcano Belisario JS, et al.** Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. *Cochrane Database Syst Rev* 2015; **2015**(7): Mr000042.

(262) **Carrasco-Escobar G, et al.** Open-Source 3D Printable GPS Tracker to Characterize the Role of Human Population Movement on Malaria Epidemiology in River Networks: A Proof-of-Concept Study in the Peruvian Amazon. *Front Public Health* 2020; **8**: 526468.

(263) **Mitchell A, Bourn, D.,Mawdsley, J., Wint,W, Clifton-Hadley, R. and Gilbert, M.** Characteristics of cattle movements in Britain - an analysis of records from the Cattle Tracing System. *Animal Science* 2005; **80**: 265-273.

(264) **Chase-Topping ME, et al.** Risk factors for the presence of high-level shedders of Escherichia coli O157 on Scottish farms. *J Clin Microbiol* 2007; **45**(5): 1594-1603.

(265) **Matthews L, et al.** Heterogeneous shedding of Escherichia coli O157 in cattle and its implications for control. *Proc Natl Acad Sci U S A* 2006; **103**(3): 547-552.

(266) **Mather AE, et al.** Factors associated with cross-contamination of hides of Scottish cattle by Escherichia coli O157. *Appl Environ Microbiol* 2008; **74**(20): 6313-6319.

(267) Anonymous. Investigation of SARS-CoV-2 variants: technical briefings. In, 2022.

(268) **McCarthy SC, et al.** An Overview of Shiga-Toxin Producing Escherichia coli Carriage and Prevalence in the Ovine Meat Production Chain. *Foodborne Pathog Dis* 2021; **18**(3): 147-168.

(269) **McCarthy SC, et al.** Prevalence and Whole-Genome Sequence-Based Analysis of Shiga Toxin-Producing Escherichia coli Isolates from the Recto-Anal Junction of Slaughter-Age Irish Sheep. *Appl Environ Microbiol* 2021; **87**(24): e0138421.

(270) **Vishram B, et al.** The emerging importance of Shiga toxin-producing Escherichia coli other than serogroup O157 in England. *J Med Microbiol* 2021; **70**(7).