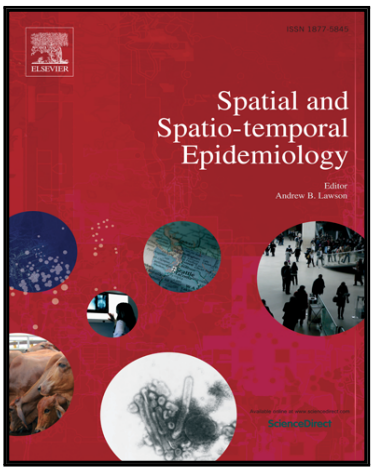
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of East Anglia digital repository

Journal Pre-proof



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

Richard Elson , Tilman M. Davies , Claire Jenkins , Roberto Vivancos , Sarah J. O'Brien , Iain R. Lake

PII: S1877-5845(19)30115-7
DOI: https://doi.org/10.1016/j.sste.2019.100305
Reference: SSTE 100305

To appear in: *Spatial and Spatio-temporal Epidemiology*

Received date: 31 January 2019
Revised date: 10 September 2019
Accepted date: 16 September 2019

Please cite this article as: Richard Elson , Tilman M. Davies , Claire Jenkins , Roberto Vivancos , Sarah J. O'Brien , Iain R. Lake , Application of kernel smoothing to estimate the spatio-temporal variation in risk of STEC O157 in England, *Spatial and Spatio-temporal Epidemiology* (2019), doi: https://doi.org/10.1016/j.sste.2019.100305

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 The Author(s). Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

Journal Pre-proof

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

Richard Elson [1,2,6], Tilman M. Davies [3], Claire Jenkins [1,2], Roberto Vivancos [1,2,4], Sarah J. O'Brien [2,5] and Iain R. Lake [2,6].

1. National Infection Service, Public Health England, United Kingdom.
2. National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections, United Kingdom.
3. Department of Mathematics & Statistics, University of Otago, Dunedin, New Zealand.
4. National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections, United Kingdom.
5. Institute of Population Health Sciences, University of Liverpool, Liverpool, United Kingdom.
6. School of Environmental Sciences, University of East Anglia, United Kingdom.

Author for correspondence:

Richard Elson

National Infection Service, Public Health England, 61 Colindale Avenue, London, NW9 5EQ.

Email: richard.elson@phe.gov.uk

Journal Pre-proof

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; *stx*1 and *stx*2 [1]. The main reservoir is ruminant animals, particularly cows and sheep.

First recognised as a human pathogen in 1982 [2], STEC are now globally distributed [3]. 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 [4]. 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 complex set of interactions between distal and proximal risk factors related to the reservoir, the environment, the pathogen, the host and opportunities for transmission [5]. The relative importance of these factors may vary at different spatial scales [6]. 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 [5]. However, these factors alone are unlikely to explain the considerable variation of infection rates between [7-11], and within [12], countries around the world, particularly when considering the comparable levels of carriage by cattle in those countries [13].

Within the United Kingdom, rates of STEC infection in Scotland are more than twice that of England [14]. 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 [6, 15] and there is evidence that this relates to living in areas with high densities of farmed animals [6]. However, the strains infecting humans are not always the same as those circulating in the 'local' ruminant reservoir [16, 17]. 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 [6,

17]. Conversely, evidence from outbreak investigations shows that transmission of highly related strains can occur via multiple routes from geographically restricted sources [18, 19].

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 [20] and the Diggle-Chetwynd statistic [21], 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) [22] and Kuldorff's scan statistic [23], 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 [24-27], and spatial interpolation methods such as inverse distance weighting [28] and kriging [29]. Modelling approaches can either take the form of empirical or mechanistic models that consider the effect of space and time alongside other factors [30-32].

Journal Pre-proof

Smith et al. [33] 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 [34-38] and local [34, 35, 37, 39-44] clustering, spatial variation in risk [36, 45, 46], modelling and other approaches [31, 35, 36, 39, 41, 42, 45, 47, 48].

The aims of this study were to first estimate the spatial variation in risk of STEC O157 in England; second, to estimate the space-time variation in risk over the study period; and, finally to explore any difference between the residential locations of cases reporting travel and those not reporting travel.

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 [15]. 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 [15]. Since 2015, all isolates have been routinely sequenced allowing identification of genetic lineage/sub-lineage and *stx* subtypes [49, 50].

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 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].

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 [49]. 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 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) [51]. 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 [52]. 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 [51, 53]. 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 [54]. 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 [55].

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) [55]. A fixed or adaptive [56, 57] 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 [26, 27, 56].

Journal Pre-proof

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 [58]. 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 [56, 57, 59], 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 [26] offers both theoretical and practical benefits for the resulting risk function estimate.

As such, we follow 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 [57], using equal global and pilot bandwidths chosen simultaneously via the likelihood cross-validation methodology described in [60]. The global bandwidth value was used for the fixed estimate in the sensitivity analyses. The far-right hand column of Table 1 reports the common case/control bandwidth found for each estimate.

Journal Pre-proof

All estimates are edge-corrected to account for kernel weight lost over the boundary of the study region [61, 62] 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 [56, 57], 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 [6]. 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 [63]. 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 [63] using two bandwidth prescriptions. The first used the maximal smoothing principle proposed by Terrell [64] applied separately to the spatial and temporal margins of the data. The second used the fixed bandwidth cross-validated likelihood method [60] 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 [65]. All subsequent analyses were performed using the contributed packages sparr [55] and spatstat [66, 67] in the R language [68]. Bandwidth selections were performed using cases and/or controls falling within a simplified polygon of the mainland boundary of England.

Results

The spatial locations of all unmarked cases and controls are shown in Figure 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 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 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.

Journal Pre-proof

The relative risk surface for all cases (including those reporting travel) is shown in Figure 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.

The relative risk surfaces for Lineages I, II and I/II are presented in Figure 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 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 here (Insert link to MP4). 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

Journal Pre-proof

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 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 Figures S1 and S2 respectively. 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 [59].

Journal Pre-proof

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 [6, 69-71]. 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 [6, 14, 72, 73]. There is evidence that the spatial distribution and relative importance of risk factors differ by pathogen sub-type [6, 45, 73] and similar findings have been produced from Northern European countries [14, 74-79], the United States [45], Canada [74, 80] and New Zealand[12].

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 [6]. 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 [81]. 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 [81, 82].

The importance of the pathways through which pathogens are transferred from the environment to humans is subject to debate [82].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* [82] and STEC [83] 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 [82].

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 [15] and our analysis demonstrates that these strains are also spatially restricted. Lineages I and II have dominated since the late nineties [15] 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[6]. 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 [82].

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 a long period of time in the over smoothed estimate (left panel in animation) 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), 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[6].

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 [84], classical fixed bandwidth estimators can be unstable and do not cope well with the smoothing requirements of highly heterogeneous patterns [55, 56, 85]. However, choosing appropriate smoothing parameters for the more sophisticated adaptive estimator is far more difficult, and this is an active area of research [59, 85] .

We used a recently developed likelihood-based selection strategy for the purely spatial analyses [60], 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 *stx*2), only emerged in significant numbers following an outbreak towards the end of 2015 [19, 86] 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 [6].

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 [87]. 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 [15]. However, a large petting farm outbreak in 2009 [88] 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.

In addition, the Health Protection (Notification) Regulations 2010 [89] 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.

Journal Pre-proof

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, 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 [6]. 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 [90, 91]) 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.

Acknowledgements

The authors are indebted to two anonymous reviewers who provided insightful feedback on the first draft of this manuscript, which in turn greatly improved the final presentation.

Journal Pre-proof

We thank everyone involved in the surveillance of gastrointestinal infections in England. Thanks are due to Lisa Byrne, Bhavita Vishram, Saira Butt, Nalini Purohit and Mike Harte who lead and manage the enhanced surveillance system for STEC in England, and Jack Wardle who proof read the final manuscript. We are grateful to the microbiologists and health protection and environmental health specialists who have contributed data and reports to national surveillance systems, and the epidemiologists and information officers who have worked on the national surveillance of intestinal infectious diseases for the National Infection Service of Public Health England. ONS census information is licensed under the terms of the Open Government Licence [http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3]. Finally, thanks are due to the R community who share their knowledge, skills and intellect freely.

The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with University of East Anglia, University of Oxford and the Institute of Food Research. Tilman M. Davies was supported in part by the Royal Society of New Zealand, Marsden Fast-start grant 15-UOO-092.

Richard Elson is based at Public Health England. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England.

Biographical sketch

Richard Elson is an epidemiologist employed by Public Health England with a particular interest in the spatial and spatio-temporal distribution of gastrointestinal infections.

Journal Pre-proof

Table 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.

Journal Pre-proof

Journal Pre-proof

Journal Pre-proof

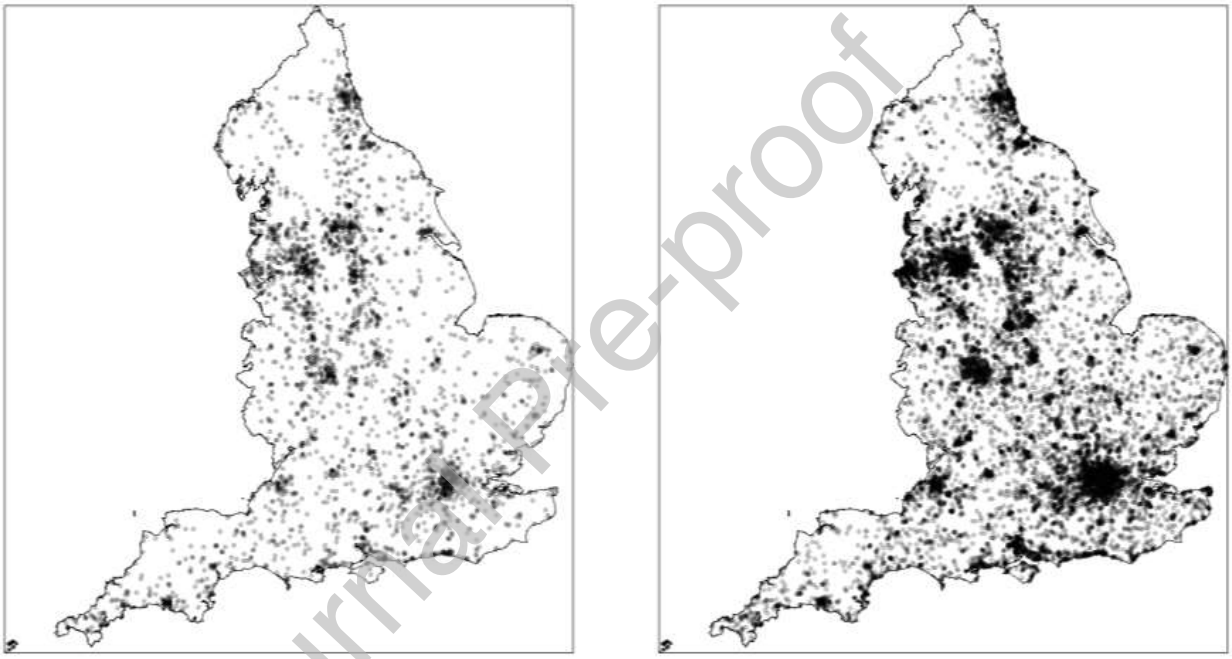


Figure 1 Spatial location of 3,592 STEC O157 cases (left panel) and 14,392 randomly selected controls (right panel).

Journal Pre-proof

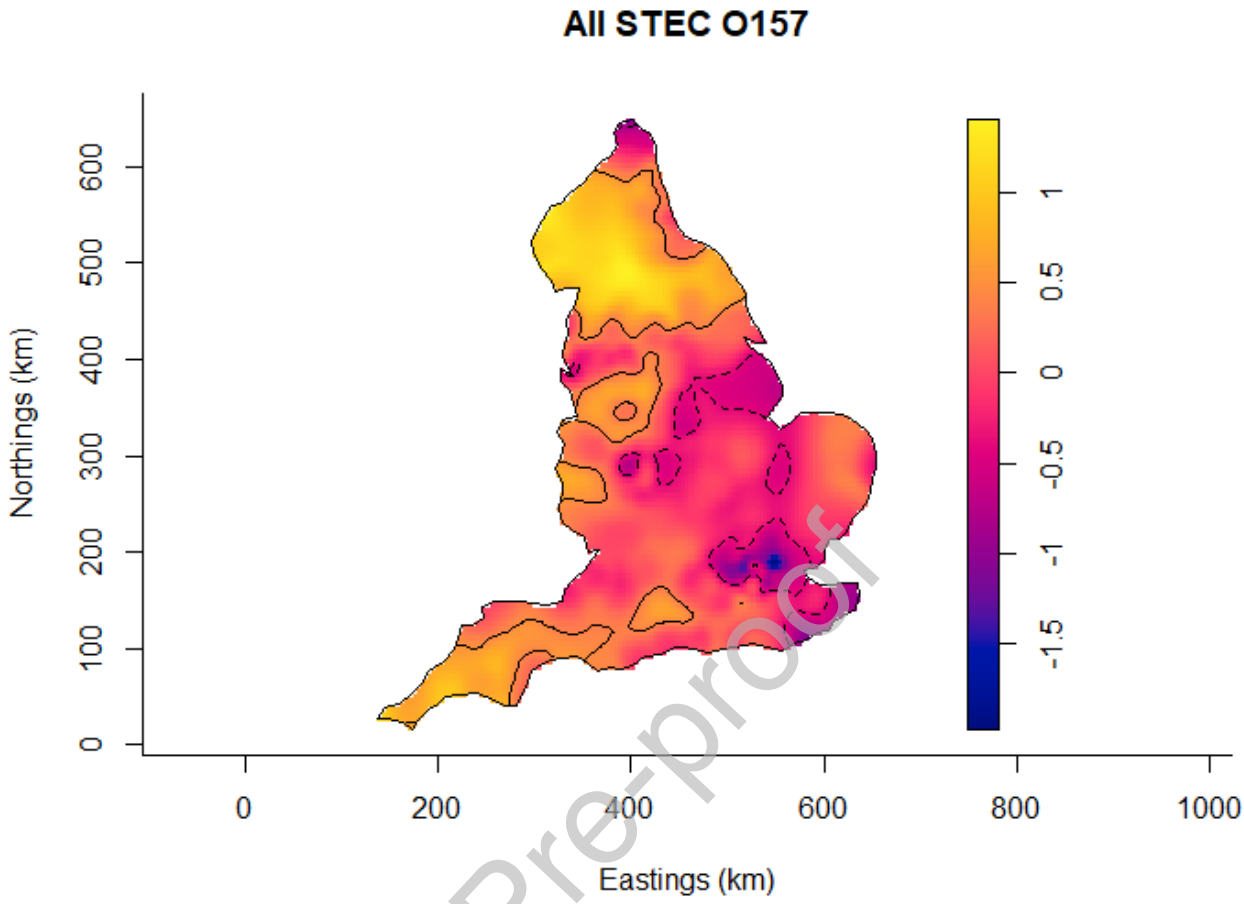


Figure 2. Estimated log relative risk for all cases of STEC O157 (including cases reporting travel). Tolerance contours are superimposed as solid lines at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk.

Journal Pre-proof

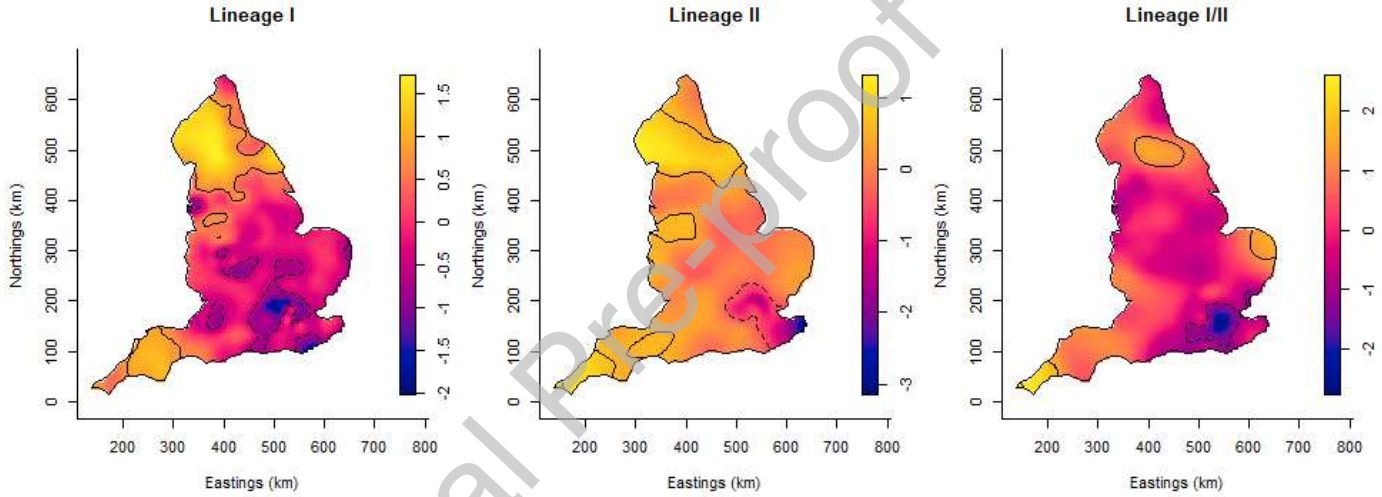


Figure 3. Estimated log relative risk for STEC O157 Lineages I, II and I/II in cases reporting no travel. Darker areas indicate areas of lower risk. Tolerance contours are superimposed as solid lines at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk.

Journal Pre-proof

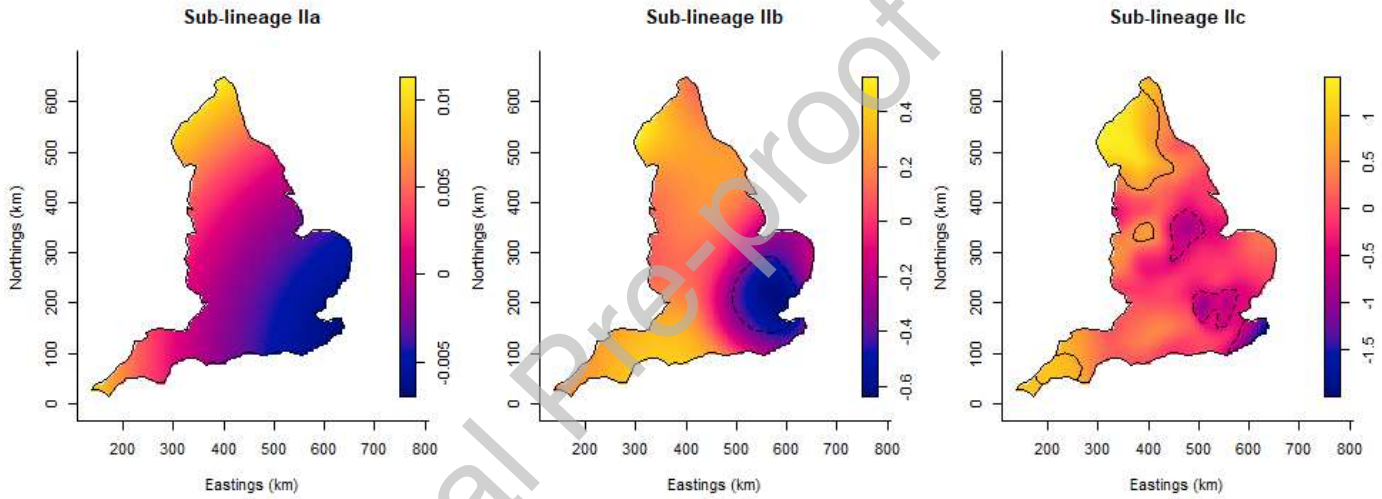


Figure 4. Estimated log relative risk for STEC O157 sub-lineages of Lineage II in cases reporting no travel. Darker areas indicate areas of lower risk. Tolerance contours are superimposed as solid lines at the 95% confidence level. Solid lines indicate areas of significantly

Journal Pre-proof

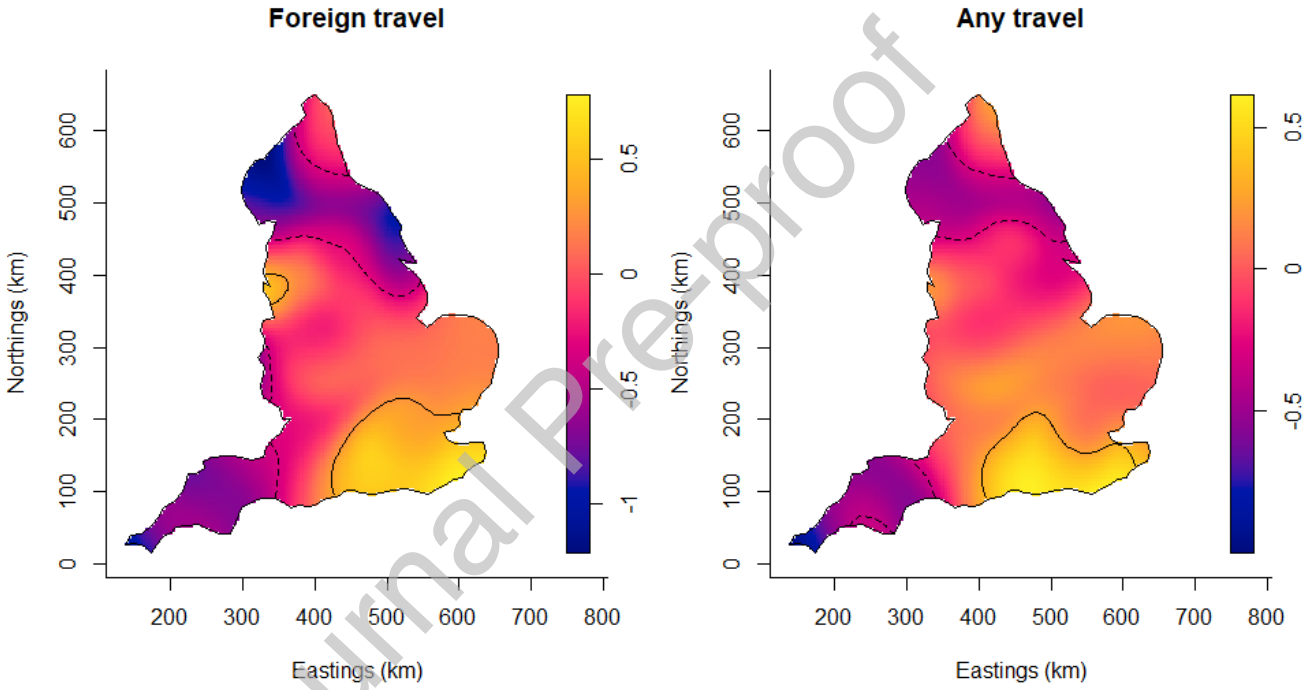


Figure 5. Estimated log relative risk for cases of STEC O157 reporting foreign travel (left panel) and those reporting any travel. Tolerance contours are superimposed at the 95% confidence level. Solid lines indicate areas of significantly higher risk and dashed lines indicate areas of lower risk higher risk and dashed lines indicate areas of lower risk.

Journal Pre-proof

References

1. Byrne L, Jenkins C, Launders N, Elson R, Adak GK: **The epidemiology, microbiology and clinical impact of Shiga toxin-producing Escherichia coli in England, 2009-2012**. *Epidemiology and infection* 2015, **143**(16):3475-3487.
2. Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, Hebert RJ, Olcott ES, Johnson LM, Hargrett NT *et al*: **Hemorrhagic colitis associated with a rare Escherichia coli serotype**. *N Engl J Med* 1983, **308**(12):681-685.
3. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M: **Super-shedding and the link between human infection and livestock carriage of Escherichia coli O157**. *Nat Rev Microbiol* 2008, **6**(12):904-912.
4. Davis MA, Hancock DD, Besser TE, Rice DH, Hovde CJ, Digiacomo R, Samadpour M, Call DR: **Correlation between geographic distance and genetic similarity in an international collection of bovine faecal Escherichia coli O157:H7 isolates**. *Epidemiology and infection* 2003, **131**(2):923-930.
5. Lal A, Hales S, French N, Baker MG: **Seasonality in human zoonotic enteric diseases: a systematic review**. *PloS one* 2012, **7**(4):e31883.
6. 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, **146**(15):1928-1939.
7. Vally H, Hall G, Dyda A, Raupach J, Knope K, Combs B, Desmarchelier P: **Epidemiology of Shiga toxin producing Escherichia coli in Australia, 2000-2010**. *BMC public health* 2012, **12**:63.
8. Locking ME, Pollock KG, Allison LJ, Rae L, Hanson MF, Cowden JM: **Escherichia coli O157 infection and secondary spread, Scotland, 1999-2008**. *Emerging infectious diseases* 2011, **17**(3):524-527.
9. Rivas M, Miliwebsky E, Chinen I, Deza N, Leotta GA: **[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.
10. Tarr PI, Gordon CA, Chandler WL: **Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome**. *Lancet (London, England)* 2005, **365**(9464):1073-1086.
11. Strachan NJ, Rotariu O, Lopes B, MacRae M, Fairley S, Laing C, Gannon V, Allison LJ, Hanson MF, Dallman T *et al*: **Whole Genome Sequencing demonstrates that Geographic Variation of Escherichia coli O157 Genotypes Dominates Host Association**. *Sci Rep* 2015, **5**:14145.
12. Jaros P, Cookson AL, Campbell DM, Duncan GE, Prattley D, Carter P, Besser TE, Shringi S, Hathaway S, Marshall JC *et al*: **Geographic divergence of bovine and human Shiga toxin-producing Escherichia coli O157:H7 genotypes, New Zealand**. *Emerging infectious diseases* 2014, **20**(12):1980-1989.
13. Mellor GE, Fegan N, Gobius KS, Smith HV, Jennison AV, D'Astek BA, Rivas M, Shringi S, Baker KN, Besser TE: **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.
14. Innocent GT, Mellor DJ, McEwen SA, Reilly WJ, Smallwood J, Locking ME, Shaw DJ, Michel P, Taylor DJ, Steele WB *et al*: **Spatial and temporal epidemiology of sporadic human cases of Escherichia coli O157 in Scotland, 1996-1999**. *Epidemiology and infection* 2005, **133**(6):1033-1041.
15. Adams NL, Byrne L, Smith GA, Elson R, Harris JP, Salmon R, Smith R, O'Brien SJ, Adak GK, Jenkins C: **Shiga Toxin-Producing Escherichia coli O157, England and Wales, 1983-2012**. *Emerging infectious diseases* 2016, **22**(4):590-597.
16. Franz E, van Hoek AH, van der Wal FJ, de Boer A, Zwartkruis-Nahuis A, van der Zwaluw K, Aarts HJ, Heuvelink AE: **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.

17. Food Standards Scotland: **E. coli O157 Super-shedding in Cattle and Mitigation of Human Risk**. 2018. Available at: https://www.foodstandards.gov.scot/downloads/Super-shedders_-_FINAL_version_for_publication.pdf. Last accessed: 6th September 2019.
18. Butcher H, Elson R, Chattaway MA, Featherstone CA, Willis C, Jorgensen F, Dallman TJ, Jenkins C, Mc LJ, Beck CR *et al*: **Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing Escherichia coli O157 associated with raw drinking milk**. *Epidemiology and infection* 2016, **144**(13):2812-2823.
19. Mikhail AFW, Jenkins C, Dallman TJ, Inns T, Douglas A, Martin AIC, Fox A, Cleary P, Elson R, Hawker J: **An outbreak of Shiga Toxin-producing Escherichia coli O157:H7 associated with contaminated salad leaves: epidemiological, genomic and food trace back investigations - CORRIGENDUM**. *Epidemiology and infection* 2018, **146**(14):1879.
20. Moran PA: **Notes on continuous stochastic phenomena**. *Biometrika* 1950, **37**(1-2):17-23.
21. Diggle PJ, Chetwynd AG: **Second-order analysis of spatial clustering for inhomogeneous populations**. *Biometrics* 1991, **47**(3):1155-1163.
22. Anselin L: **Local Indicators of Spatial Association—LISA**. *Geographical Analysis* 1995, **27**(2):93-115.
23. Kulldorff M: **A spatial scan statistic**. *Communications in Statistics - Theory and Methods* 1997, **26**(6):1481-1496.
24. Bithell JF: **Estimation of relative risk functions**. *Stat Med* 1991, **10**(11):1745-1751.
25. Bithell JF: **An application of density estimation to geographical epidemiology**. *Stat Med* 1990, **9**(6):691-701.
26. Kelsall JE, Diggle PJ: **Non-parametric estimation of spatial variation in relative risk**. *Stat Med* 1995, **14**(21-22):2335-2342.
27. Hazelton ML, Davies TM: **Inference based on kernel estimates of the relative risk function in geographical epidemiology**. *Biom J* 2009, **51**(1):98-109.
28. Shepard D: **A two-dimensional interpolation function for irregularly-spaced data**. In: *Proceedings of the 1968 23rd ACM national conference.* ACM; 1968: 517-524.
29. 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.
30. Diggle P, Rowlingson B, Su T-l: **Point process methodology for on-line spatio-temporal disease surveillance**. *Environmetrics* 2005, **16**(5):423-434.
31. Sanderson RA, Maas JA, Blain AP, Gorton R, Ward J, O'Brien SJ, Hunter PR, Rushton SP: **Spatio-temporal models to determine association between Campylobacter cases and environment**. *International journal of epidemiology* 2018, **47**(1):202-216.
32. Diggle PJ: **Spatio-temporal point processes, partial likelihood, foot and mouth disease**. *Stat Methods Med Res* 2006, **15**(4):325-336.
33. Smith CM, Le Comber SC, Fry H, Bull M, Leach S, Hayward AC: **Spatial methods for infectious disease outbreak investigations: systematic literature review**. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2015, **20**(39).
34. Valcour JE, Charron DF, Berke O, Wilson JB, Edge T, Waltner-Toews D: **A descriptive analysis of the spatio-temporal distribution of enteric diseases in New Brunswick, Canada**. *BMC public health* 2016, **16**:204.
35. Tang F, Cheng Y, Bao C, Hu J, Liu W, Liang Q, Wu Y, Norris J, Peng Z, Yu R *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.
36. Gabriel E, Wilson DJ, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart CA, Diggle PJ: **Spatio-temporal epidemiology of Campylobacter jejuni enteritis, in an area of Northwest England, 2000-2002**. *Epidemiology and infection* 2010, **138**(10):1384-1390.

Journal Pre-proof

37. Varga C, Pearl DL, McEwen SA, Sargeant JM, Pollari F, Guerin MT: **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.
38. Varga C, Pearl DL, McEwen SA, Sargeant JM, Pollari F, Guerin MT: **Area-level global and local clustering of human Salmonella Enteritidis infection rates in the city of Toronto, Canada, 2007-2009**. *BMC infectious diseases* 2015, **15**:359.
39. Zhang T, Zhang X, Ma Y, Zhou XA, Liu Y, Feng Z, Li X: **Bayesian spatio-temporal random coefficient time series (BaST-RCTS) model of infectious disease**. *Mathematical biosciences* 2014, **258**:93-100.
40. Bowie C, Campbell M, Beere P, Kingham S: **Social and spatial inequalities in Rotaviral enteritis: a case for universally funded vaccination in New Zealand**. *The New Zealand medical journal* 2016, **129**(1431):59-66.
41. Ma Y, Zhang T, Liu L, Lv Q, Yin F: **Spatio-Temporal Pattern and Socio-Economic Factors of Bacillary Dysentery at County Level in Sichuan Province, China**. *Sci Rep* 2015, **5**:15264.
42. Xiao G, Xu C, Wang J, Yang D, Wang L: **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.
43. Seixas R, Nunes T, Machado J, Tavares L, Owen SP, Bernardo F, Oliveira M: **Demographic characterization and spatial cluster analysis of human Salmonella 1,4,[5],12:i:- infections in Portugal: A 10year study**. *Journal of infection and public health* 2018, **11**(2):178-182.
44. Varga C, Pearl DL, McEwen SA, Sargeant JM, Pollari F, Guerin MT: **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.
45. Tarr GAM, Shringi S, Phipps AI, Besser TE, Mayer J, Oltean HN, Wakefield J, Tarr PI, Rabinowitz P: **Geogenomic Segregation and Temporal Trends of Human Pathogenic Escherichia coli O157:H7, Washington, USA, 2005-2014(1)**. *Emerging infectious diseases* 2018, **24**(1):32-39.
46. Inaida S, Shobugawa Y, Matsuno S, Saito R, Suzuki H: **The spatial diffusion of norovirus epidemics over three seasons in Tokyo**. *Epidemiology and infection* 2015, **143**(3):522-528.
47. Xu C, Xiao G, Wang J, Zhang X, Liang J: **Spatiotemporal Risk of Bacillary Dysentery and Sensitivity to Meteorological Factors in Hunan Province, China**. *International journal of environmental research and public health* 2017, **15**(1).
48. Lal A, Marshall J, Benschop J, Brock A, Hales S, Baker MG, French NP: **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.
49. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ *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.
50. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A *et al*: **Whole-genome sequencing for national surveillance of Shiga toxin-producing Escherichia coli O157**. *Clin Infect Dis* 2015, **61**(3):305-312.
51. Health and Safety Executive: **RR678: Updating and improving the National Population Database to National Population Database 2**. In. United Kingdom:; 2008. Available at: http://www.hse.gov.uk/research/rrpdf/rr678.pdf . Last accessed: 6th September 2019 .
52. Office for National Statistics: **2011 Census aggregate data.** June 2016. Available at: http://dx.doi.org/10.5257/census/aggregate-2011-1. Last accessed: 6th September 2019.
53. Health and Safety Executive: **RR 297: A National Population Data Base for Major Accident Hazard Modelling**. In. United Kingdom: Health and Safety Executive; 2005. Available at: http://www.hse.gov.uk/research/rrpdf/rr297.pdf . Last accessed: 6th September 2019 .

54. Prince MI, Chetwynd A, Diggle P, Jarner M, Metcalf JV, James OF: **The geographical distribution of primary biliary cirrhosis in a well-defined cohort**. *Hepatology* 2001, **34**(6):1083-1088.
55. Davies TM, Marshall JC, Hazelton ML: **Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk**. *Stat Med* 2018, **37**(7):1191-1221.
56. Davies TM, Hazelton ML: **Adaptive kernel estimation of spatial relative risk**. *Stat Med* 2010, **29**(23):2423-2437.
57. 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.
58. Abramson IS: **On Bandwidth Variation in Kernel Estimates-A Square Root Law**. *Ann Statist* 1982, **10**(4):1217-1223.
59. Davies TM, Baddeley A: **Fast computation of spatially adaptive kernel estimates**. *Statistics and Computing* 2018, **28**(4):937-956.
60. 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.
61. Jones MC: **Simple boundary correction for kernel density estimation**. *Statistics and Computing* 1993, **3**(3):135-146.
62. Gelfand AE, Diggle PJ, Fuentes M, Guttorp P: **Nonparametric methods**. In: *Handbook of spatial statistics.* Edited by Guttorp P. Boca Raton, Fla.: CRC Press; 2010.
63. Sarojinie Fernando WT, Hazelton ML: **Generalizing the spatial relative risk function**. *Spat Spatiotemporal Epidemiol* 2014, **8**:1-10.
64. Terrell GR: **The Maximal Smoothing Principle in Density Estimation**. *Journal of the American Statistical Association* 1990, **85**(410):470-477.
65. Environmental Systems Research Institute (ESRI): **ArcGIS Deskto**p Release 10.2. 2012. Redlands, California.
66. Baddeley A, Turner R: **spatstat: An R Package for Analyzing Spatial Point Patterns**. *Journal of Statistical Software* 2005, **12**(6):42.
67. Baddeley A, Rubak E, Turner R: **Spatial Point Patterns: Methodology and Applications with R**: Chapman and Hall/CRC Press; 2015.
68. R Core Team (2017). **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
69. Locking ME, O'Brien SJ, Reilly WJ, Wright EM, Campbell DM, Coia JE, Browning LM, Ramsay CN: **Risk factors for sporadic cases of Escherichia coli O157 infection: the importance of contact with animal excreta**. *Epidemiology and infection* 2001, **127**(2):215-220.
70. Kintz E, Brainard J, Hooper L, Hunter P: **Transmission pathways for sporadic Shiga-toxin producing E. coli infections: A systematic review and meta-analysis**. *International journal of hygiene and environmental health* 2017, **220**(1):57-67.
71. 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**. *Emerging infectious diseases* 2001, **7**(6):1049-1051.
72. Brehony C, Cullinan J, Cormican M, Morris D: **Shiga toxigenic Escherichia coli incidence is related to small area variation in cattle density in a region in Ireland**. *The Science of the total environment* 2018, **637-638**:865-870.
73. Ohaiseadha C, Hynds PD, Fallon UB, O'Dwyer J: **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.

Journal Pre-proof

74. Valcour JE, Michel P, McEwen SA, Wilson JB: **Associations between indicators of livestock farming intensity and incidence of human Shiga toxin-producing Escherichia coli infection**. *Emerging infectious diseases* 2002, **8**(3):252-257.
75. Kistemann T, Zimmer S, Vagsholm I, Andersson Y: **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.
76. Frank C, Kapfhammer S, Werber D, Stark K, Held L: **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.
77. Friesema IH, Van De Kassteele J, CM DEJ, Heuvelink AE, Van Pelt W: **Geographical association between livestock density and human Shiga toxin-producing Escherichia coli O157 infections**. *Epidemiology and infection* 2011, **139**(7):1081-1087.
78. Jalava K, Ollgren J, Eklund M, Siitonen A, Kuusi M: **Agricultural, socioeconomic and environmental variables as risks for human verotoxigenic Escherichia coli (VTEC) infection in Finland**. *BMC infectious diseases* 2011, **11**:275.
79. Haus-Cheymol R, Espie E, Che D, Vaillant V, H DEV, Desenclos JC: **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.
80. Pearl DL, Louie M, Chui L, Dore K, Grimsrud KM, Martin SW, Michel P, Svenson LW, McEwen SA: **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 and public health* 2009, **56**(8):455-464.
81. National Parks.**Our challenges:Tourism.** Available at:https://nationalparks.uk/students/ourchallenges/tourism. Last accessed 6th September 2019.
82. Jones NR, Millman C, van der Es M, Hukelova M, Forbes KJ, Glover C, Haldenby S, Hunter PR, Jackson K, O'Brien SJ *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).
83. Kintz E: **Regional Differences in Presence of Shiga toxin-producing E. coli in the Environment in England.** *NIHR HPRU in Gastrointestinal infections Conference.* Liverpool. March 2017.
84. Davies TM: **Jointly optimal bandwidth selection for the planar kernel-smoothed density-ratio**. *Spat Spatiotemporal Epidemiol* 2013, **5**:51-65.
85. 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.
86. Byrne L, Dallman TJ, Adams N, Mikhail AFW, McCarthy N, Jenkins C: **Highly Pathogenic Clone of Shiga Toxin-Producing Escherichia coli O157:H7, England and Wales**. *Emerging infectious diseases* 2018, **24**(12):2303-2308.
87. Tam CC, Rodrigues LC, Viviani L, Dodds JP, Evans MR, Hunter PR, Gray JJ, Letley LH, Rait G, Tompkins DS *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.
88. Ihekweazu C, Carroll K, Adak B, Smith G, Pritchard GC, Gillespie IA, Verlander NQ, Harvey-Vince L, Reacher M, Edeghere O *et al*: **Large outbreak of verocytotoxin-producing Escherichia coli O157 infection in visitors to a petting farm in South East England, 2009**. *Epidemiology and infection* 2012, **140**(8):1400-1413.
89. The Stationery Office Limited: **The Health Protection (Notification) Regulations 2010**. UK:; 2010.
90. 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.

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

Journal Pre-proof