Mathematics for Real-World Systems CDT
2016

Detection of campylobacter outbreaks using genomics and epidemiological data

L. Guzmán*, N. McCarthy*, S.E.F. Spencer*

Abstract
Campylobacter is the most common bacteria causing gastroenteritis in UK. Although near 280 thousand cases per year are estimated, detected outbreaks are relatively rare. Therefore, mathematical approaches are suitable when exploring notification data to identify epidemics. This report proposes four different methods for detecting potential outbreaks. Three of them are based on the spatio-temporal distribution of reported cases, whilst the last one uses the genetic distances among Campylobacter samples. A sentinel surveillance dataset containing 743 incidents in Oxfordshire in the period April 2010 - March 2011 is used, dataset funded and collected by Public Health England and the Food Standards Agency. Information about spatial and temporal locations is included, as well as the structure of the whole genome. Although some outbreaks are identified in the spatio-temporal analysis, a genome-to-genome comparison among cases is not conclusive. Results obtained based on a genetic distance analysis suggest that the methodology potentially detects outbreaks but future testing and improvements are required.

Keywords
Outbreak detection — Campylobacter — Spatio-temporal clustering ... Genetic distance clustering.

*Mathematics of Real-World Systems Centre for Doctoral Training, University of Warwick, Coventry CV4 7AL, United Kingdom

Contents

1. Introduction

Bacteria are always present in the human body. Some of them are harmless to humans while others are the main causes of diseases, especially when transmitted through consumption of contaminated food or contact with infected animals. In UK, the most common bacteria implicated in food-borne illnesses and gastroenteritis are Campylobacter (between 50% and 80%), Salmonella and E. Coli, as reported by the Food Standards Agency. Particularly, Campylobacter is responsible for more than 280 thousand cases per year in UK, causing at least 100 deaths and a cost of £900 million [3]. Therefore, several strategies have been conducted to reduce the incidence of cases. Since raw poultry is the common source of infection [10], several sporadic cases are associated with undercooked meat. However, the effectively identification of outbreaks is relatively rare. Consequently mathematical approaches to attack the problem are required. The main goal of this project is to develop methodologies for outbreak detection based on the disease notification databases.

Some statistical techniques have been studied for outbreak detection, based on the spatio-temporal information of reported cases. M. Kulldorff first proposed a spatial statistic for the detection of clusters, assuming a multi-dimensional point process on data [4]. Later, an alternative model was proposed by the same author, suggesting a space-time permutation scan statistic for finding regions with high number of cases when compared with the remaining geographical areas [5]. An study of the notification data from a region in New Zealand was described by [12]. It assumes that the number of cases per cell follows a Poisson distribution and the estimation of the risk associated is calculated performing a Bayesian hierarchical model. On the other hand, the analysis of outbreak detection based on the genome has been focussed on detecting the host of the bacteria before transmission to humans. In [9] and [11] genotyping models are proposed to determine the source of infections.

This study had access to the reported cases of Campylobacteriosis in Oxfordshire, for a period of April 2010 - March 2011. The dataset included demographic, spatial and temporal information, as well as the whole genome of collected samples. None of the models developed in literature have had access to both sources of data. This document is organised as follows. Chapter 2 describes the databases and gives an overview of the data. Chapter 3 describes the methodologies proposed. Chapter 4 detailes the results obtained when applying the procedures to the databases. Finally, in Chapter 5 the results are discussed

2. Data

Public Health of England and the Food Standards Agency monitored *Campylobacteriosis* cases through a sentinel surveillance in Oxfordshire, UK. A first part of the data collected included information of the location and date of confirmed cases from April 2010 to March 2011, as well as certain demographic variables. Additionally, a second database was constructed, comprising the genome sequences for the cases with available processed isolates. Only 745 of the 999 cases were considered, excluding those who traveled from abroad in the previous weeks of the report. A detailed description of the data is described in the following subsections. First an overview of the epidemiological data is presented, giving a qualitative insight into the possible outbreaks. Finally a detailed description of the genomic data is covered. Some notation is also introduced.

2.1 Epidemiological data

For each case, age, gender, address (postcode) and reported date are included. When date was missing, it was replaced by the date of notification to Public Health England. For the analysis in this study, the area of Oxfordshire was partitioned in the 104 postcode sectors according to the 2011 census, obtained from the Office of National Statistics. This division is the smaller possible based on postcodes, where each sector comprises a population average of 5000 inhabitants, compared to the total of 612827 in Oxfordshire. Therefore, postcode sectors are smaller in urban areas than in rural zones, as shown in Figure 1.

First, the distribution of occurrences as a function of time is analysed. In Figure 2 the points represent the amount of cases per week, starting in April 2010, while each bar corresponds to the average amount of cases per month. The peak observed in May, June and July represents a typical seasonal rise, result that coincides with previous Campylobacteriosis seasonality studies [7]. Although there are some peaks in weeks 13 and 26, it is not strong evidence for classifying them as outbreaks. Nevertheless, October and February have a number of cases larger than expected compared to the overall, events that should be tested in the quantitative analysis.

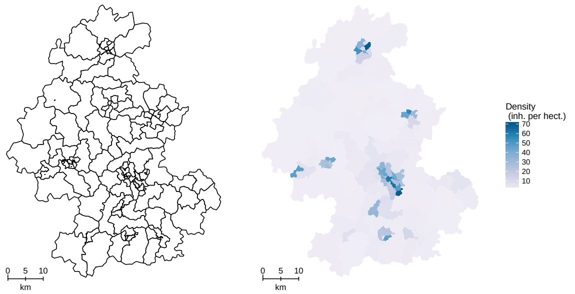

Figure 1. (Left) Partition of Oxfordshire in postcode sectors. (Right) Population density in Oxfordshire, according to the 2011 census. Sectors with high population density are enclosed in smaller areas.

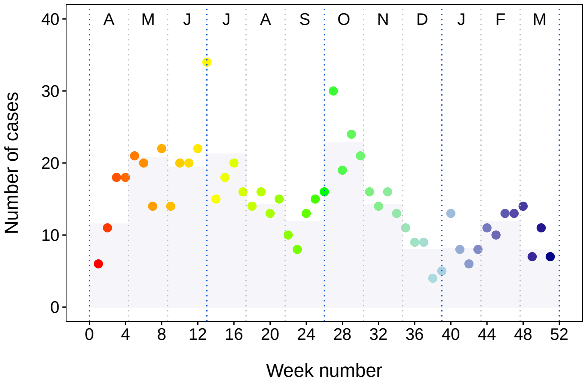

Figure 2. Number of reported cases per week, from April 2010 to March 2011. The average of occurrences per month are represented with bars.

Second, an evaluation of the frequency of cases as a function of age is performed to infer potential outbreaks. In Figure 3 the distribution of age is visualised on a box plot for each week, starting in April 2010. Weeks 1, 38, 39 and 42 are not displayed since they are composed by few cases (less than 7 cases) and the box plots were not significative. In spite of the fact that median for some weeks are larger than the average (e.g week 5 and 20), no inference could be drawn that they were potential outbreaks. However, in other studies there has been notified breakouts where all individuals were within the same age-group, as in [13]. Therefore, these observations potentially support results obtained from quantitative analysis.

Finally, a study of the incidence of gender and type of location (rural or urban) is performed. Figure 4 (top) shows the

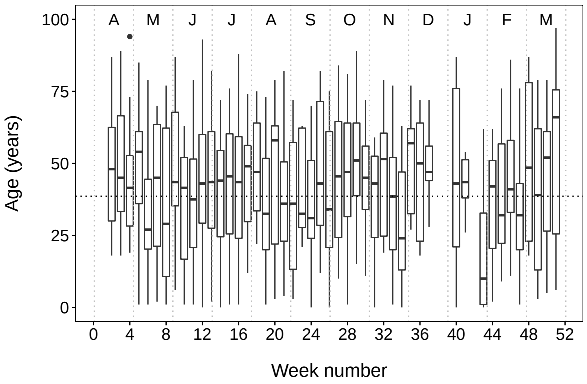

Figure 3. Distribution of age of reported cases per week, from April 2010 to March 2011, excluding those with less than 7 cases. The horizontal dotted line corresponds to the median of age in UK (38.6 years), as calculated in the 2011 census. For each box plot, the first and third quartiles are represented by the bottom and the top of the rectangle, while the middle line shows the median. The upper whisker extends from the third quartile to the maximum value below 1.5 times the inter-quartile distance. Similarly for the lower whisker.

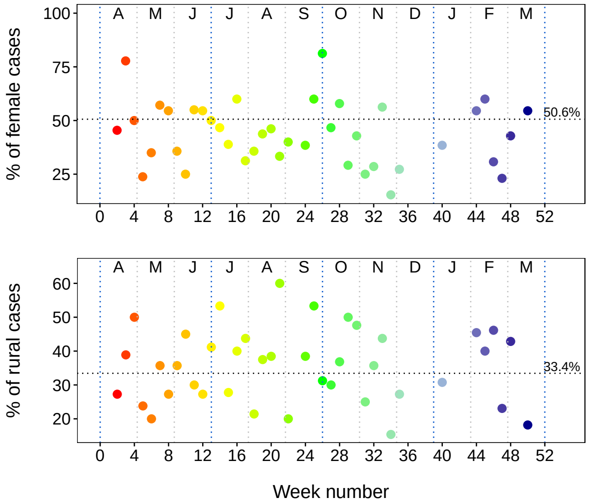

Figure 4. (Top) % of reported female cases cases per week, from April 2010 to March 2011. The dashed line represents the female/total ratio of Oxfordshire in the 2011 census. (Bottom) % of reported rural cases cases per week, from April 2010 to March 2011. The dashed line represents the rural/total ratio of Oxfordshire in the 2011 census.

percentage of female cases per week, compared to the ratio of female/total of Oxfordshire in the 2011 census (50.6%). In general, males reports are more frequent than females ones as observed in previous analysis performed in UK [7]. Two possible outliers are observed in weeks 3 and 26, where the ratio is larger than 75%. On the other hand, Figure 4 (bottom) displays the percentage of rural cases, compared to the Oxfordshire ratio (33.4%). Reported incidences in rural sectors exhibited several peaks in contrast to urban cases. [7] reported a significant correlation between Campylobacter samples and agricultural environments. Generally, agricultural activities includes manipulation of poultry, cattle and other host animal for Campylobacter, having an effect on incidence rates. Although some weeks as the 34 have extreme values, no inference can be made about the existence of outbreaks. Oxfordshire is mostly rural, where the agriculture managed around 80% of the county [3]. Therefore, several rural cases are not necessarily confined to the same geographical region.

2.2 Genomic data

The epidemiological data described in the previous section contains a link to an additional database covering the whole Campylobacter genome of each reported case. However, this database collected by the Food Standards Agency only includes samples (*isolates*) for the 71.4% of the reported incidences (532 cases). The analysis of the sequence data was performed by the software platform BIGSdb, comparing every isolate within the database and providing a distance matrix D of size 532×532, based on the amount of shared *alleles*. Each $d_{ij} \in \mathbb{R}$ represent the distance between the isolates c_i and c_j, where $d_{ij} = 0$ indicates that both genomes are identical and $d_{ij} = 1643$ is the maximum possible value. A cumulative plot of the obtained distances is shown in Figure 5, where the maximum value obtained is 1637. Most of the cases are not closely related.

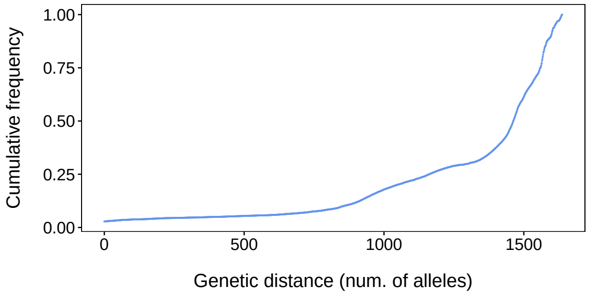

Figure 5. Cumulative frequency of distances between the genomes, for the 532 cases with available genetic information.

3. Methodology

The methodologies proposed in this study are divided into the spatio-temporal ones, based in the epidemiological data, and the genomic-based ones, derived from the genomic data. Although the starting point of both approaches are separated, they use the information both datasets.

3.1 Spatio-temporal analysis

In this section some methodologies for spatio-temporal data clustering are introduced. The *discrete Poisson model* [4] is described in the first subsection. Two approaches based on GLM are proposed in subsections 3.1.2 and 3.1.3. Based on the partition mentioned previously, the following notation is introduced and shall be used throughout the remainder of this document. Let $S \subseteq \mathbb{R}$ be the spatial area of Oxfordshire and let $I = [0,T]$ be the interval of time of the observed data (T is one year). S is divided into $K = 104$ postcode sectors S^k such that $S = \dot{\bigcup}_{k=1}^{K} S^k$. Hence, the total spatio-temporal space considered in this study is $S \times I$.

3.1.1 Discrete Poisson model

As described in [4], the *discrete Poisson model*, or DPM, scans a collection of subsets in the spatio-temporal space $S \times [0,T]$ to determine whether they are possible clusters, and it assumes that the number of cases per region follows a Poisson distribution. The model fixes a subset $R \subseteq S \times [0,T]$ to examine whether it is a cluster. Let p be the probability of having the disease for an individual in R, and q for an individual in R^C. Then, the number of cases observed for every region $A \subseteq S \times [0,T]$ follows the distribution $O(A) \sim Poi(p|A \cap R| + q|A \cap R^C|)$. A permutation test is defined according to this result: the null hypothesis consists of an homogeneous Poisson process while the alternative one is inhomogeneous; that is, $H_0 : p = q$ while $H_a : p > q$.

To apply the model to the epidemiological data, the software SaTScan$^{\text{TM}}$ v9.4 was used [5]. The software evaluates the hypothesis using a likelihood ratio test to determine if the region R is a cluster or not. However, the scan is run only for a set of regions with certain predetermined characteristics: only regions with cylindrical shape are considered, that is, $C_k \times [a,b]$, where C_k is a circle with centre in c_k and $0 \le a \le b \le T$. For the set of regions considered, the model chooses the most significants according to the p-value of the hypothesis test.

The SaTScan$^{\text{TM}}$ settings used for the epidemiological data analysis were the maximum temporal length of each cylinder was 28 days (i.e. it only considers possible clusters of maximum one month of duration). Additionally, each c_k is located in the centroid of a corresponding sector S^k in Oxfordshire and the circle cannot include more than 25% of the total population.

3.1.2 Poisson GLM approach

The *Poisson GLM approach*, or PGLM, considers a partition in cells of the total spatio-temporal window $S \times I$ and assumes that the number of cases per cell follows a Poisson distribution. Moreover, it estimates the parameter of each distribution applying a Poisson Generalised Linear Model. Namely, let $\{S^k \times I^l\}_{l=1,\ldots,L}^{k=1,\ldots,K}$ be the partition of the space $S \times I$ where $\{S^k\}$ are the sectors in Oxfordshire, n_k is the population per sector and $\{I^l\}$ is a partition of the time interval I in subintervals of four weeks. For each $S^k \times I^l$, the observed number of cases in each cell follows a Poisson distribution $X_{kl} \sim \mathcal{P}oi(\lambda_{kl})$ where $\lambda_{kl} = n_k \lambda_k \lambda_l$ is composed by a spatial rate λ_k (quantifying the spatial risk of sporadic cases) and a temporal rate λ_l (containing the information of seasonality).

For estimating the parameters λ_{kl} a PGLM is applied; that is, based on the epidemiological data the λ_{kl} is given by

$$\log\left(\frac{\lambda_{kl}}{|S^k|}\right) = \alpha + \beta_k + \gamma_l. \tag{1}$$

Each coefficient β_k is related with the epidemiological risk factor of the spatial region S^k whilst each γ_l depicts the temporal trend of the disease for the interval I^l. To enable identifiability, one of the β's and γ's is defined to be zero.

Once the λ_{kl}'s are estimated the probability of having at least x_{kl} cases in the cell $S^k \times I^l$ is calculated, where x_{kl} is the number of cases reported in the epidemiological data. This probability is given by

$$p_{kl} := \mathbb{P}(X \ge x_{kl} | \lambda_{kl}). \tag{2}$$

To determine which cells are possibly containing an cluster, a threshold $c \in (0,1]$ is chosen such that $S^k \times I^l$ is a cluster if and only if $p_{kl} < c$. If more than one neighbouring region was detected, it would be considered one single cluster.

3.1.3 Robust Poisson GLM approach

Similarly to the previous procedure, the *robust Poisson GLM approach*, or RGLM assumes that the number of cases per cell follows a Poisson distribution with a λ_{kl} estimation described by (1). However, the estimation of β proposed in the subsection 3.1.2 is highly sensitive to outliers [1], including potential outbreaks. Consequently, a robust GLM for determining β is proposed to avoid the influence of extreme values in the estimation. The estimated parameters would then represent the distribution of sporadic cases only.

3.2 Genomic analysis

In this section a methodology for genomic data clustering is introduced. Let $c_1,\ldots,c_N$ be the isolates included in the analysis, and let $D \in \mathbb{R}^{N \times N}$ be the matrix of distances d_{ij} between each pair of genomes c_i, c_j. An *agglomerative hierarchical clustering* algorithm, or AHC, is applied to D, with the aim of finding clusters of cases according to their genetic similarity. In the first step $w = 1$ the set of clusters $\{c_1\},\ldots,\{c_N\}$ is considered. For the step w the method merges the two closest groups C_1^w, C_2^w into one set $\mathcal{C}^w := C_1^w \cup C_2^w$. The iterations continue until all data conforms a single cluster $\{c_1,\ldots,c_N\}$ (in the last step $w = W$). Nevertheless, each iteration requires a measure for calculating the distance between clusters, known as *linkage criteria*. In this study, the *average linkage* was chosen. That is, for two clusters A, B, the distance is given by:

$$\text{dist}(A,B) = \frac{1}{|A||B|} \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} d_{ij}. \tag{3}$$

Although more sophisticated linkage functions have been proposed, the simplicity of the one in equation (3) fulfills the goal of this analysis.

The standard representation of the output of this algorithm is a binary tree or *dendogram*, where the bottom of the tree corresponds to the set of singletons $\{c_1\},...,\{c_N\}$ and the top to the single cluster $\{c_1,...,c_N\}$. Selecting a procedure for cutting the tree determines the definitive set of clusters. In genetic clustering several methods for - have been defined, as in [6]. However, none of them includes the spatio-temporal information of each isolate. Hence, a methodology for cutting the dendogram based on the temporal data had to be developed and is given in Algorithm 1. Although only temporal data is included; the procedure could be adapted to incorporate information of the spatial distribution.

```
input  : A series of sets: $\{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^W\}$
output : A series of maximal clusters $\mathcal{M}^{\max}$
initialise $\mathcal{M} \leftarrow$ empty set;
for $w$ in $W{:}1$ do
    $\mathcal{T}^w \leftarrow$ reported day of each $c_i$ in $\mathcal{C}^w$;
    $v_w \leftarrow \mathrm{var}(\mathcal{T}^w)$;
    if $v_w < \theta$ then
        $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{C}^w$ ;
    end
end
$\mathcal{M}^{\max} \leftarrow$ maximal elements of $\mathcal{M}$;
```

Algorithm 1: Methodology for cutting the tree obtained by the *agglomerative hierarchical clustering*.

The algorithm starts from the bottom of the tree, when $w = W$. The day in the time interval I when each isolate $c_i \in \mathcal{C}^w$ was reported is stored in a set $\mathcal{T}^w$. To determine if the set $\mathcal{C}^w$ is a possible outbreak, the variance of $\mathcal{T}^w$ is calculated. If the value is less than a threshold parameter θ, the set $\mathcal{C}^w$ is marked as *cluster*. The algorithm continues until $w = 1$. Finally, a set of *clusters* is obtained. However, only the *maximal* ones are chosen. That is, the *clusters* contained in another one are not taken into account. It can be easily proved that every *cluster* is either contained in another one or is *maximal*.

4. Results

4.1 Spatio-temporal analysis

The methodologies proposed in Section 3.1 were applied to the epidemiological data. The clusters detected by the DPM are given in Table 1. The geographical location and the period of time of each cluster was included, along with the p-value in order to evaluate the hypothesis test. The matrix of genetic distances between cases were computed to assess the genetic proximity between isolates. For the Campylobacter genome, a pair of samples is *genetically similar* if the distance is less than 30 *loci* approximately. Therefore, the percentage of pairs of isolates significantly related was calculated using a threshold of 30 [1].

Two outbreaks were detected by the DPM, such that the p-value was less than 0.05. Each of them contains at least %20 of the total population. One occurred in October, while the other only in three days of July; peaks observed in the plot of cases in Figure 2. The distribution of age in both cases was irrelevant and it is not included in Table 1. Moreover, the genomes differ considerably (most of the distances were larger than 1300 loci). However, two cases in the second outbreaks were genetically correlated, with a distance of 20 loci. These two isolates could be part of a real outbreak. It is important to take into account that the source of the infection could contain non-related Campylobacter bacteria. Then, more isolates in this cluster could be part of the same outbreak. Although the genetic data gave a feedback of the results, the model does not provide any other method for testing its performance.

Comparatively, the PGLM was applied to the epidemiological data. The area of Oxfordshire was partitioned into postcode sectors as mentioned in Section 2.1. Nevertheless, the large number of cells obtained in the spatio-temporal space (1352 cells) were not suitable for estimating the parameters of the model. Consequently, the area was divided again into postcode districts, all of which are 20 times larger on average than the postcode sectors. The output of the model is included in Table 2, similarly to the previous approach. In addition to the geographical and temporal location, the probability p_{kl} defined in equation (2) was computed. The percentage of pairs significantly related was also calculated using a threshold of 30.

Two clusters were detected by the PGLM, when a threshold of $c = 0.02$ was fixed (for smaller thresholds no outbreaks would be detected). The size of the population for each region is smaller than the ones detected by the DPM, since the regions here are fixed. Again, the distribution of age was not conclusive and therefore it is not shown. Additionally, the distance matrix of both cases do not provide any similarity between isolates (the closest distance was 300 loci approx.).

In addition to the outbreaks overview, the parameters β_k and γ_l of the Poisson GLM were obtained. Based on these quantities, the rates λ_{kl} can be computed according to equation (1). Moreover, the expected number of cases in the region S_k are calculated, quantifying the risk of acquiring Campylobacteriosis in each area, as shown in Figure 6. Analogously, Figure 7 represents the expected number of incidents γ_l for every interval of time. These quantities are useful for improving the understanding of the underlying epidemiological risk associated to each particular location and period of time.

4.2 Robust GLM approach

In this section, a RGLM procedure is applied to the epidemiological data. As opposed to the PGLM approach, the RGLM algorithm was aim to ignore outliers in the calculation of the parameters β_k and γ_l. For performing this test, the function `glmrob` of the package `robust` in `R` was applied. In this algorithm, some predefined robust estimators are computed iteratively, as described in [1]. However, the algorithm did

[1] Actually, the choice of the threshold was not relevant, since the obtained distances were considerably larger to represent related isolates.

Outbreak No.	Number of cases	% of total population	Period	p-value	% pairs signif. related
1	37	21.74%	02 Oct - 26 Oct	0.001	0.0%
2	12	21.23%	02 Jul - 04 Jul	0.014	6.6%

Table 1. Detected outbreaks applying the DPM. The minimum genetic distance between the cases is calculated to validate if they are possible clusters.

Outbreak No.	Number of cases	% of total population	Period	Probability p_{kl}	% pairs signif. related
1	8	7.64%	26 Apr - 23 May	0.010	0.0%
2	12	9.45%	08 Nov - 05 Dec	0.020	0.0%

Table 2. Detected outbreaks applying the PGLM. The minimum genetic distance between the cases is calculated to validate if they are possible clusters.

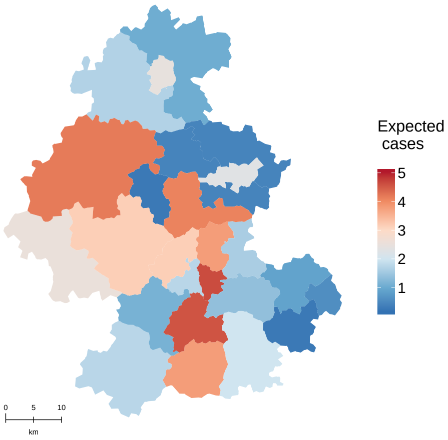

Figure 6. Expected number of reported Campylobacter cases in each postcode district in an interval of 4 weeks, obtained with the PGLM approach.

not converge for the epidemiological data, even if the sectors partition or the district were applied to the area of Oxfordshire. That result could indicate that no outliers were found by the algorithm. To support this statement, a *quasi Poisson GLM* was fitted to the data, since this model is able to quantify whether a Poisson distributed data has overdispersion. The result obtained showed no dispersion, suggesting that the number of cases per cell could not have spatio-temporal significant clusters. That would explain the small and differing number of results obtained by previous methods.

4.3 Genomic analysis

The AHC algorithm was applied to the genomic data to construct a tree with possible clusters. The resulting dendrogram is shown in Figure 8, where the top of the tree corresponds to a unique cluster containing all points (when the maximum distance for being a cluster is equal to 1643 alleles), while the bottom represents the set of clusters of size 1 (maximum distance equal to 0 alleles). Then, a dynamical cutting is performed as proposed in Algorithm 1. The detected outbreaks

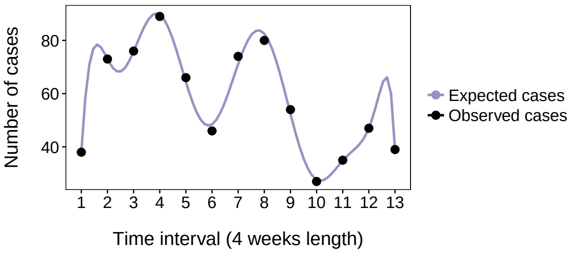

Figure 7. Expected number of reported Campylobacter cases in Oxfordshire, obtained with the PGLM approach.

are described in Table 3, where the number of cases and the period of time is included. Moreover, the percentage of significantly related isolates is calculated with the aim of testing the cutting performance. Nine possible outbreaks were detected, six of which were genetically close between them.

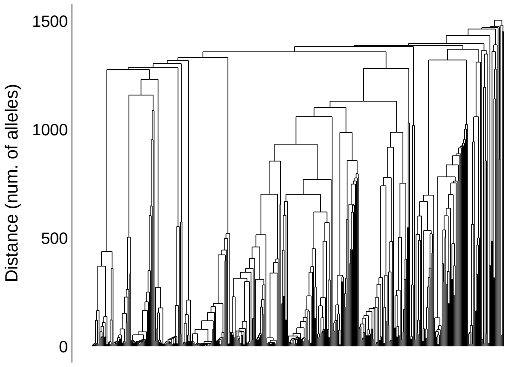

Figure 8. Dendrogram obtained when a Hierarchical Clustering is applied to the matrix of distances between reported Campylobacter cases.

Previous reports suggested that Campylobacters are genetically stable for short-term epidemiological investigations [8]. Therefore, the genetic proximity between isolates found in this study is conclusive to determine their connection, even if there is no other mechanism to validate the results. Nevertheless, an outbreak could be composed of two or more unrelated

Campylobacters. Then, the results in Table 3 could be a subset of a larger epidemic.

Outbreak No.	Number of cases	Period	% pairs sign. related
1	3	22 Apr - 07 May	100.0%
2	3	30 Apr - 17 May	100.0%
3	3	31 Oct - 09 Nov	100.0%
4	3	05 Jan - 12 Jan	100.0%
5	4	16 Jan - 03 Mar	100.0%
6	4	07 Jun - 24 Jun	66.6%
7	3	10 Jun - 08 Jul	0.0%
8	3	14 Aug - 04 Sep	0.0%
9	3	01 Sep - 20 Oct	0.0%

Table 3. Detected outbreaks applying the genomic analysis.

It is important to realise that none of the spatio-temporal algorithms proposed in Section 3.1 was able to detect the clusters in 3. The spatial distance between the points composing the outbreak is a possible explanation for this affirmation. However, a exhaustive analysis showed that cases in outbreak No. 2 in Table 3 were relatively closed in space, as shown in Figure 9. The DPM could fail in the detection of these cases since any circular region including these locations would also include areas with small number of incidents. In that case, the cluster would be not significant. In a similar way, the PGLM approach was limited only to neighbouring district regions, including those with small number of Campylobacter reports.

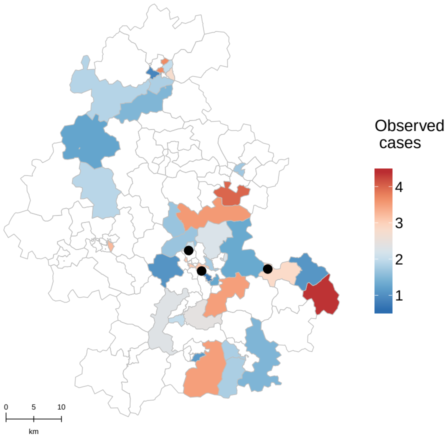

Figure 9. Location of cases of the outbreak No. 2 in Table 3 (black points). Colors indicate the number of cases in the period of time the cluster occurred (30 Apr - 17 May).

5. Discussion

Three spatio-temporal approaches were proposed to face the problem of outbreak detection. First, the DPM was considered, with the aim of measuring spatio-temporal regions with significantly more risk when compared with the others. The method found two outbreaks occupying large areas (covering 20% of population each one). Proposed by Kulldorff [4] and commonly used for cluster detection, this approach is useful for comparing any new procedure with previous works. Second, the PGLM calculated the parameters of a GLM and computed the risk associated to each cell of a spatio-temporal partition. Although it found two possible clusters, the probability of observing them was not considerably low (it was greater than 0.01). The Figures 7 and 6 of expected cases per time and region respectively, are useful tools for understanding the epidemiological risk associated to each area. However, the similarities between the observed and expected cases confirm that possibly there were no noteworthy outbreaks. Finally, the RGLM attempted to calculate the parameters of the GLM ignoring potential outliers. However, the method did not converge, supporting the presumption that no significantly outbreaks occurred in the interval considered in this study.

The results achieved by the spatio-temporal methods were substantially different. Whereas no comparison test was specified, the genetic data was crucial in assessing the effectiveness of each approach. Then, the distance between pairs was calculated per each outbreak. Nonetheless, the large differences among them indicated that the testing could not be conclusive. Moreover, no all the isolates were available; then some missing information could correspond to related cases. Therefore, there are no strategies to compare or to test the efficacy of the methods proposed. Also, this project did not have access to estimated outbreaks defined by Food Standard Agency or other governmental instances. A simulation of sporadic and outbreak cases would compare the performance of these methods. Finally, although the DPM is commonly used for outbreak detection, the small size of Oxfordshire compared with usual studied regions is a limitation.

In contrast to the performance of the spatio-temporal analysis, an inspection of the genetic data detected five genetically related clusters, as shown in Table 3. The Algorithm 1 effectively identified similar isolates, showing that the mixture of epidemiological with genome sequences is substantial for the identification of potential outbreaks. However, as noted previously, real clusters could contain bacteria with non-related genomes. Therefore, further investigation by experts should be considered to determine other possible individuals affected by the epidemic.

In literature, most of the algorithms used for gene sequence clustering are based in the *hierarchical clustering model* [2]. However, a model combining spatio-temporal and genetic information has never been suggested. The AHC together with the cutting procedure successfully incorporates both sources of data. In addition, some improvements are suggested. The function used for performing the cutting is based in the variance of the temporal distribution of cases. However, a more sophisticated criteria could be developed, including spatial or demographic variables. Also, a test should be included to determine the performance of the cutting. For instance, it could help to estimate the impact of the chosen linkage function (in this study only the average linkage was taken into account).

Some limitations of the approaches of this study are considered. Not all Campylobacteriosis cases are reached notification. In fact, Food Standards Agency estimates that there are 280000 cases per year in UK (approximately 3000 in a region of the size of Oxfordshire). However, the database used in this analysis had only 743 reports. Moreover, not all cases had sequencing data or they are incomplete (restricting the reliability of genetic comparisons). On the other hand, the PGLM and RLGM are sensible to the spatio-temporal partitions. Large regions could include more sporadic than outbreak cases, whilst small ones could no register any incident. Additionally, small size outbreaks are not easily detected by any of the spatio-temporal methods.

6. Conclusions and Further work

In this study four methods were applied to detect Campylobacter outbreaks: a *discrete Poisson model*, a *Poisson GLM approach* and a *Robust Poisson GLM approach*. The first three were based on the spatio-temporal distribution of cases and were tested using the genetic distances between the samples. Although some potential outbreaks were detected by the first two methods, the test was not conclusive for determining the validity of these approaches. The last model, based on an *agglomerative hierarchical clustering*, performed an algorithm to find genetically related cases while testing temporal proximity. Further modifications to this approach are suggested for improving performance.

Acknowledgements

Finally, the author would like to thank specially Noel McCarthy and Simon Spencer for their invaluable guidance through the fields of Statistics and Epidemiology

References

[1] E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96:455, 2001.

[2] G. Dong and J. Pei. *Sequence data mining*. Springer, 2007.

[3] F. S. A. (Food Standards Agency). The joint government and industry target to reduce campylobacter in uk produced by chickens by 2015. december 2010. Available online, December 2010.

[4] M. Kulldorff. A spatial scan statistic. *Commun. Statist. - Theory Meth.*, 26, 1997.

[5] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunc, and F. Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLoS Med*, 2, 2005.

[6] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24:719–720, 2007.

[7] V. R. Louis, I. A. Gillespie, S. J. O'Brien, E. Russek-Cohen, A. D. Pearson, and R. R. Colwell. Temperature-driven campylobacter seasonality in england and wales. *Applied and Environmental Microbiology*, 71, 2005.

[8] G. Manning, B. Duim, T. Wassenaar, J. Wagenaar, A. Ridley, and D. Newell. Evidence for a genetically stable strain of campylobacter jejuni. *Applied and Environmental Microbiology*, 67, 2001.

[9] N. D. McCarthy, F. M. Colles, K. E. Dingle, M. C. Bagnall, G. Manning, M. C. Maiden, and D. Falush. Host-associated genetic import in campylobacter jejuni. *Emerg Infect Dis*, 13, 2007.

[10] G. L. Nichols, J. F. Richardson, S. K. Sheppard, C. Lane, and C. Sarran. Campylobacter epidemiology: a descriptive study reviewing 1 million cases in england and wales between 1989 and 2011. *BMJ Open*, 2, 2012.

[11] S. K. Sheppard, J. F. Dallas2, N. J. C. Strachan, M. MacRae, N. D. McCarthy, D. J. Wilson, F. J. Gormley, D. Falush, I. D. Ogden, M. C. J. Maiden, and K. J. Forbes. Campylobacter genotyping to determine the source of human infection. *Clin Infect Dis*, 48, 2009.

[12] S. E. Spencer, J. Marshall, R. Pirie, D. Campbell, and N. P. French. The detection of spatially localised outbreaks in campylobacteriosis notification data. *Spatial and Spatio-temporal Epidemiology 2*, pages 173–183, 2011.

[13] J.-H. Yu, N.-Y. Kim, N.-G. Cho, J.-H. Kim, Y.-A. Kang, and H.-G. Lee. Epidemiology of campylobacter jejuni outbreak in a middle school in incheon, korea. *J Korean Med Sci*, 25, 2010.