NIHR | Health Protection Research Unit in Gastrointestinal Infections at University of Liverpool

Joint analysis of national surveillance and external data modelling of disease spread

Ellie Brown

Abstract

Gastrointestinal infections are a major public health concern. There is enhanced interest in understanding what major environmental and individual patient factors influence patterns of gastrointestinal infections in England, and how these characteristics may be used to benefit disease prevention. The spatio-temporal patterns of infections will be explored using intensive statistical analysis of surveillance and external data to determine what factors influence infections. Weather station data, animal data, and data on the COVID-19 pandemic will be among the external data sources. This PhD will examine various models, analysing their effectiveness in understanding how different covariates, such as temperature, influence cases. Beginning with basic regression models, building to non-frequency domain approaches. Methods discussed will include regressions models, decomposition methods, smoothing methods and ARIMA models.

Results

The following results are from a sample Campylobacter data set featuring case numbers from October 2015 to July 2018. A weekly plot of cases is seen in Figure 4.1.

There is some interesting behaviour. There is no clear long-term increase or decrease in the data. There seems to be seasonal variation in the number of cases per week - there is a peak every summer, and a trough every winter. The seasonal fluctuations are roughly constant in size over time and do not seem to depend on the level of the time series, and the random fluctuations also seem to be roughly constant in size over time; the series could be described using an additive model. If an additive model is not appropriate for describing this time series, the time series may be transformed to get a transformed time series that can be described using an additive model.

To look more into the seasonality of the data, a monthly seasonal plot is created where the data is plotted against individual seasons, as represented by Figure 4.1 There is distinctively different seasonality in the spring-summer period. Each year has a marked increase from April when temperatures begin to increase. For 2016 and 2017 the peak is around June or July, whereas there is a forward shift in the seasonal pattern in 2018 with the peak occurring around mid-April. 2017 and 2018 both feature decreases in cases from their peaks, whereas 2016 features a second peak during November. It would be beneficial to find the cause of this peak during the autumn.

Plots of lag are produced where the horizontal axis shows the lagged values of the time series. Each graph shows the time series observation plotted against the observation k-periods behind for different periods k. The colour indicates the month of the recorded case. The relationship is positive at lags 1 and 12, reflecting the strong seasonality in the data (the number of cases one month apart is similar, as well as the number of cases one year apart). There is a negative relationship at lag 5, this is likely due to the different seasons experienced five months apart, supporting seasonality argument.

A standard STL decomposition is applied. The two parameters to be chosen are the trend window and the seasonal window. These control how the trend and seasonal components change, the smaller the value the more rapid the change. Both trend and seasonal windows should be odd numbers; the trend window is the number of consecutive observations to be used when estimating the trend-cycle; and the season window is the number of consecutive months to be used in estimating each value in the seasonal component. Setting the seasonal window to be infinite is equivalent to forcing the seasonal component to be periodic. Figure 4.6 shows the output. The trend follows the overall movement of the series, ignoring any seasonality and random fluctuations. There is a steady decrease until mid-2016, where it then begins to increase for a few months during summer, remaining at similar levels towards the end of 2016 where there is a decreasing trend until mid-2017 where there is a sharp increase. A sharp decrease occurs during mid-2018. The remainder component is what is left when the seasonal and trend components are removed

Conclusion

Using surveillance data from England and publicly available external data, the regional and temporal risk of gastrointestinal infection was computed. The temporal and spatial variation has been investigated using a combination of regression approaches, ARIMA models, and spectral analysis. The hypothesis at the start of this study was that seasonal variations would increase the incidence of gastrointestinal illnesses. Previous studies have indicated environmental factors such as variations in temperature and sunlight hours. Despite the result of this report proving that during summer months Campylobacter cases increase, the exact cause of this increase remains unknown. It does, however, point to some potential new areas of inquiry for locating the source of increasing infections. It is integral to this report to understand the relationship behind the seasonality. If exposure terms can replace the seasonality, then they are the driving force behind infection. A high resolution spatial temporal linkage of external data parameters and cases is required.

A better knowledge of the dynamics of gastrointestinal infections is critical in the current setting. Time-series analysis has emerged as an intriguing method for studying the dynamics of a variety of diseases that follow stationary patterns. However, the time series discussed in this report are non-stationary and require better suited methods such as Wavelet analysis. Spectral analysis may also be applied to account for trends and cycles within the time series. These methods will prove crucial for distinguishing modes of transmission in this report and answering central research questions, where normal statistical methods will fail.

Questions

1. To identify the spatial risk associated with various factors including area-level deprivation, proximity to livestock, and overpopulation by combining surveillance data with external data; to address the disparity in case numbers between urban and rural locations across demographic groups.

(a) Can increased case incidence in rural areas be explained by the spatial distribution of livestock? (b) Is there a correlation between case numbers and social economic variables like income, education, or area-level deprivation?

(c) Are case numbers correlated with population density or urbanisation, when controlling for demographic factors?

2. To establish how to capture the seasonal trends that have been observed in gastrointestinal infections, including the timing and amplitude variability. Subsequently, to evaluate whether covariate or external data can explain these seasonal patterns.

(a) Can seasonal patterns in notified case data be explained by covariate data, such as temperature and rainfall?

(b) Can seasonal patterns in case data be explained by extreme weather events, such as heatwaves?

(c) Is there remaining seasonality after taking out covariate data such as temperature and precipitation?

3. To explore the hypothesis that the COVID-19 pandemic resulted in substantial changes in infection, and to establish how the subsequent restrictions on personal freedoms impacted case counts.

(a) How has the closure and re-introduction of restaurants, specifically through the 'Eat Out to Help Out' scheme, influenced the rate of infection?

(b) How has the reduction in international air-travel impacted gastrointestinal cases, and what can be learnt about transmission routes into England from this?

(c) How has the systematic reintroduction of businesses provided insight into the importance of risk variables, such as human-to-human transmission?

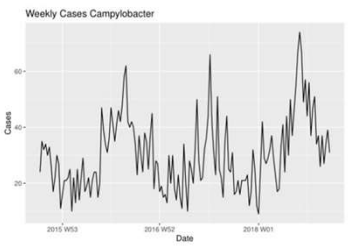

Figure 4.1: A plot of weekly *Campylobacter* cases from October 2015 to July 2018

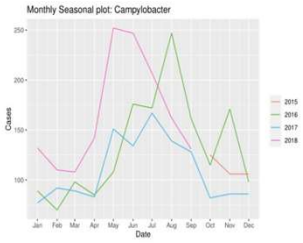

Figure 4.2: A seasonal plot of monthly *Campylobacter* cases

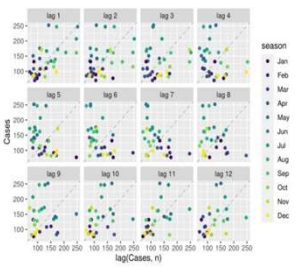

Figure 4.5: Lag Plot of monthly *Campylobacter* cases

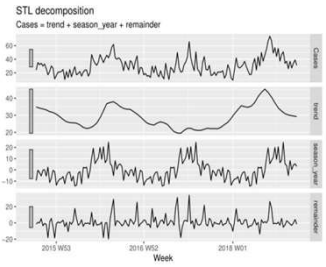

Figure 4.6: Decomposition using STL

The research was funded by the National Institute for Health Research Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at University of Liverpool in partnership with the UK Health Security Agency (UKHSA), in collaboration with University of Warwick. The views expressed are those of the author(s) and not necessarily the NIHR, the Department of Health and Social Care or the UK Health Security Agency.