Space-time modelling of sexually transmitted infections in London with focus on *Chlamydia trachomatis*

Desislava Petrova[1], Prof Tao Cheng[1], Dr Ian Simms[2]

[1]Department for Civil, Environmental and Geomatic Engineering, UCL
[2]Public Health England, HIV and STI Department

September 2016, 2016

Summary

Space-time modelling techniques were used as an innovative method to model spatio-temporal variation of STIs to identify areas and periods at high risk so as to inform better prevention and control strategies by public health bodies. This study presents the implementation of three predictive models of *Chlamydia trachomatis* in London – SVR, ETS and Croston's method. The results show that varying spatial and temporal scales have an impact on the accuracy of the models, thus suitable combination of such scales is essential. The project revealed that space-time modelling of chlamydia can significantly improve understanding of disease behaviour in space and time.

KEYWORDS: Space-time modeling, machine learning, sexually transmitted infections, forecast accuracy

1. Introduction

The World Health Organisation (WHO) estimated that 357 million new cases of curable STIs occur annually of which chlamydia has the highest incidence at 131 million (WHO, 2016). Monitoring to decrease incidence in STIs is a global priority due to the strain imposed on national health budgets and the overall hell-being of individuals.

In the UK, the area with the highest rate of reported new STIs is London (PHE, 2014) (Fig 2). The rate of new STIs in 2014 was estimated to be 1,347 per 100,000 population which is 65% higher than rates in any other regions in Great Britain. Monitoring and scientific research of STIs in the UK is executed by Public Health England (PHE) which found that chlamydia is the most common sexually transmitted infection diagnosed in the UK (Adams, Charlett, Edmunds and Hughes, 2004) with 200,288 positive diagnoses in 2015 which is nearly half of all STIs diagnoses in England (PHE, 2016). As an asymptomatic disease its monitoring is crucial so as to avoid further costs for the National Health Service which can be avoided if contracted individuals are treated on time. Hence, it is vital to identify areas experiencing the highest rate of the infection so that health services in these regions can be prioritised.

Transmission of infectious diseases is related to the concepts of spatial and temporal proximity of case events as spread of a disease is more likely to occur in infected individuals who are located closely in spatial and temporal sense (Pfeiffer et al., 2012; Waller and Gotway, 2004; Lawson, 2010). Therefore, spatial mapping of STIs was developed as an essential component of identifying

origin areas of disease spread (Becker, Glass, Brathwaite and Zenilman, 1998). Further technological development in spatial analytical systems allowed for spatial modelling of STIs using techniques such as kriging, spatial clustering and Bayesian models (Pfeiffer et al., 2012; Waller and Gotway, 2004; Lawson, 2010). However, spatio-temporal modelling has been limited.

An early application of GI systems was used in Baltimore, United States of America (USA) to explore geographic patterns of gonorrhea (Becker et al., 1998) where geographically defined "core areas" of disease occurrence were identified. More advanced techniques have been adopted by Law et al. (2004) calculating spatial variability of chlamydia, gonorrhea, syphilis and HIV in Wake County, North Carolina, USA rates using covariance functions which helped in the assessment of their spatial patterns. The study went further to predict STI rates using kriging. Jakob et al. (2015) have conducted a study using space-time cluster analysis (Kulldorf's spatial scan statistic) to explore space-time variation of infectious syphilis epidemic in England among men who have sex with men (MSM). The analysis helped understand the separation of endemic and outbreak areas of syphilis among MSM and how the disease develops over time.

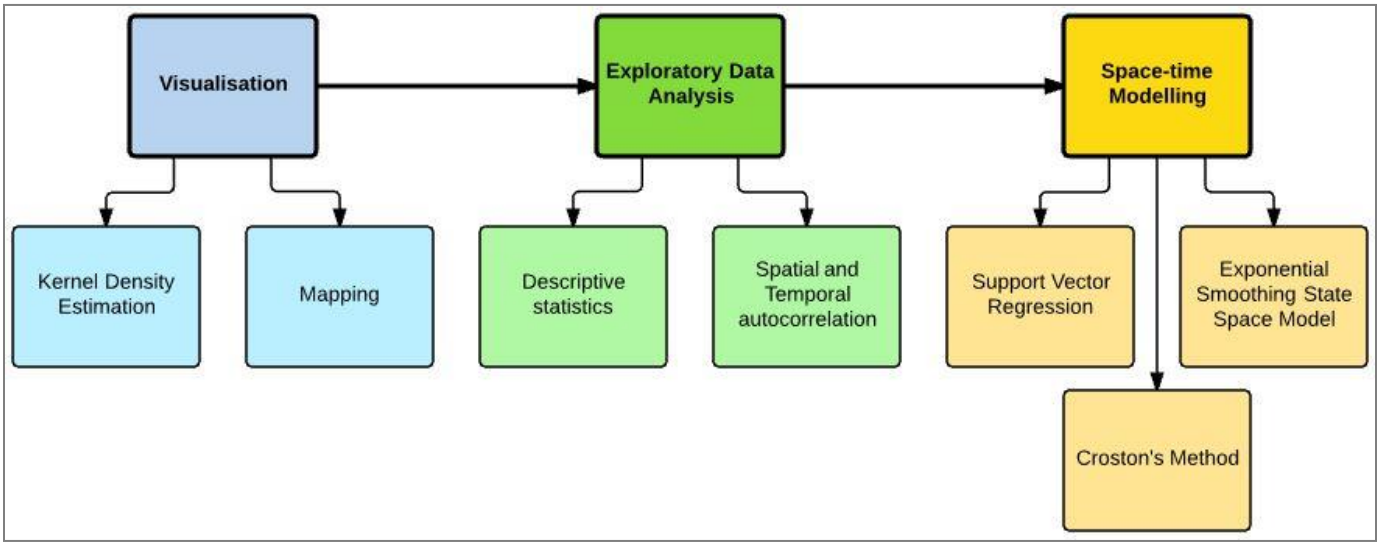
Garnett & Anderson (1996) argue that the use of mathematical models which simplify assumptions and complexity of an existing process can lead to significant insights to the factors or processes that affect the epidemiological pattern of a disease, especially in STIs. Understanding the factors that control the spread of STIs can be very complex due to varying social and behavioural drivers of individuals thus, including such complexity in a predictive model may not yield satisfying results. Therefore, the predictive models in this study do not take into account any demographic or behavioural factors.

Space-time analysis of existing cases can be very powerful as it would provide understanding of patterns in a temporal dimension as well. Such information can be vital for the planning of an adequate prevention strategy by public bodies. The aim of this study was to investigate spatio-temporal variation of chlamydia in London and subsequently identify the appropriate and accurate prediction model for future outbreaks of chlamydia in London and areas at high risk of repeat infection.

2. Methodology

A robust methodology was designed and divided into three stages to achieve the aim of the study (Figure 1). The first stage was the visualisation allowing recognition of spatio-temporal patterns. The next step was exploratory data analysis (EDA) allowing an insight to underlying structure of the data and informing appropriate types of modelling (Kanevski and Maignan, 2004). Space-time modelling was the third and the most significant stage at which suitable space-time models were selected and performed. The most valuable advantage of this approach is that contrary to previous work in STIs modelling, where the causation of the infection was of primary focus of the analysis, this approach is data driven. It does not make any previous assumptions about the causation of the disease but results of analysed data can be used to find an explanation as to why certain patterns occur. Models were built at different spatial (borough and postcode) and temporal (monthly and weekly) scales to assess which combination those provides the highest accuracy.

Figure 1 Diagrammatic representation of the methodological framework.



3. Results and discussion

Visualisation of *Chlamydia* spread reveal which boroughs of London experience higher rates and which lower. On figure 2 it is seen that boroughs located in east London such as Lewisham and Tower Hamlets are more problematic than those in the west. To explore more granular temporal variation in space monthly map for 2015 was created (Figure 3). Lewisham once again exhibits the highest rates followed by Southwark, Lambeth, Tower Hamlets and Hackney. What is interesting is that all boroughs show a rise in testing rate during August, October, November and December. A very interesting occurrence is that Hillingdon which had low rates of 0.1 to 0.4 during most of the two years, experienced a peak in rate of 0.7 in September.

Figure 2 Map of GLA boroughs displaying total ratio of all chlamydia tests performed during 2015.

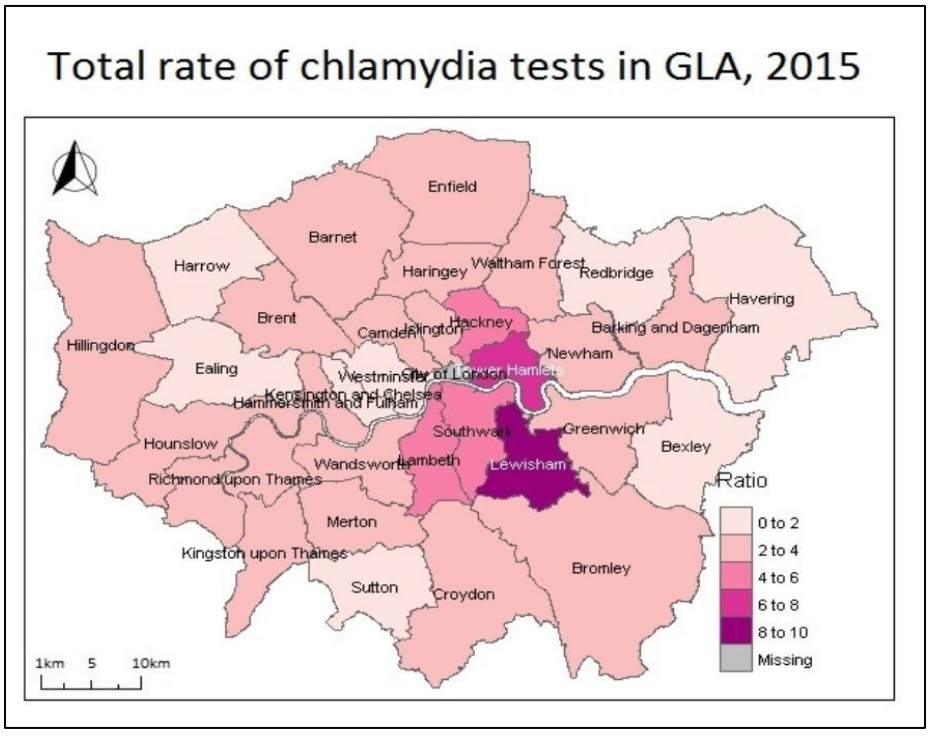


Figure 3 Monthly rate in chlamydia testing in GLA during 2015.

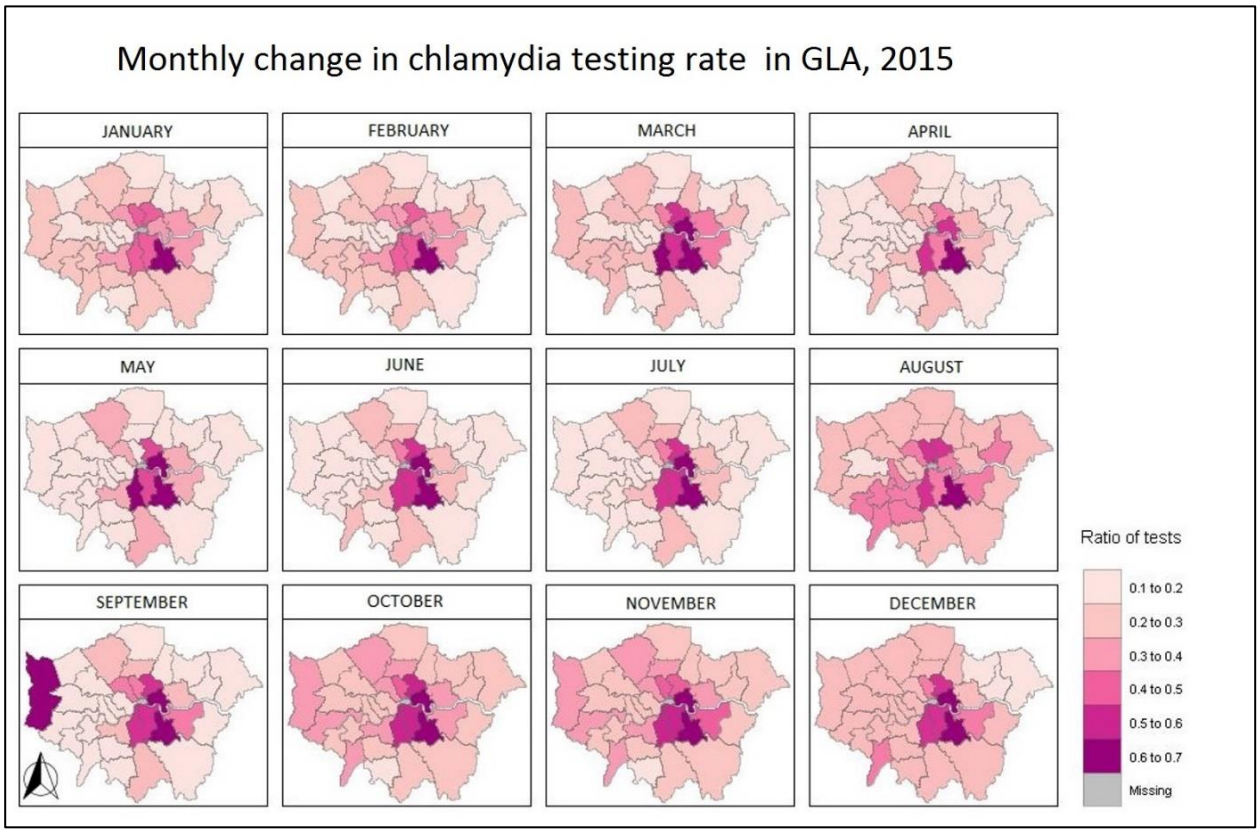
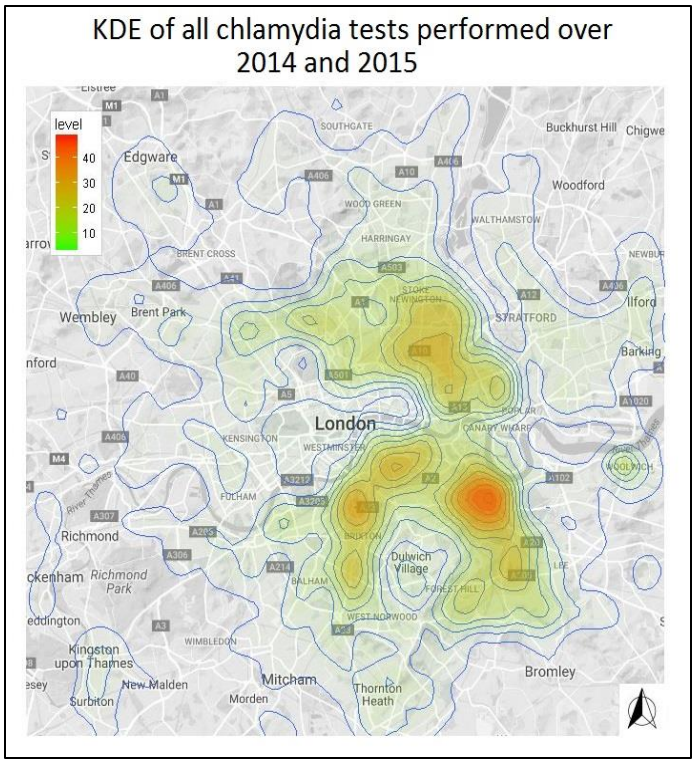
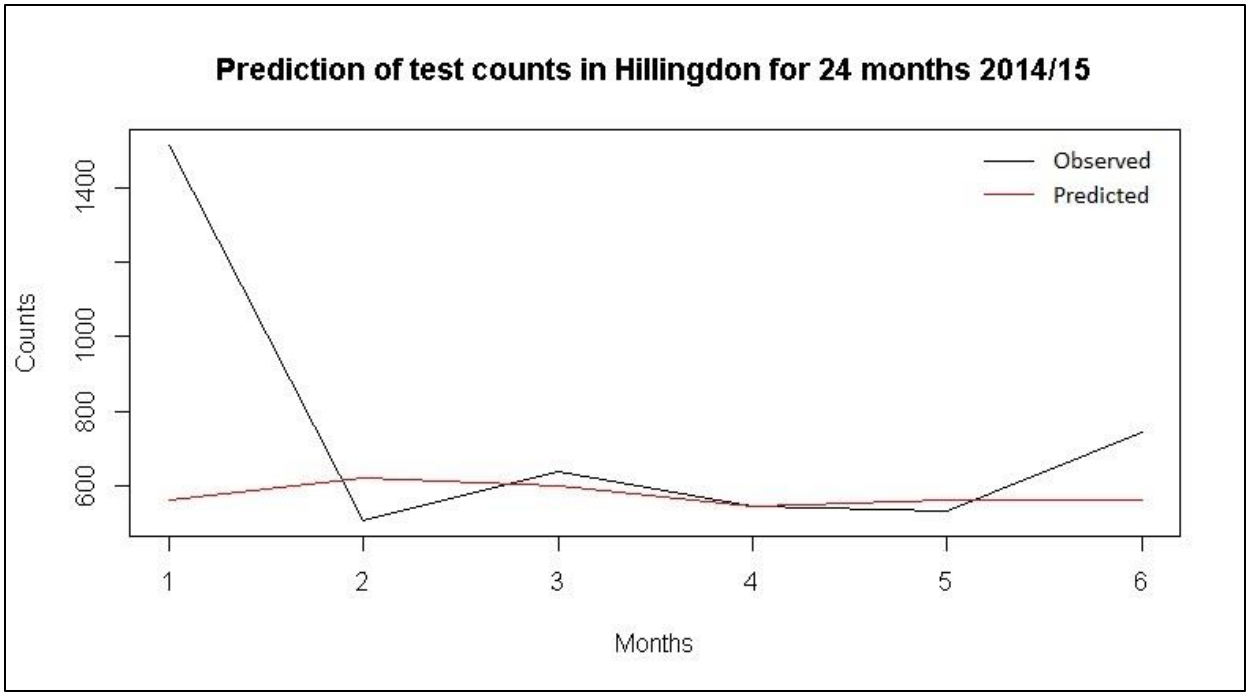


Figure 4 Figure 13 KDE of all chlamydia tests performed during 2014 and 2015 altogether.



To establish whether aggregation of tests to administrative areas has added any bias to the spatial spread of chlamydia KDE maps were produced. Figure 4 shows KDE of all chlamydia tests performed during 2014 and 2015. Three main clusters were identified 1) in Lewisham, 2) Lambeth and Southwark and 3) in the north between Tower Hamlets and Hackney which matched the general trend shown in the choropleth maps.

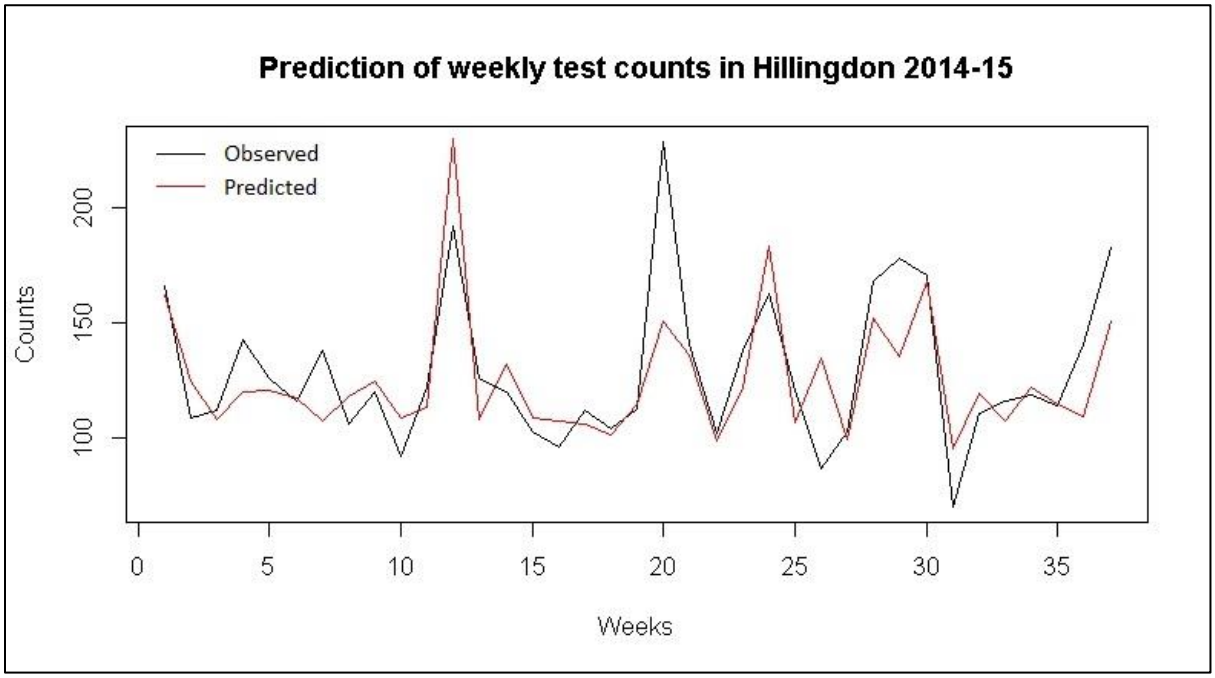
Figure 3 Predicted test counts for monthly time scale in Hillingdon borough with an SVR model.



3.1.SVR

EDA revealed that data were not normally distributed allowing for the choice of non-traditional statistical models. Support Vector Regression (SVR) of time series was performed only at borough level due to the small number of test counts at smaller scales. Two models were created –for monthly and weekly time series, to assess at which time scale the model would perform better for the borough of Hillingdon. First monthly SVR model was built (Fig 5) which did not predict the initial peak in the data, followed the general trend afterwards correctly, however, the prediction does not have high accuracy (RMSE = 157.5376). To check if model performance improves with more granular data, SVR prediction of weekly test counts was performed (Fig 6). Peaks in test counts were accurately predicted providing a good forecast. Reduced number of training data showed improvement in prediction. RMSE of weekly forecast was much lower at 27.80915.

Figure 4 Predicted test counts for weekly time scale in Hillingdon borough with an SVR model.



3.2. ETS

The second model used for test counts prediction was exponential smoothing state space model – ETS; only for monthly temporal scale due to large presence of 0s in weekly data which the model is not good at dealing with (Hyndman, Koehler, Snyder and Grose, 2002). ETS allows for the best model combination of error, season and trend to be chosen based on the AIC which serves to compare performance of different models. Figures 7 and 8 show forecast at borough and postcode level, respectively. It is seen that higher accuracy was achieved at borough level where temporal trend can be detected.

Figure 7 ETS time series forecasting model of Hillingdon borough.

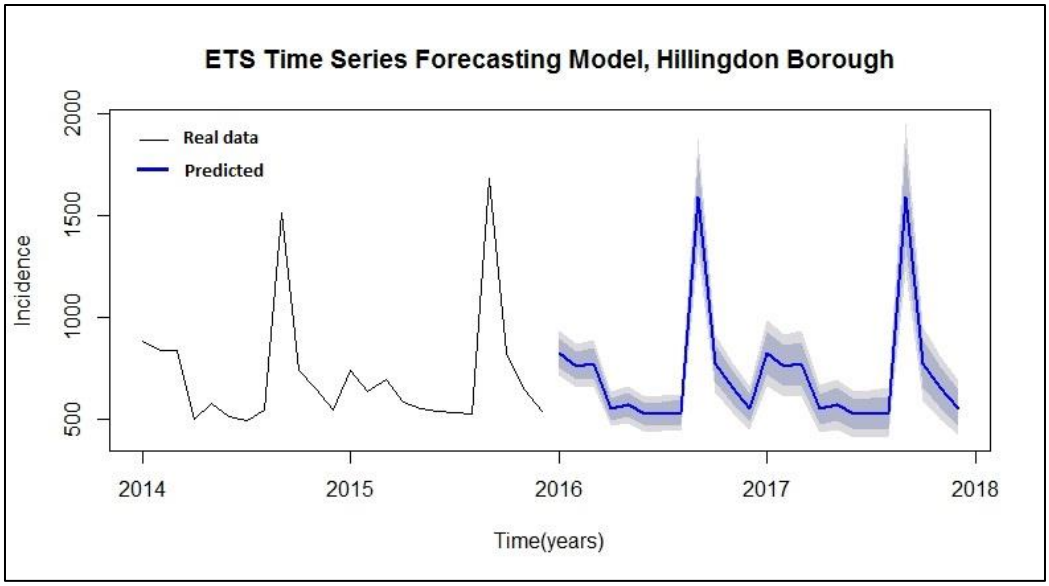
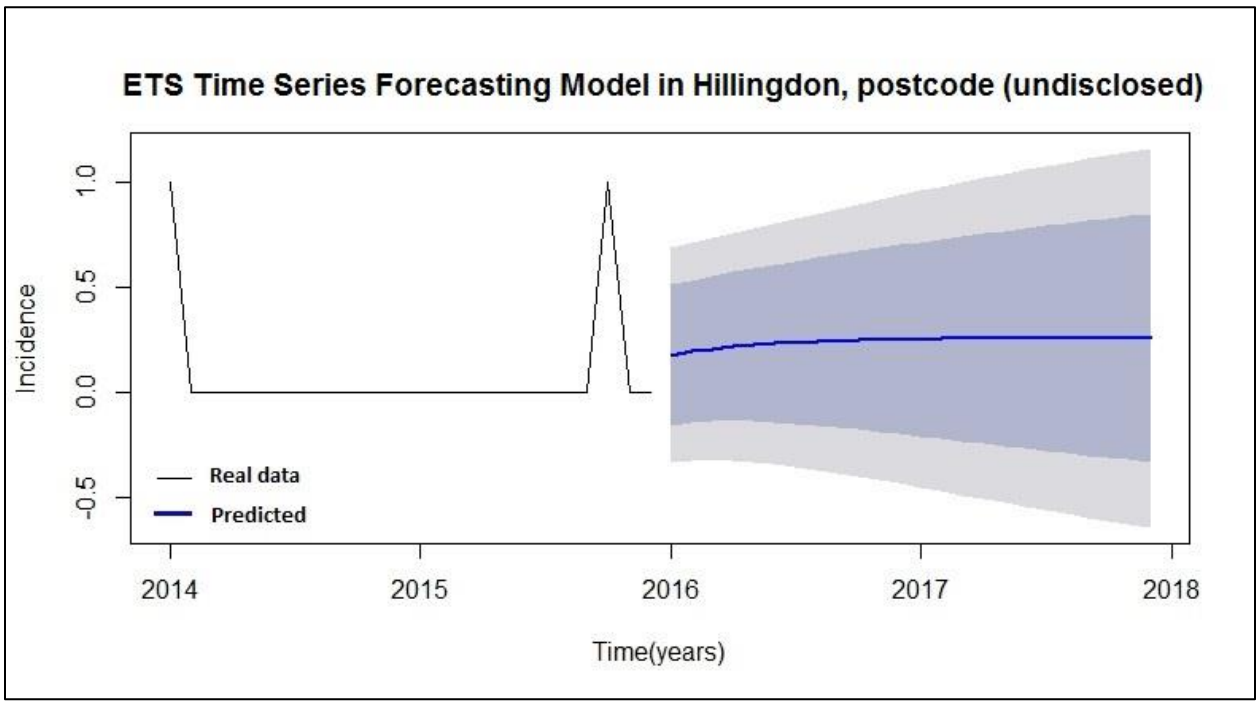


Figure 8 ETS time series forecasting model of Hillingdon, postcode (undisclosed).



3.3.Croston's method

In contrast to ETS, Croston's method was designed to forecast intermittent time series. Therefore, test counts only at post code level were used for prediction as presence of 0s in the time series is a requirement of the model. The model was performed on a monthly and a weekly scale. The constant multi-step ahead monthly forecasts are shown in Figure 9 where it is seen that counts in an undisclosed postcode in Hillingdon are too few causing the smooth factor to be inefficient and thus, the forecast is unreliable. The same procedure was followed for the weekly forecast of the same location. In Figure 10 it is seen that again Croston's method provides an unreliable forecast.

Figure 9 Croston's monthly forecast for a Hillingdon postcode.

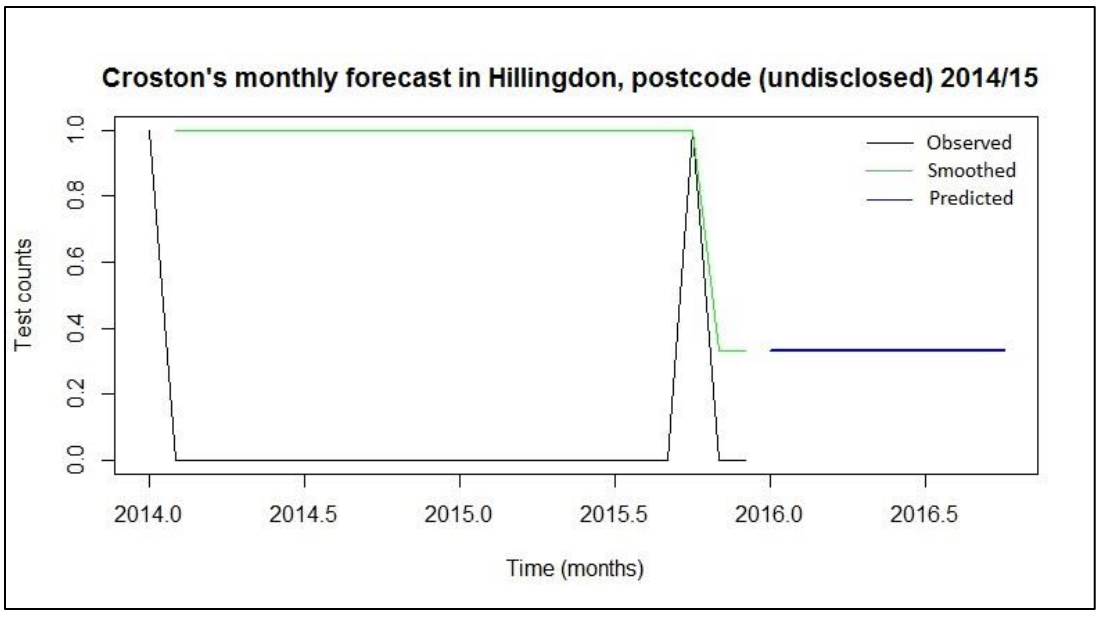
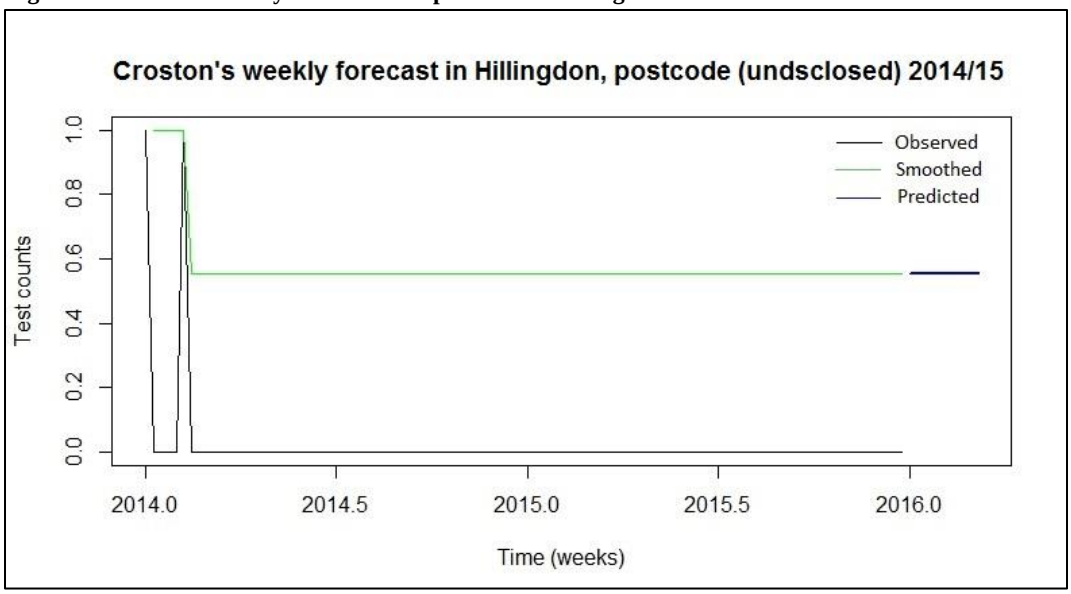


Figure 10 Croston's weekly forecast for a postcode in Hillingdon.



4. Summary

Comparing accuracy of forecasts over two separate spatial and temporal scales can be challenging when several methods based on different methodological principles are used. The most effective approach to compare forecasts within one model and between models was the use of RMSE for SVR and ETS, and MASE for ETS and Croston (Table 1).

It is seen that the combination of spatial and temporal scale is very important for the choice of appropriate prediction model. SVR and ETS showed promising forecasts at weekly and monthly time series at borough level, while at post code level ETS proved better than Croston. Developing an established framework for the prediction of STIs outbreaks can be challenging, however, this study shows that space-time modelling is a promising direction that should be even further explored.

Table 1 Summary of residuals' errors across all applied space-time models and scales.

Model	Error type	Lewisham Borough level		Lewisham Post code level		Hillingdon Borough level		Hillingdon Postcode level	
		weekly	*monthly*	*weekly*	*monthly*	*weekly*	*monthly*	*weekly*	*monthly*
SVR	RMSE	27.9997	120.335	NA	NA	25.4488	463.613	NA	NA
ETS	RMSE	NA	46.5692	NA	0.88662	NA	48.7618	NA	0.62124
	MASE	NA	0.24073	NA	0.55637	NA	0.47906	NA	0.52369
Croston	MASE	NA	NA	0.96446	1.01017	NA	NA	14.1975	5.39130

5. Acknowledgements

Special gratitude is given to Public Health England (PHE) and all the staff of the STI and HIV Department for providing the dataset, workspace, and computers for the project.

6. Biography

Desislava Petrova has a background in geography and the civil service. She graduated from the MSc GIS programme at UCL in September 2016. She currently hold the position of a Data Scientist at Satalia, a company solving hard problems in data science, optimization and artificial intelligence.

Tao Cheng is a Professor in GeoInformatics at UCL whose research interests span network complexity, Geocomputation, space-time analytics and Big data mining (modelling, prediction, clustering, visualisation and simulation) with applications in transport, crime, health, social media, and natural hazards. She has published over 180 refereed papers, and secured research grants over £11 million in the UK, EU and China. She has worked with many industrial partners including Transport for London, the London Metropolitan Police and Arup.

Dr Ian Simms is an Epidemiologist at Public Health England. His professional focus area has been on the geospatial monitoring of syphilis and geodemographic factors affecting the spread of the disease.

7. Bibliography

Adams, E.J., Charlett, A., Edmunds, W.J. and Hughes, G., 2004. Chlamydia trachomatis in the United Kingdom: a systematic review and analysis of prevalence studies. *Sexually transmitted infections*, [online] 80(5), pp.354–62. Available at: <http://sti.bmj.com/content/80/5/354.short#aff-1>.

Becker, K.M., Glass, G.E., Brathwaite, W. and Zenilman, J.M., 1998. Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system. *American journal of epidemiology*, [online] 147(7), pp.709–16. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9554611>.

Garnett, G.P. and Anderson, R.M., 1996. Sexually Transmitted Diseases and Sexual Behavior: Insights from Mathematical Models. *The Journal of Infectious Diseases VO - 174*, [online] 174, p.S150. Available at: <https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.30126281&site=eds-live&scope=site>.

Hyndman, R.J., Koehler, A.B., Snyder, R.D. and Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3), pp.439–454.

Jakob, P., Maurizio, G., Bersabeh, S. and Simms, I., 2015. Identifying and interpreting spatial temporal variation in diagnoses of infectious syphilis amongst men England: 2009 to 2013. *Sexually Transmitted Infections*.

Kanevski, M. and Maignan, M., 2004. *Analysis and Modelling of Spatial Environmental Data*. Lausanne: EPFL Press.

Law, D.C.G., Serre, M.L., Christakos, G., Leone, P.A. and Miller, W.C., 2004. Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. *Sexually transmitted infections*, [online] 80(4), pp.294–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1744854&tool=pmcentrez&rendertype=abstract>.

Lawson, A., 2010. *Statistical Methods in Spatial Epidemiology*. *Wiley Series in Probability and Statistics*. Chichester: John Wiley and Sons.

Pfeiffer, D., Robinson, T., Stevensn, M., Rogers, K. and Clements, A., 2012. *Spatial Analysis in Epidemiology*. 4th ed. Oxford: Oxford University Press.

PHE, 2014. *Annual Epidemiological Spotlight on HIV in London 2014 data About Public Health England*.

PHE, 2016. *Infection report HIV-STIs Sexually transmitted infections and chlamydia screening in England, 2015*.

Waller, L.A. and Gotway, C.A., 2004. *Applied spatial statistics for public health data*. Hoboken: John Wiley & Sons.

WHO, 2016. *Global Health Sector Strategy on Sexually Transmitted Infections 2016-2021 towards ending STIs*.