Title
Analysis of Point Process and SEIR Models for the Spread of Mumps in Pennsylvania

Author
Gao, Yueyan

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Analysis of Point Process and SEIR Models for the Spread of Mumps in Pennsylvania

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Yueyan Gao

2019

ABSTRACT OF THE THESIS

Analysis of Point Process and SEIR Models for the Spread of Mumps in Pennsylvania

by

Yueyan Gao

Master of Applied Statistics

University of California, Los Angeles, 2019

Professor Frederic R. Paik Schoenberg, Chair

Mumps has been long gone from public attention due to developing vaccine programs. In recent year, however, there are random outbreaks of mumps in US, all of which are strongly associated with college campus settings. It is necessary to find out and develop a statistical model with strong forecasting ability to help avoid massive contagion and help surveillance of this epidemic disease in the future. Pennsylvania mumps data is collected from Project Tycho to fit both the point process and SEIR models. Different methods of model evaluation are applied to help determine which one is the best performed. All three methods show quite similarity, but the Recursive model outweighs all others slightly. It is chosen to perform an 75% training vs 25% testing forecasting to see if it is able to catch the dispersal trend of mumps in Pennsylvania. Although the Recursive model predicts well in general, the whole comparison process shed some insights about what could be further done to evaluate and utilize various models.

The thesis of Yueyan Gao is approved.

Nicolas Christou

Jingyi Li

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2019

To my parents…

For their love, support and understanding

Table of Contents

List of Figures

ACKOWLEDGEMENTS

I would like to acknowledge everyone who generously helped me during my academic accomplishments. Foremost, I would like to express my sincere gratitude to each of my thesis committee chair, Dr. Frederic R. Paik Schoenberg, for his guidance with patience, enthusiasm and profound knowledge throughout this thesis.

I would also like to thank the rest of my committee members, Dr. Nicolas Christou and Dr. Jingyi (Jessica) Li, for their generous encouragement and insightful comments.

Finally, I must express my genuine appreciation to my parents and my family for their unwavering support, understanding and faith in me.

CHAPTER 1

Introduction

Mumps is an acute viral disease caused by the mumps virus [1]. During first two to three weeks after exposure, it typically starts with a few days of general unwell feelings such as fever, headache, muscle aches, tiredness, and loss of appetite. Then most people will have swelling of their salivary glands, which causes the puffy cheeks and a tender, swollen jaw. [2].

Mumps is extremely contagious, whose virus can spread through coughing and sneezing and close-contact activities. Mumps virus can be spread by an infectious people before or after several days of their swelling symptoms begins [2]. Although the highest risk of contracting mumps is to children [3], symptoms are often more severe in adults and possible further illness include meningitis, pancreatitis and testicular atrophy [4]. Before the U.S. mumps vaccination program started in 1967, about 186,000 cases were reported each year, but the actual number of cases was likely much higher due to underreporting [5].

The best precautionary measure is to get vaccine and mumps is preventable with the safe and effective MMR vaccine. Since the introduction of improved MMR vaccination program in 1989, U.S. mumps cases decreased more than 99%. However, there are mumps outbreaks reported increasingly in US since 2006 and shows a pattern of 5-year cycle [5]. Recent mumps outbreaks in US have been strongly associated with college campus settings during 2005 -2019 [6].This unusual outbreak of mumps within nationwide college campus kindled my interest to look into the

infectious process of mumps and to figure out how statistical models can be applied for research and forecast of similar type of diseases.

There are three models applied and compared in this paper. Hawkes models is a self-exciting point process proposed by Alan G. Hawkes [7]. Hawkes models are currently commonly seen in applications to seismology but hardly in research of infectious diseases [8]. In order to extend Hawkes models in cases where productivity (expected number if transmission) is not static, a new type of point process model, recursive point process model [9], was introduced to offer a more precise account of clustering. Also, a compartmental model named the SEIR (Susceptible-Exposed-Infected-Recovered) is included in discussion because it is shown effective to describe the dynamics of the Ebola virus [8]. The comparisons among three models may help us develop a better insight and predictions of the spread of mumps, especially between the first two as point process and the last as compartmental model.

The structure of this paper is as follows. Following a description of the mumps dataset in Chapter 2, we briefly review all three models, including Hawkes self-exciting point process models, recursive point process models and SEIR models, in Chapter 3. Model fitting and evaluation methods are discussed and explained in Chapter 4. Followed by Chapter 5, we would apply the best model to train part of data and conduct forecasting using remaining test data. Finally, Chapter 6 contains some concluding remarks and future discussion.

CHAPTER 2

Dataset

The United States of America Mumps dataset [10] is obtained from Project Tycho, an open-access research platform for global health, particularly disease surveillance data compiled from reputable sources such as the United States Centers for Disease Control or the World Health Organization.

US Mumps dataset contains case counts for mumps in United States reported during 1923 to 2017 with 165242 records. As stated in its description, this dataset also includes information about these attributes, such as the location, age group, the source where Project Tycho team obtained case counts, and etc [10]. Because of our interests, we use cases happened in Pennsylvania in this paper.

After selecting Pennsylvania as targeted location from the original dataset, the filtered data contains counts of confirmed cases of mumps in Pennsylvania by week. In order to prepare for model fitting, the onset time for each individual case was drawn uniformly within each seven-day time interval [9], therefore transferred as a list of sorted numbers starting from 0. Further inspection shows that the dataset consists of continuous time periods without any missing weekly data input, so no more data cleaning process is complete. There are total 13948 cases from Jan. 1970 to Dec. 2017.

CHAPTER 3

Point Process & SEIR Models

3.1 Point Process

3.1.1 Introduction

A point process is a locally finite collection of random elements in some space *S* [11]. Assuming *S* is bounded [0, T] in time, *B* is a subset of some complete separable metric space equipped with Lebesgue measure, *L*, and the spatial region is scaled so that $L(B) = 1$, then $K = B \times [0,T]$ is extended to a bounded region in space-time [9].

A point process on the real line is particularly amenable to study [12] because the whole process can be described naturally by the randomness between the points. Application of point process can be found in a wide range of fields from epidemiology to economics.

A temporal point process is typically modeled via its conditional intensity, $\lambda(t)$, which represents the infinitesimal expected rate at which points are accumulating at time t, given information on all points occurring prior to time t [9].The simplest and most important example of a point process is the Poisson point process.

3.1.2 Hawkes Models

A point process *N* may be called self-exciting [13] if $cov\{N(s,t), N(t,u)\} > 0$ for $s < t < u$. *N* is self-correcting if instead this covariance is negative. Thus the occurrence of points in a self-exciting

point process causes other points to be more likely to occur, whereas in a self-correcting process, the points have an inhibitory effect [13].

A purely temporal Hawkes process is a self-exciting point process model. Let t_i be the ith occurrence of the point process prior to time t, then the conditional intensity $\lambda(t)$ of Hawkes model is given by

$$\lambda(t) = \mu + K \sum_{i:t_i<t} g(t - t_i), \mu > 0$$

Apparently, occurrence t_i contributes a secondary series of occurrences (aftershocks) occurring at a time varying rate $Kg(t - t_i) \geq 0$, which in turn produces its own aftershock sequence, and so on [8]. K represents the expected number of new infections directly attributable to each case, thus, to be stable, $0 \leq K < 1$ [8]. This "branching" attribution, that is, occurrence of some points makes other more likely to happen, is an indication that Hawkes process may be used for the following infectious diseases study.

For many processes, the triggering density $g(u)$ decays gradually as the time delay u increases [8] and a common choice for $g(u)$ could be an exponential density function.

3.1.3 Recursive Models

Recursive point process model is a model where the productivity for a subject infected at time t is inversely related to the conditional intensity at time t [9]. Similarly, the conditional intensity $\lambda(t)$ of recursive models is given by

$$\lambda(t) = \mu + \sum_{i:t_i<t} H(\lambda(t_i))g(t - t_i), \mu > 0$$

where the function $g(u) \geq 0$ is still the triggering function, but the productivity of any point t_i is given by $H(\lambda(t_i))$. Thus the total productivity, for n points $t_1, t_2, ..., t_n$ is $\sum_{i=1}^{n} H(\lambda(t_i))$ [9].

This shows major difference between Hawkes models and Recursive models. Hawkes process has a static productivity, whereas Recursive process has changing productivity in compliance with situation. Since assumption of static productivity seems unreal for actual infectious disease due to circumstances such as human intervention, recursive models may outperform in terms of precision when dynamic situation occurred.

3.2 SEIR Models

The SEIR (Susceptible-Exposed-Infected-Recovered) compartmental model takes the period of time during which individual is infected but not yet infectious into consideration. It is a model commonly used in forecasting the dynamics and duration of an epidemic. The infectious rate at time t, $\beta(t)$, represents the probability of transmitting disease between a susceptible and an infectious individual. The incubation rate, σ, is the rate of latent individuals becoming infectious (average duration of incubation is $1/\sigma$). Recovery rate, $\gamma = 1/D$, is determined by the average duration, D, of infection [14].

In a closed population with no births or deaths, N(total population) = S + E + I + R where S, E, I, R are all population of each step [8]:

$$\frac{dS}{dt} = -\beta(t)\frac{SI}{N}$$

$$\frac{dE}{dt} = -\beta(t)\frac{SI}{N} - \sigma E$$

$$\frac{dI}{dt} = \sigma E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Under this model, the transmission rate, β(t), is assumed to decline exponentially at rate κ [8]:

$\beta(t) = \beta e^{-\kappa t}$ (t=number of days from the start of outbreak)

Also, the reproductive number, $R_0(t)$, represents the average number if new infections generated by an infected individual until death or recovery: $R_0(t) = \beta/\gamma$. It has a critical value of 1: if $R_0(t) > 1$, the epidemic can spread massively; otherwise, the epidemic is unsustainable [8].

The application of SEIR model is reasonable because of the special spreading behavior of mumps. As mentioned during introduction in Chapter 1, it is believed that a lag that fits SEIR model may exist because mumps virus can be spread by an infected people some days before or after their swelling symptoms begins. In next chapter, I will compare those two genres of models to see which fits and predicts better by examining different evaluation factors.

CHAPTER 4

Model Fitting & Evaluation

As discussed in Chapter 2, weekly cases of mumps in Pennsylvania are recorded from Jan 1970 to Dec 2017. Figure 4.1(a) displays a histogram of the cases. There is no obvious gap but there is a rebound around 1994 after the disease almost died out. Overall, most peaks appear in the early years and we can see the number of events in recent years clearly becomes fewer than previous years, which may be related with popularization of vaccines.

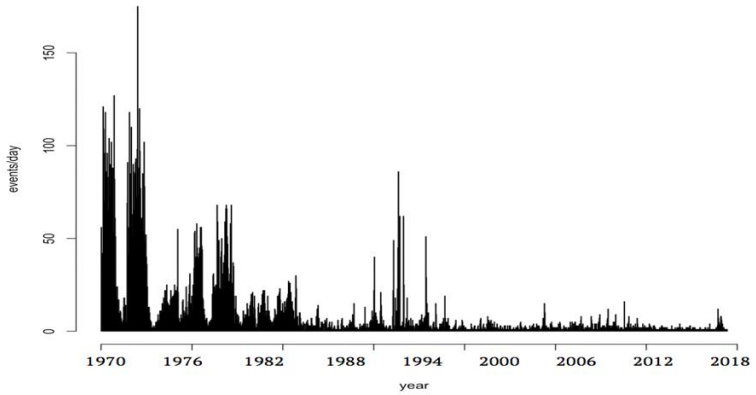


Figure 4.1(a): Histogram of Mumps in Pennsylvania from Jan 1970 to Dec 2017

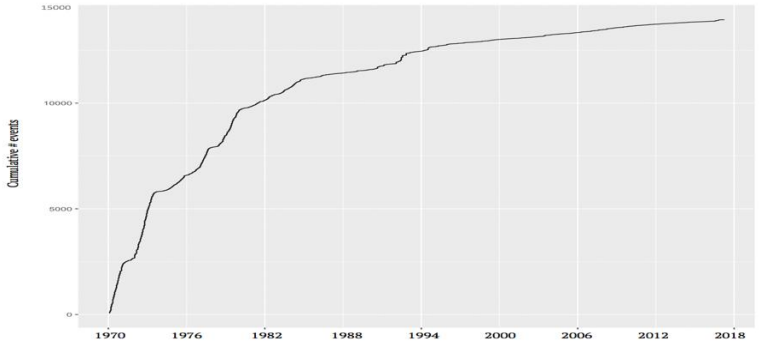
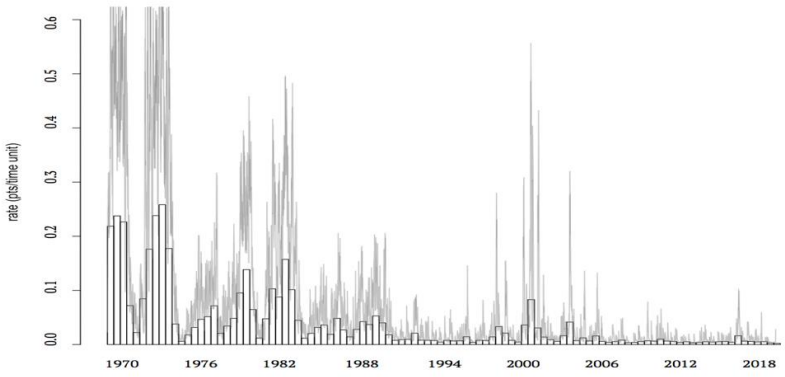


Figure 4.1(b): Cumulative Number of Mumps in Pennsylvania from Jan 1970 to Dec 2017

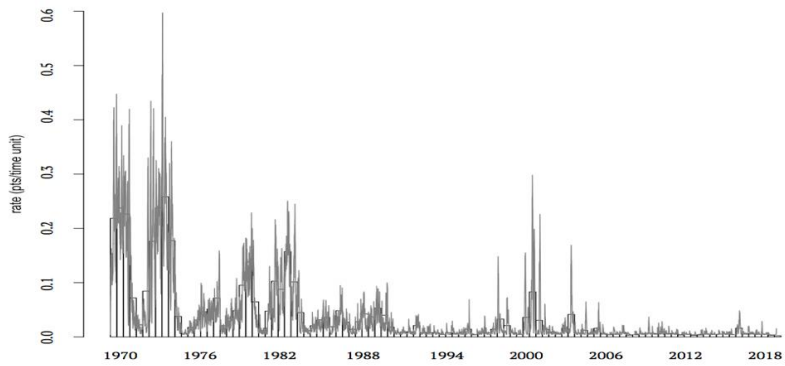
Figure 4.1: Visualization of Mumps in Pennsylvania dataset (Jan 1970 to Dec 2017)

4.1 Model Fitting

In this section, we fit three different types of models: Hawkes models with power-law triggering function, recursive model with exponential function, as well as SEIR model. It would be interesting to see performance difference both within different point process models and between point process and SEIR models.



(a) Estimated rate of Hawkes model with power-law function



(b) Estimated rate of recursive model with exponential function

Figure 4.2: Histogram of Mumps in Pennsylvania along with different estimated rates of point process models

Firstly, a Hawkes model is used with power-law function $g(u) = (p-1)c^{(p-1)}(u+c)^{(-p)}$. The estimated parameters are (μ, κ, c, p) = (0.1659192 points/day, 0.9354747 triggered points/observed point, 2.0829354 points/day, 2.4667649), with corresponding standard error estimates (0.013547852, 0.009649899, 0.314763178, 0.170221767). The estimated conditional intensity of the fitted Hawkes model is shown in Figure 4.2(a).

To compare with the Hawkes models, a recursive model is also fitted with exponential function to this data with $g(u)= \beta e^{-\beta(t-t(i))}$ with $\lambda = \mu + \sum K \beta e^{-\beta(t-t(i))}$, where $K = c\lambda^{-p}$. The recursive model with exponential function is identical to Hawkes models with exponential function but with one more parameter. The estimated parameters are (μ, c, β, p) = (0.2501784 points/day, 0.7271388 triggered points/observed point, 0. 6523949 points/day, -0.1091504), as well as corresponding standard error estimates (0.010684196, 0.015015177, 0.025570487, 0.008689334). Figure 4.2(b) shows the estimated rates from recursive model.

From Figure 4.2, it is easy to observe and deduce that Recursive model should perform much better than the Hawkes model since the trend is much more captured and followed in the second plot. Therefore, in order to prove our conjecture, we will proceed to evaluate the models in several statistical ways.

Using completely different parameters compared to the point process models, SEIR models are also applied with an outcome of $(\beta_0, \kappa, f, \sigma, \gamma)$ = (0.2446231918, 0.008056205, 0.0094300521, 0.1886792453, 0.1782531194). The reproductive number, $R_0(t)$, is calculated to be 1.372336 with an 95% confidence interval as (1.370786, 1.373888). Since it is over 1, it proves that mumps are very infectious and able to spread massively.

4.2 Model Evaluation

Since estimates of parameters and corresponding standard error are calculated in the previous part, it is essential to evaluate and compare each model, so that we can obtain the best fitted model to conduct further forecasting on the dataset. We shall start with the evaluation between two point-process models, followed by a comparison among them and SEIR model.

4.2.1 Point Process Models Comparisons

First of all, it is important to figure out what values can be tested to see if the results are reasonable and what methods can be applied accordingly.

There are two properties that can be used to check reasonability:

(1) Stoyan-Grabarnik diagnostic: Stoyan and Grabarnik [15] were first to exploit a diagnostic formula for point process model checking. Let x denote a point pattern dataset, consisting of the time x_1, ..., x_n of events observed in a special region W. Then attach weight/mark $x_i = 1/\lambda(x_i, X)$ to each x_i where λ denotes the conditional intensity of the model. Stoyan and Grabarnik proved that $E\sum_{i=1}^{n}\frac{1}{\lambda(x_i,X)} = E\int\frac{1}{\lambda(x)}du = E\int 1du = |B|$, where B denotes an area in W. In our case, B is the time set T, and hence, $\sum\frac{1}{\lambda_i}/T \sim 1$. They suggested this property to be used for exploratory data analysis and goodness-of-fir testing [16].

(2) Harte's Ratio: Another way to check that the estimates are reasonable is developed by Harte [17]. It is a commonly computed ratio $\int_0^T\frac{\hat{\lambda}(t)dt}{N(0,T)}$ which should be close to 1 as well. The whole proof process is basically similar to what we have mentioned above.

Both criteria could be applied to the point process models to measure if their results are of good fit. The Harte's ratio for Hawkes models with power-law function is 0.9998850 and Stoyan-Grabarnik diagnostic is 0.9729088. For the recursive model, we got a Harte's ration equal to 1.000064, and Stoyan-Grabarnik diagnostic is equal to 1.005441. Both of these two diagnostics for all three models are approximate to 1, but recursive model's numbers perform better by being closer to 1, especially on the Stoyan-Grabarnik diagnostic value, than the other Hawkes model. This result indicates the estimates provided by both point process models are all good and reasonable, whereas recursive model shows a better fit of parameters.

4.2.2 Akaike Information Criterion (AIC)

Then, in order to statistically support our selection of model, Akaike information criterion [21] is a critical indicator that can be used. It represents the relative quality of statistical models for a given dataset. When given a collection of models, AIC, which values on both the goodness of fit and the simplicity, estimates not only the quality of each model, but also their relative performance to each other.

Given a statistical model of some data, let p be the number of estimated parameters in the model, and $\hat{L}$ be the maximum value of the likelihood function for the model. AIC is defined as $AIC = 2p - 2\ln(\hat{L})$ [21]. The preferred model is the one with the minimum AIC value.

Here we calculate the log-likelihood for each model first. For the Hawkes model with power law function, the log-likelihood is 8433.404; the log-likelihood for Recursive model is 8717.235; the log-likelihood for SEIR model is -5260.655. From above, we get AIC value is equal to -16858.808 for Hawkes, -17426.47 for Recursive and 10527.31 for SEIR. Hence, with minimum

AIC value for this Mumps dataset, Recursive model performs relatively better than the other two. However, there are some possible dispute of AIC value comparisons in this case, which would be covered and discussed in Chapter 6.

4.2.3 Superthinning

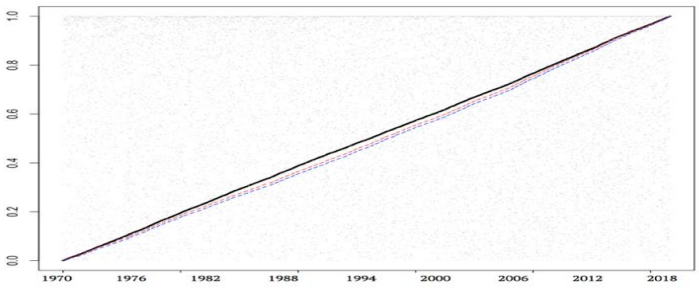
Next step, super-thinning process is applied to evaluate all three models including the SEIR. Super-thinning residuals [18] is a hybrid approach where one thins [19] in areas of high intensity and superposes [20] simulated points in areas of low intensity, resulting in a homogeneous point process if the model for λ used in the thinning and superposition is correct.

Thinning is defined as a process such that each observed point is retained independently with probability $\frac{\mathrm{b}}{\lambda(\mathrm{x}_i,\mathrm{t}_i)}$, where $b = \inf\{\lambda(x,t)\}\, for\ (x,t) \in S$ [19]. If b is small, the power of thinning may suffer from too few points and little power to detect inhomogeneity. Superposition, on the other hand, has weakness when its assumption c= $sup\{\lambda(x,t)\}\, for\ (x,t) \in S$ is large. Too many points would be generated with the simulated rate $c - \lambda(\mathrm{x}_i, \mathrm{t}_i)$ [20].

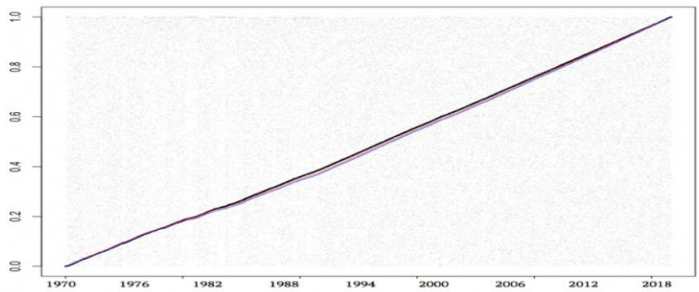
Therefore, it is necessary to introduce a combined method, super-thinning, so that it contains neither too few nor too many points. In super-thinning, it requires an initial choice of the tuning parameter, *b*, and as suggested in Clements et al. [18], we used the simple default value of the total number of cases divided by the length, in days, of the observation period [8]. Then, original data are thinned so that all are independently random with probability min{1, $b/\hat{\lambda}(t)$} and new points are superposed over with rate $(b - \hat{\lambda}(t))^+$ [8]. The outcome is a homogeneous Poisson process with rate b if and only if the estimated conditional intensity $\hat{\lambda}$ is correct [18] and, consequently, the resulting residuals can be used as an assessment for model's uniformity.

In the following page, Figure 4.3 shows the result of super-thinned residuals for all three models. The solid black line shows the cumulative sum of the standardized interevent times for each residual. The dotted blue and red lines show lower and upper 95% confidence interval separately, based on 1000 simulations of same amount of uniformly distributed random variables. These plots are graphical descriptions expression of how well the models fit in the most straightforward way.

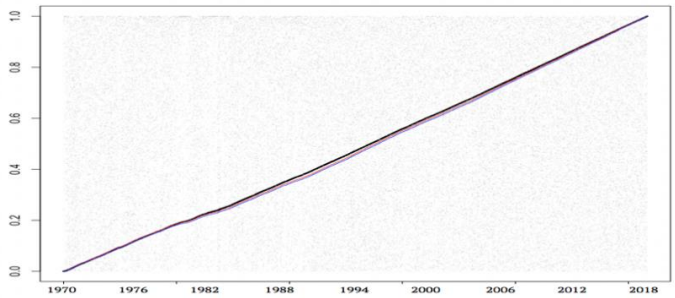
In general, the super-thinned residuals for three models appear to be well scattered without any unusual gap. However, in figure 4.3(a), the normalized cumulative sum is overall higher than its upper confidence bound with an obvious disparity from 1982 to 2006. In figure 4.3(b), the normalized cumulative sum is also slightly higher than the upper bound from 1988 to 2006. And we also can see the similar pattern in figure 4.3(c) from 1988 to 2000 where the cumulative sum almost coincides with the upper bound line. From the graphs, we find out similar problem with different level of visibility, which might be related to the popularity of improved MMR vaccine program since 1989.



(a) Super-thinned residual plot for Hawkes model



(b) Super-thinned residual plot for Recursive model



(c) Super-thinned residual plot for SEIR model

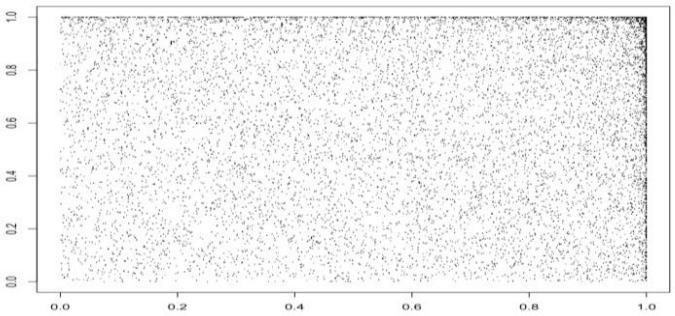
Figure 4.3: Super-thinned residuals plots with their 95% confidence interval for three models. Y-coordinates are uniform(0,1) random variables which represent standardized interevent times.

Figure 4.4 shows the lag plots of the standardized interevent times of the super-thinned residuals. A lag plot is used to examine whether the values in a dataset are random [22]. If no identifiable pattern is shown in the graph, then the data are random; otherwise, the data are not random if some obvious patterns can be found. The type of pattern helps in identifying the non-random portion and outliers. For example, the level of autocorrection can be decided according to how tight the points tend to cluster along the diagonal [23].

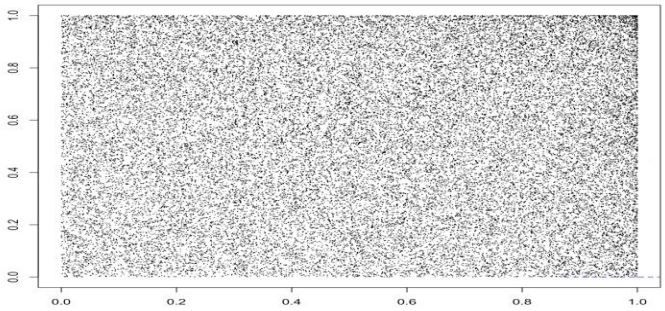
Here, in figure 4.4(a), for Hawkes model, there are more points on the upper and right borders, especially in that corner, than randomness. Comparatively, in figure 4.4(b) and (c), the lag plots for the Recursive and SEIR models look alike and evenly scattered, although points are a little more clustered than expected in the upper right corner and, especially in SEIR models, points tend to concentrate more on the right hand side of the graph.

4.3 Summary

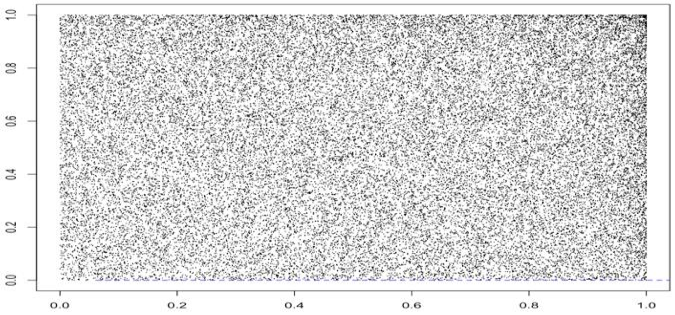
In conclusion, the Recursive model outperforms both Hawkes model and SEIR model in modeling spreading process of mumps. As a result, we will be focusing on Recursive model in the following chapter to forecast using train and test data to see how well it predicts.



(a) Lag plot for Hawkes model



(b) Lag plot for Recursive model



(c) Lag plot for SEIR model

Figure 4.4: Lag plot of the standardized interevent times of the super-thinned residuals for three different models

CHAPTER 5

Out-of-Sample Forecasting

5.1 Training & Testing Dataset

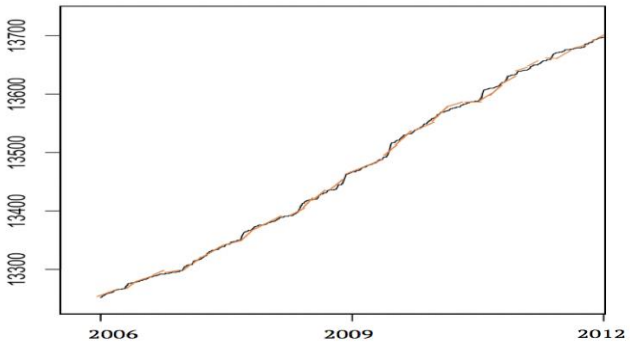
In order to simulate out-of-sample forecast with recursive model, we separate the total mumps data into two different parts, the first 75% of dates, from Jan. 1970 to Jan. 2006, as training data and the rest 25% of dates, from Jan 2006 to Dec. 2017, as testing data. We will use the former part to fit the Recursive model, and then use the latter part to evaluate our models.

To begin with, the recursive model is fitted to the training data, which returns a set of estimated parameters (μ, c, β, p) = (0.27465275 points/day, 0.75966956 triggered points/observed point, 0.65001191 points/day, -0.08979429). For any given week, we would build a model with the parameters from training, then fit all the data up to the beginning of the week that needs to be predicted. To do forecast, we should calculate the product of the mean lambda and days of week (7). Since the model is exponentially distributed, the result is not only the total estimated weekly lambda, but also the variance of the number of predicted events in the week.

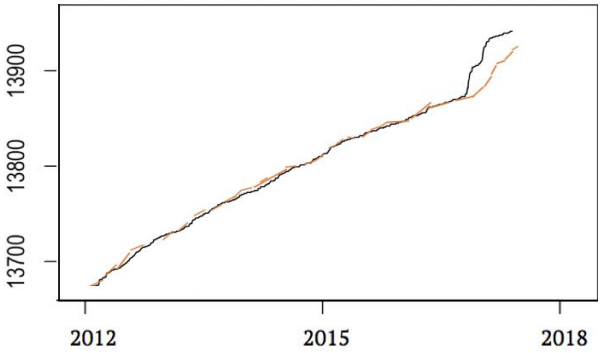
5.2 Weekly Forecasting

There is a long twelve-year gap in between 2006 and 2018, which includes a large amount of weeks. In order to present the details more clearly, we would divide the forecasting into two parts, each with 3 years. Figure 5.1 (a) shows the cumulative forecasts during 2006 and 2012 and Figure 5.1(b) shows that of 2012 to 2018. The red line in the plot is our forecasted number of events from

recursive model, and the black solid line is the plot of actual data. In general, the estimations and real data points are matched well.



(a) Cumulative forecasts with the observed number of events (2006-2012)



(b) Cumulative forecasts with the observed number of events (2012-2018)

Figure 5.1: Prediction by fitting recursive model (2006-2018)

Apparently, the cumulative forecasts follow the observed number of events more precisely during the first half of test data. In Figure 5.1(a), there are some minor deviations at the beginning of 2006, which marks the re-appearance of contagious mumps disease in recent years. Also, we can see some results are lower than actual numbers when unexpected sharp increase happened

around 2010-2011. We realize that forecasts may not do well when dramatic changes or sudden fluctuations appear in a short period.

In figure 5.1 (b), although most of the estimated values are quite accurate before 2017, there are substantial differences between actual number of events and our estimated values. In 2012-2017, the curve for the cumulative number of events tends to be smoother and follow the trend from previous years which makes the forecasting fit. However, after 2017, there is a sharp growth which, as we deducted in last paragraph, largely affects the accuracy of the recursive model. The forecasting followed the overall trend but did not catch up with actual figures. It may suggest that there is a changing pattern of mumps-infecting process in recent years.

Lastly, to evaluate our forecasting statistically, RMSE can be calculated, by comparing the estimated cumulative predicted to the observed number on the last day of each week, to measure the accuracy of our model. The total RMSE is approximately 20.42, indicating that the difference between our forecasts and the actual observed cases is around 20 events/day and this number could be largely improved if the forecasting in 2017 is not counted. Given the long period of time contained in the testing dataset, I believe we have achieved a pretty good forecasting model.

CHAPTER 6

Concluding Remarks

6.1 Conclusions

The application of Recursive point process model to predict the spread of mumps in Pennsylvania indicates that these point process methods have the potential to be a useful addition into disease forecasting research. In all aspects of fitting and evaluation of the mumps spread that we performed, Recursive point process models with exponential function performed as well as, or better than, both Hawkes point process models and SEIR models. However, it does not mean that the other two models show no use in this situation; rather, they actually perform well enough that it may be worthwhile to look into and may shed new insights to how outbreaks and infecting process develop.

Based on AIC value and super-thinning results, recursive models show a consistent better performance over the other two. By comparing the estimated rates of model fitting between two point-process models, we can see an effective improve of accuracy from traditional Hawkes models to Recursive models, which indicates that, due to flexibility, Recursive models is easier to adjust according to dynamics of the outbreaks. From forecasting through a long period of time, we further prove the good prediction ability of Recursive models. Hence, I believe that recursive models could help us to investigate how spreading and outbreaks of mumps in Pennsylvania would work in advance.

6.2 Future Discussion

Although there is no obvious gap in the mumps in Pennsylvania data, there are some unused part of data in the original dataset. They are not clearly labeled, some of which are duplicate of the others. With that being said, they, very likely an ongoing update, may create a difference to the dataset and model if properly added to the existing dataset. This reminds me of the importance to verify the accuracy and integrity of our data in future research. In addition, since data are originally collected in a weekly manner, it is necessary but not ideally precise to artificially generate randomized times of events within each week.

Another aspect that worth investigate is about how SEIR model can be applied with fuller assumptions. As mentioned in Chapter 3, SEIR models take "lag" effect into consideration, that is, the time between being infected and becoming infectious. Although mumps virus may fit into this lag effect, there is hardly any further details to elaborate. In addition, it may inappropriate to compare the AIC value, which is calculated from loglikelihood of each, of point process and SEIR model in Chapter 4. The former is a sum of each point of events while the latter is computed using total events on weekly basis. This method is questionable to take place in comparing two utterly different types of models. Therefore, future research into this case may provide better insights into model selection.

Last but not least, when comparing the two types of point process models, Hawkes model is formulated with power-law function while Recursive model is with exponential and both use four parameters. SEIR is the only model constructed with three variables. One might object that, in retrospective analysis, the improvement in fit might due for utilizing a more complex model with

more free parameters, in which case overfitting could be a potential problem and the improvement would be unlikely to be maintained in further applications, particularly in forecasting [8]. In my future research, I would like to evaluate out-of-sample simulating performance during an unusual outbreak to see which one would perform more accurately in forecasting the spread of epidemic diseases, which could help prevent an outbreak in advance.

References

[1] Atkinson, William. Public Health Foundation. *Mumps Epidemiology and Prevention of Vaccine-Preventable Diseases* (12th ed.). pp. Chapter 14. Archived July 2016.

[2] Centers of Disease Control and Prevention. Signs & Symptoms of Mumps. https://www.cdc.gov/mumps/about/signs-symptoms.html. Mar 2019.

[3] John Mersch, MD, FAAP. Mumps. Medicine Net: https:/medicinenet.com/mumps/article.html. Aug 2019.

[4] Weekly Epidemiological Record contributor. "Mumps virus vaccines" (PDF). Weekly Epidemiological Record. 82 (7): 49–60. February 2007. Archived (PDF) March 2015.

[5] Centers of Disease Control and Prevention. Mumps: cases and outbreaks. https://www.cdc.gov/mumps/outbreaks.html. Jul 2019.

[6] Wikipedia contributors. Mumps outbreaks in the 21st century. https://en.wikipedia.org/wiki/Mumps_outbreaks_in_the_21st_century. Sep 2019.

[7] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83, 1971.

[8] Junhyung Park, Adam W. Chaffee, Ryan J. Harrigan, and Frederic Paik Schoenberg. A non-parametric hawkes model of the spread of ebola in west africa. December 2018.

[9] Frederic Schoenberg, Marc Hoffmann, and Ryan Harrigan. A recursive point process model for infectious diseases. *arXiv e-prints*, page arXiv:1703.08202, March 2017.

[10] Van Panhuis W., Cross A., Burke D., Counts of Mumps reported in UNITED STATES OF AMERICA:1923-2017(version 2.0). Project Tycho data release, DOI: 10.25337/T7/ptycho.v2.0/US.36989005. April 2018.

[11] Wikipedia contributors. Point Process. https://en.wikipedia.org/wiki/Point_process. Sep 2019.

[12] Last, G., Brandt, A. Marked point processes on the real line: The dynamic approach. *Probability and its Applications*. Springer, New York. 1995.

[13] Frederic Paik Schoenberg. Introduction to point processes. 2000.

[14] https://institutefordiseasemodeling.github.io/Documentation/general/model-seir.html

[15] Dietrich Stoyan and Pavel Grabarnik. Secondorder characteristics for stochastic structures connected with gibbs point processes. *Mathematische Nachrichten*, 151:95–100, 1991.

[16] Adrian Baddeley. Residuals and Diagnostics for Spatial Point Processes. https://pdfs.semanticscholar.org/646c/b35f930135c61dfe7962073482e2c8a1db91.pdf. 2019.

[17] D.S. Harte. Log-likelihood of earthquake models: Evaluation of models and forecasts. *Geophysical Journal International*, 201:711–723, March 2015.

[18] Robert Alan Clements, Frederic Paik Schoenberg, & Alejandro Veen. (2011). "Evaluations of Space-time Point Process Models Using Super-thinning." In: *Environmetrics* 23.7 (2013), pp.606-616

[19] Frederic Paik Schoenberg. Multidimensional residuals analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98(464):789–795, December 2003.

[20] Pierre Br´emaud. Point Processes and Queues: Martingale Dynamics. *Springer Series in Statistics.* Springer, 1981 edition, September 1981.

[21] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, *Budapest: Akadémiai Kiadó*, pp.

267-281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), *Breakthroughs in Statistics*, I, Springer-Verlag, pp. 610-624.

[22] NCSS Statistical Software. “Chapter 164: Lag Plots”. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Lag_Plots.pdf. 2019

[23] NIST SEMATECH contributors. “1.3.3.15. Lag Plot”. *Engineering Statistics Handbook*. https://www.itl.nist.gov/div898/handbook/eda/section3/lagplot.htm#examplesThe%20lag%20plot%20for%20three%20models. 2019.