SANDIA REPORT

SAND2018-11533
Unlimited Release
Printed October 2018

Synthetic data generators for the evaluation of biosurveillance outbreak detection algorithms

Drew Levin, Patrick D. Finley

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

SAND2018-11533
Unlimited Release
Printed October 2018

Synthetic data generators for the evaluation of biosurveillance outbreak detection algorithms

Drew Levin
Operations Research and Computational Analysis
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1188
dlevin@sandia.gov

Patrick D. Finley
Operations Research and Computational Analysis
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1188
pdfinle@sandia.gov

Abstract

The research and development of new algorithmic and statistical methods of outbreak detection is an ongoing research priority in the field of biosurveillance. The early detection of emergent disease outbreaks is crucial for effective treatment and mitigation. New detection methods must be compared to established approaches for proper evaluation. This comparison requires biosurveillance test data that accurately reflects the complexity of the real-world data it will be applied to. While the test and evaluation of new detection methods is best performed on real data, it is often impractical to obtain such data as it is either proprietary or limited in scope. Thus, scientists must turn to synthetic data generation to provide enough data to properly evaluate new detection methodologies. This paper evaluates three such synthetic data sources: The WSARE dataset, the Noufilay equation-based approach, and the Project Mimic data generator.

Acknowledgment

This work was supported by Laboratory Directed Research and Development funding from Sandia National Laboratories.

Contents

Figures

Tables

Motivation

The research and development of new algorithmic and statistical methods of outbreak detection is an ongoing research priority in the field of biosurveillance [16, 17, 3]. The early detection of emergent disease outbreaks is crucial for effective treatment and mitigation. Unfortunately, outbreak warning signs can be obfuscated by noisy and missing data and can be overshadowed by higher baseline counts of other systemic diseases. Modern approaches to biosurveillance monitor data streams that may report a variety of information including daily hospital and clinic visits [12, 9], prescription and over-the-counter medicine sales [4], and natural language processing of written hospital reports [1]. The analysis of these data must balance the trade-off between high sensitivity and specificity. A high rate of outbreak detection will likely lead to a similarly high rate of false positives that may prove to be both time consuming and costly. Conversely, a low number of false positives will lead to a lower outbreak detection rate. A new detection method will strive to have a higher rate of detection while maintaining a similar or lower false positive rate when compared to previous approaches.

To properly evaluate a new detection method, it must be compared to established approaches. This comparison requires biosurveillance test data that accurately reflects the complexity of the real-world data it will be applied to. While the test and evaluation of a new detection method is best performed on real data, it is often impractical to obtain such data as it is either proprietary or limited in scope. Thus, scientists must turn to synthetic data generation methods to provide enough data to properly evaluate their new detection methodologies.

Limitations

Recent implementations of outbreak detectors are often designed and tested against a limited set of real-world data [2, 15, 10]. While ideal in principle, the use of real-world data has several limitations in practice.

- Real data is often difficult and expensive to obtain. Sources of data may not want to release potential personally identifying or HIPAA protected information to public researchers. Other sources may require significant financial compensation for their records, an unreasonable request for many researchers funded by preallocated research grants.
- When obtained, read-world data are often limited in size and scope. Small data sets that contain outlier events may unintentionally bias detectors and future studies. Small data sets that do not contain improbable yet existent features may not allow for the design of detectors with appropriate levels of specificity.
- Most real data sets are proprietary and thus may not be directly included in publications. Studies without available data sets are not reproducible and thus suspect by nature.
- Organizations that do release records will be constrained to specific sources of data. Therefore, multivariate data sets are usually an unrealistic objective when attempting to procure real data.

Until real data sets that are affordable, public, and comprehensive can be made available, researchers will need to make use of synthetic data generators in their studies.

Challenges

The generation of synthetic time series data dates back to at least the 1960s [8], yet most data generators are designed for use in climate studies [13, 6, 14]. While the techniques found in previous studies may prove useful, their direct application to the field of biosurveillance is not realistic. Generation of realistic biosurveillance time series introduces unique challenges not always present in the study of physical systems. These challenges include the following:

- Real-world biosurveillance data is noisy and incomplete. Pervasive infections such as the common cold and seasonal influenzas contribute seemingly random infection counts that may dwarf the size of a local outbreak of a novel pathogen. Data collectors such as hospitals, clinics, and pharmacies may forget to file timely reports or may misdiagnose illnesses.
- Detectable symptoms and effects of an outbreak may lag behind the actual onset of the infection process by an unspecified amount of time.
- The dynamics of infection spread is an ongoing research topic in the field of epidemiology. Simple SIR models of infection spread [5] are useful, but do not capture the complex spatial interactions present in the modern world. Temporal outbreak distribution can be modeled by a basic probability distribution such as a negative binomial or log normal, but this approach may prove overly simplistic and unrealistic.
- Infection data taken from a dynamic subpopulation may be non-stationary in that infection counts mirror the increase or decrease of the population being monitored.
- Pervasive illnesses such as influenza have a known seasonal component. Generated data should be able to replicate seasonal oscillations in disease prevalence.
- Data obtained from clinical and commercial sources are affected by weekly behavioral patterns. Mimicked data from these sources must accurately reflect week-weekend effects such as decreased clinical visits during the weekend and increased visits on Mondays due to the accumulated backlog of the weekend.
- Similar to day-of-the-week effects, official and unofficial holidays must be accurately taken into account.
- Modern approaches to outbreak detection often rely on a multitude of independent data sources. Each individual source may have noise and irregular behavior as described in the previous points that are unique to its domain.
- The detection of spatial concentration and spread of disease is of particular interest to biosurveillance researchers. Accurately representing such data likely requires a complex model as human movement and interaction is not well represented by simple diffusion processes.

While it is not required that a data generator represent all of the above behavior, outbreak detectors that are trained and evaluated on data that contains only a subset of these factors will be

ID	Label	Description
1	XY	Spatial region of record (3×3 grid)
2	age	Patient age: child, working, or senior
3	gender	Patient gender
4	flu	Global influenza prevalence: none, low, high, or decline
5	day_of_week	Saturday, Sunday, or a weekday
6	weather	Hot or cold
7	season	Winter, Spring, Summer, or Fall
8	action	Record type: purchase, evisit, or absent
9	reported_symptom	None, respiratory, nausea, or rash
10	drug	Drugs administered: none, nyquil, apririn, or vomit-b-gone
11	date	From Jan-01-2002 to Dec-31-2003
12	daynum	Date converted to a single integer index

Table 1: **Fields included in a single WSARE record.** 100 unique data sets exist. A single data set spans two years and contains multiple records per day. Each data set contains one simulated anthrax outbreak.

inherently limited in scope. Development of comprehensive modern detection algorithms demands data that reflects the full complexity of the real world.

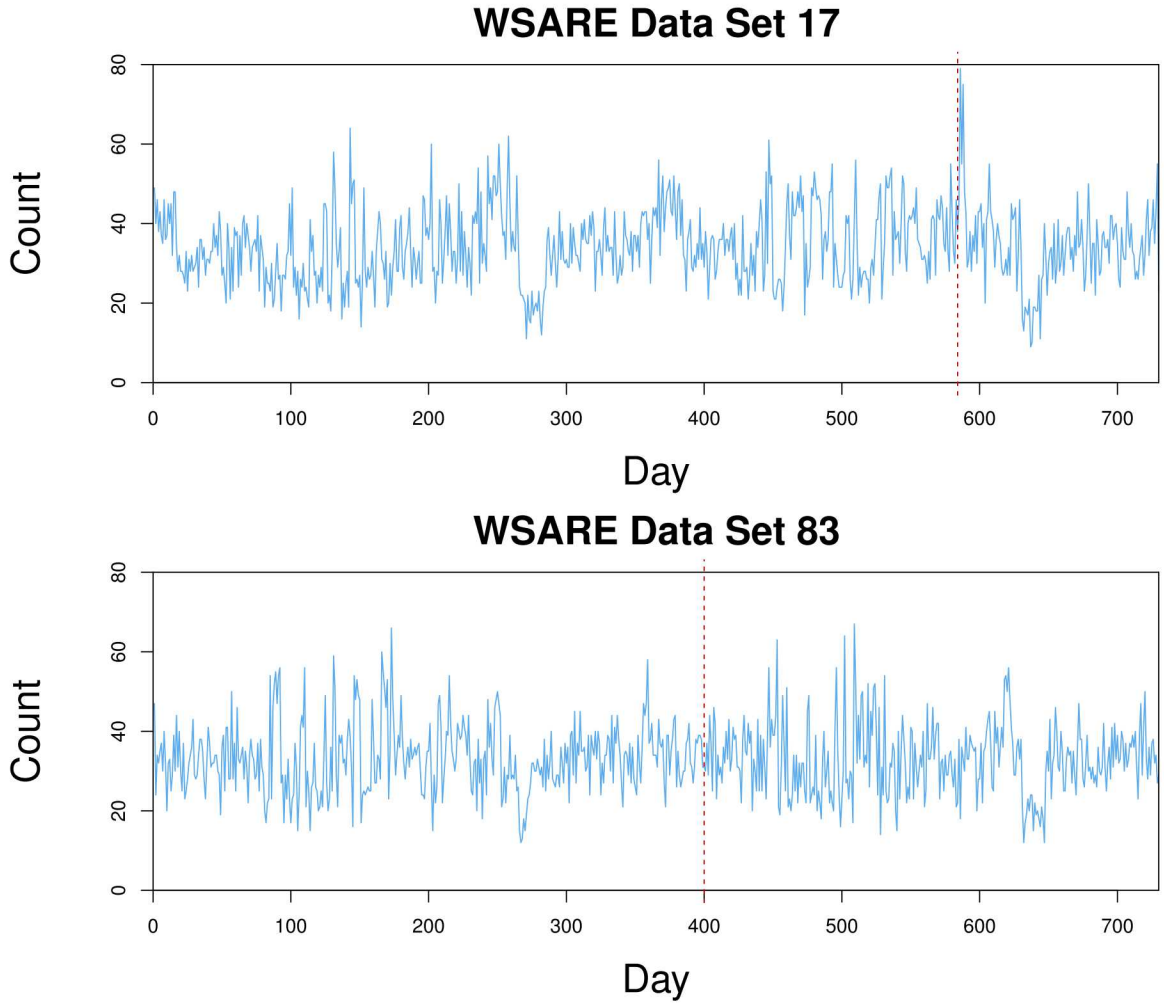

Figure 1: **Selected WSARE Timeseries.** Two randomly selected time series from the WSARE data sets. The blue line represents aggregated daily incident reports while the dashed red line marks the occurence of an anthrax outbreak. Note that the outbreak in Series 17 represents the global maximum, while the smaller outbreak in Series 83 is not directly visible.

Implementations

Despite the need for methods to generate realistic biosurveillance data, there are few established approaches in literature. Here, we review three sources of synthetic data: the What's Strange About Recent Events (WSARE) static database maintained by Weng-Keen Wong [18], a data generation method proposed by Noufaily et. al [11], and a second data generation method hosted by Project Mimic [7].

WSARE

The WSARE data set is maintained by Weng-Keen Wong and is hosted by the Auton Lab at `https://www.autonlab.org/datasets`. The data set consists of 100 generated samples, each with a simulated anthrax outbreak included in the time series, and was used to compare the WSARE 2.0, WSARE 2.5, and WSARE 3.0 detectors to a baseline algorithm [18].

The data contains lists of health care events as described in Table 1. An event is defined as either a clinical visit, a purchase of medication, or an irregular absence from school or work. Each data set contains approximately 25,000 individual records of patient visits, a subset of which are caused by a simulated anthrax infection event. Records contain a coarse spatial component consisting of a single location on a 3×3 grid along with general symptom information. Because each record is an independent event, daily counts must by obtained by aggregating individual records (Fig. 1).

The WSARE data series are public and static, allowing researchers to compare outbreak detectors on the same data. The data contains multivariate sources including three types of records, seasonal data, and a variety of reported symptoms. Each time series contains a single 1-day labeled outbreak event of various sizes with which to test detection algorithms. While the data do not contain 'missed' records, the data does reflect the random and noisy nature of real biosurveillance data series.

The main drawback of the WSARE data is the lack of a generation mechanism. Without a generator, data sets cannot be constructed that fit the exact needs of a specific scientific study. While 100 unique data sets are valuable, the two year window is limiting in its duration, and the data does not include non-stationary components. Further, the generated anthrax outbreaks appear to be single day events that do not reflect a spread of infection through a population over the course of several days. Finally, because the generation method is unknown, it is difficult to evaluate the accuracy of the data in isolation.

Noufaily et. al.

Noufaily et. al. [11] describe an equation-based approach to generating synthetic data. Mean infection counts (μ) are defined by an exponential function with baseline (θ), growth (β), and seasonality components (γ_1 and γ_2). The number of seasonal cycles per year is defined by m. Variance is controlled by the parameter ϕ and a single weekly infection count $C(t)$ is chosen at each time point t from a negative binomial distribution with mean $\mu(t)$ and variance $\phi\mu(t)$. The system is described by Equations 1 and 2:

$$\mu(t) = exp\left[\theta + \beta t + \sum_{j=1}^{m}\left(\gamma_1 cos\left(\frac{2\pi jt}{52}\right) + \gamma_2 sin\left(\frac{2\pi jt}{52}\right)\right)\right] \tag{1}$$

$$C(t) \sim NB(\mu(t), \phi\mu(t)) \tag{2}$$

Parameter	Description	Min	Max
θ	Baseline	-2	5
β	Growth Rate	0	0.005
γ_1	1^{st} Seasonality	0	1
γ_2	2^{nd} Seasonality	-0.4	0.6
ϕ	Variance	1	5
m	No. of Seasons	0	2

Table 2: **Parameter ranges used in parameter selection by Noufaily et. al.** The minimum and maximum values used by Noufilay et. al. [11] when generating a time series. Data taken from values included in the 42 sample scenario set.

It is important to note that Noufaily et. al. and Eq. 2 describe the negative binomial random variable $C(t)$ in terms of the mean and variance, whereas a negative binomial distribution is generally defined in terms of a single trial probability and number of failures.

Once the baseline series is generated, outbreaks may be retroactively added to the time series through simple addition. Outbreak size is controlled using the parameter k, typically in the range of 2 through 10. An outbreak at time t, $O(t)$, has its size chosen as a Poisson distributed random variable with mean proportional to k times the standard deviation of the baseline count:

$$O(t) \sim Pois(k\sigma_{C(t)}) \tag{3}$$

Once an outbreak size is selected, each individual infection of the outbreak is added to the count $C(t+i)$ where i is a random variable chosen from a log normal distribution with mean 0 and standard deviation of 0.5:

$$i \sim LN(0, 0.5) \tag{4}$$

This results in outbreaks that generally last from 2 to 8 weeks, depending on the choice of k

We have implemented these equations in the form of a function written in the R programming language, presented in Listing 1 at the end of this document. Given this function, Noufaily et. al. describe 42 parameter combinations for the parameter set (θ, β, γ_1, γ_2, *phi*, m) that they claim to represent realistic infection scenarios. The parameter sets contain a mix of stationary and non-stationary time series, as well as a mix of seasonal and non-seasonal time series. Since outbreak generation is independent of the base series generation, outbreaks can be added dynamically at any point to any generated time series, which is useful for testing purposes.

While an improvement of the WSARE static data set, this approach also has several drawbacks. First, the equation only generates a single time series rather than more useful multivariate data. Second, the resolution of the time series is defined as weekly and oscillating components represent seasonal trends but not weekly ones. Finally, the parameters that define the generator (Table 2) are not grounded in real-world units and are therefore difficult to interpret in isolation. Further, it is

ideal that synthetic data mirror trends seen in real data, yet we must take the authors' word that the predefined 42 parameter sets properly reflect the dynamics of real infection counts.

Project Mimic

The Project Mimic data, data generator, and documentation can be found at the project's website, `projectmimic.com`. Project Mimic's defines a random multivariate time series in terms of vectors containing the series' means, variances, and 1-step autocorrelations, as well as a 2-dimensional covariance matrix. Once a baseline series is created from these parameters, it can be modified to reflect weekly and seasonal trends, as well as effects from holidays. Finally, emerging outbreaks can be randomly generated and added into the baseline series. An example time series with three components and one outbreak created by the Project Mimic generator can be seen in Figure 3.

Project Mimic's generator is fully coded as a complete R package that can be downloaded from the website and installed locally (it is not in the CRAN remote package library). Because the generator was written for R versions 2.7.0, it must be updated for use with R versions 3.0 and later. Specifically, in the `extractSeriesCharacteristics.R` file, the use of the `mean()` function has been deprecated and must be replaced with the newer `colMeans()` function.

The Project Mimic generator comes with documentation describing its design as well as usage instructions for the R package [7]. In general the usage instructions to generate baseline outbreak times series are accurate, but minor corrections need to be made. The example date used in the instructions, `"Feb-01-2000"`, must be changed to an ISO-8601 standard format: `"2000-02-01"`. The instructions also contain minor typos on page 18: weekly trend vectors each have an erroneous comma that must be removed. Baseline multivariate series can be generated using the `generateBaseSeries()` function as described in the instructions.

The generator can also simulate emerging outbreaks, but the documentation describing the process is incorrect. The user must specify the following elements:

1. `outbreakType` - `'logNormal'` or `'spike'` depending on the desired outbreak time distribution.
2. `numcases` - A vector containing the number of outbreak events for each of the separate data sources.
3. `numdays` - The number of days contained in the original baseline time series.
4. `outbreakStartDay` - A vector containing the outbreak start days. Multiple values correspond to multiple independent outbreaks.
5. `outbreakInSeries` - A matrix with a row for each independent outbreak. Each row is the same length as the `numcases` vector, and contains either a 1 (true) or a 0 (false) indicating whether the specific data source is affected by the outbreak corresponding to that matrix row.

6. `mean` - For log normal outbreaks only: the average location in time past the initial infection of a single outbreak event.
7. `sd` - For log normal outbreaks only: the standard deviation in time past the initial infection of a single outbreak event.
8. `trimPercent` - For log normal outbreaks only: the percent of days to remove from the end of a log normal outbreak to prevent unrealistic long-tail effects.

These arguments can be supplied to the function `generateOutbreakWithLabels()` in the order specified to create a new random multivariate outbreak that can be combined with a baseline time series.

The software package also has the ability to mimic established time series. This can be useful when the static time series is either too short to be useful or if the data contains personally identifying information that needs to be removed. Project Mimic can examine an existing time series and extract the parameters needed to generate a similar but random time series. Use the `extractSeriesCharacteristics()` function as described in the instructions to generate a proper parameter set.

The Project Mimic package is powerful and relatively easy to use. Of the three options presented here, it is the most powerful as it is the only one that can generate weekly trends and infer parameter sets from static data series. The main drawback of the Project Mimic generator is its inability to natively generate non-stationary data, but that capability would be an easy manual addition.

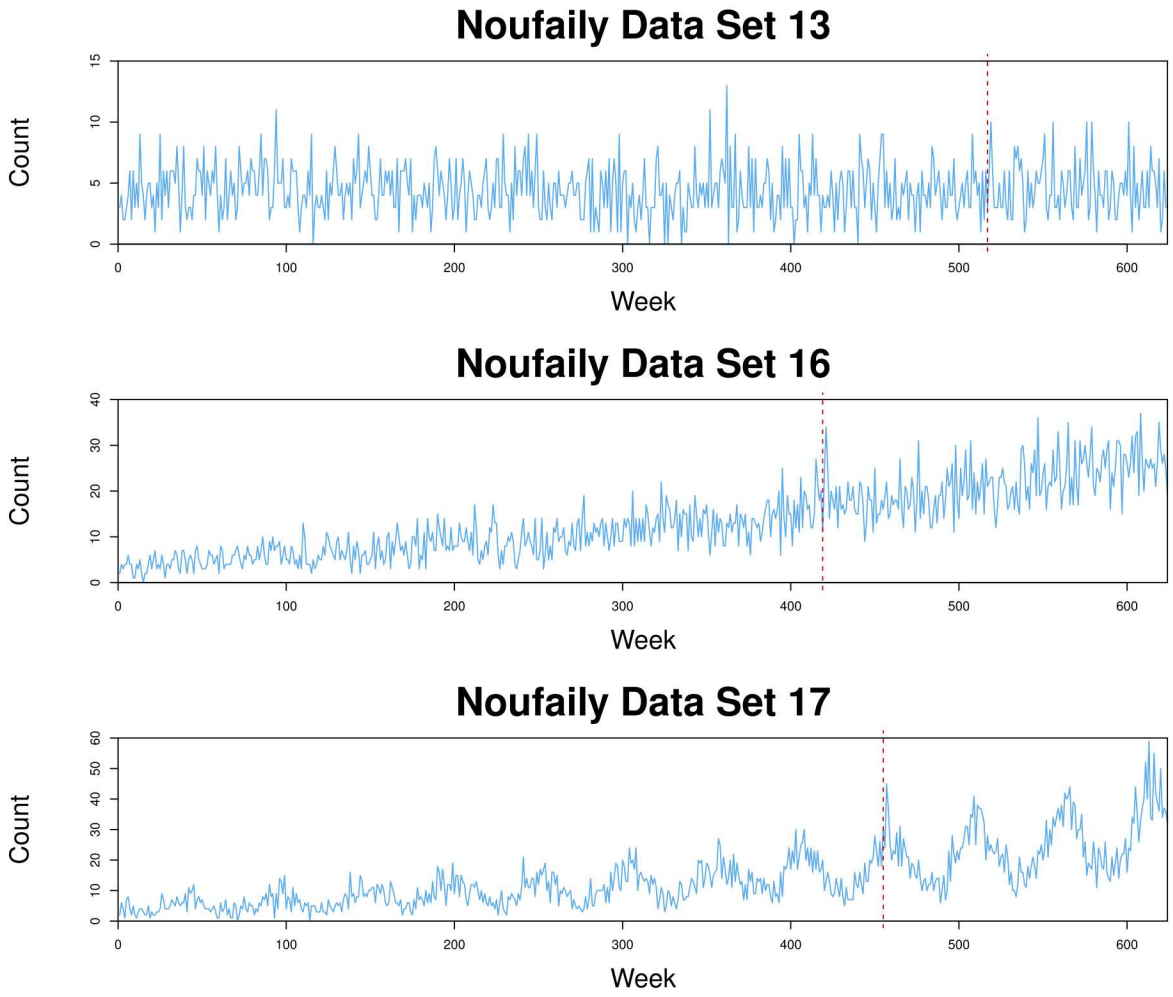

Figure 2: **Selected Noufaily et. al. Generated Timeseries.** Three randomly generated time series from the Noufaily et. al [11] generation functions (Eqs. 1-4). The blue line represents weekly infection counts while the dashed red line marks the occurrence of an outbreak. Data Set 13 generates a stationary distribution, Data Set 16 generates a non-stationary distribution, and Data Set 17 generates a non-stationary distribution with a seasonal component.

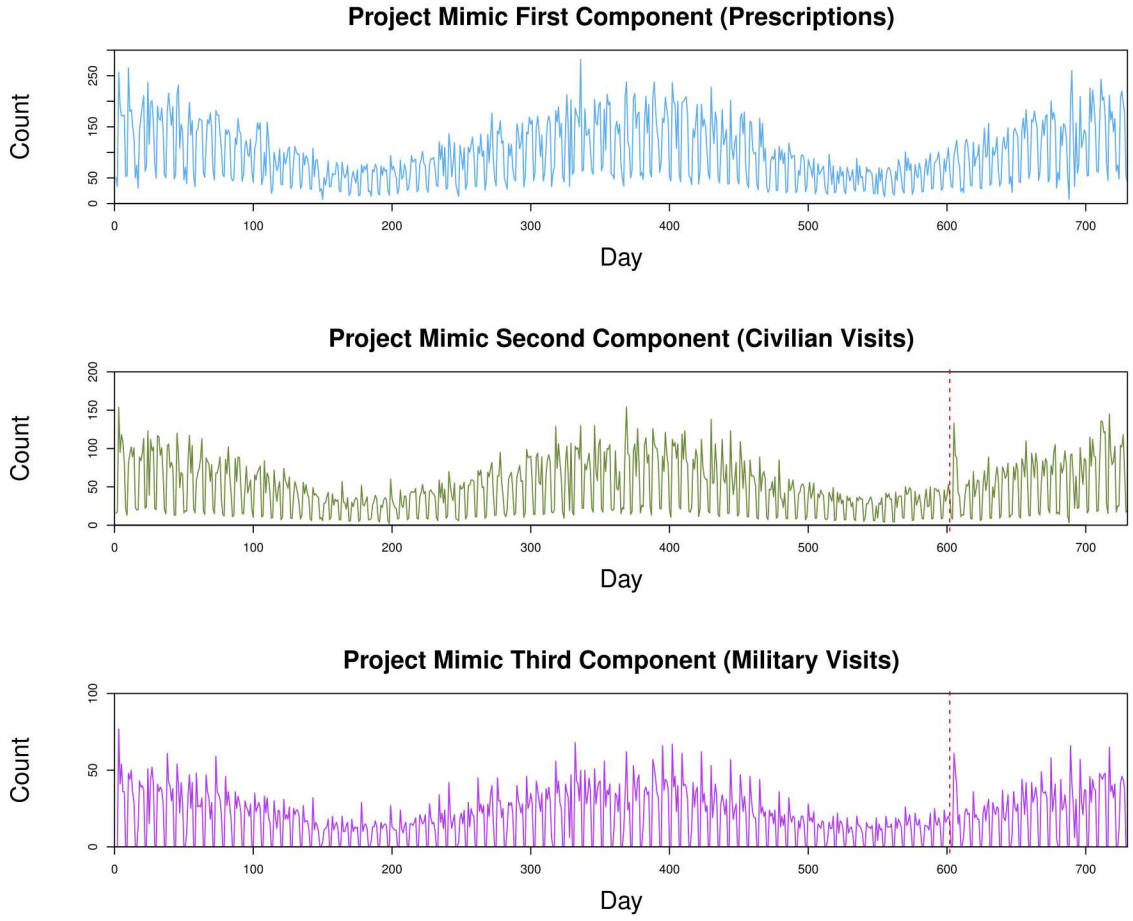

Figure 3: **Project Mimic Generated Trivariate Timeseries.** Three correlated time series generated using Project Mimic. Each component represents a separate data source (prescription sales, civilian clinic visits, and military clinic visits). The series exhibit both weekly and seasonal effects. A single outbreak even occurs at day 600. Note that the outbreak event is customized to only affect the clinic visit counts and not prescription sales.

Future Work

The three data sources presented in this paper provide a diverse set of synthetic biosurveillance data. Specifically, the Project Mimic generator presents a state-of-the-art data generator that can produce correlated multivariate time series that mimic static real-world data sets.

A key component missing from each of the sources is the ability to represent spatial information. Of the three, only the WSARE data sets contain any spatial component, and they are limited to a simple 3-by-3 grid. Real biosurveillance data is spatially distributed by nature, and the geographic spread of disease is a key component of any infection outbreak. As future studies will aim to make use of the multitude of spatially diverse biosurveillance data sources to better detect and localize outbreaks, synthetic data sources will need to be able to represent realistic spatial effects. Because most diseases spread through a form of human contact, disease propagation must be modeled with human activity patterns in mind. To do so will likely require an explicit agent-based model of human movement over a simulated map. The drawback of such an approach will be the increased computational complexity of such a simulation.

Conclusions

This paper presents three sources of synthetic biosurveillance data. Each source has unique advantages not seen in the other two. The WSARE data set [18] contains multivariate data in the forms of individual event records (not just daily counts). The data also contains demographic and medical information specific to each individual event for retrospective analysis. Further, the WSARE data set is the only one of the three to contain a spatial component. The data generation method presented by Noufaily et. al [11] is elegant in its simplicity as it is able to be implemented in a limited amount of code (Listing 1). The Noufaily et. al. generator is also the only one of the three that can represent non-stationary data, an important feature present in most real-world time series. Finally, the Project Mimic generator has the unique ability to dynamically generate data with both weekly and seasonal components, and can also generate new time series that mimic the statistical properties of static real-world time series.

Of the three data sources reviewed in this paper, the Project Mimic time series generator stands out as the most powerful and accurate source of synthetic data. Future work may look to extend established data generation techniques to allow for spatially explicit components, although a spatial data generator may introduce an infeasible level of complexity to a basic data generator.

Listing 1: Sample R code to generate a synthetic data series using the approach of Noufilay et. al.

```
# genDiseaseSeriesSingle
#   Generates a single synthetic infection data series based on algorithm in
#   Noufaily 2012. Series contains one outbreak.
# Inputs:
#   theta     - Baseline frequency
#   beta      - General linear trend
#   gamma1    - Seasonality trend 1
#   gamma2    - Seasonality trend 2
#   phi       - Variance
#   m         - Number of seasons (0 - none, 1 - annual, 2 - biannual)
#   k         - Size of outbreak relative to series standard deviation
#   days      - Number of time points to generate
#   curstart - Start of current time period, outbreak will occur after this
    day
# Output:
#   A list containing the observed time series ($observed) and the outbreak
    day ($day)
genDiseaseSeriesSingle <-
  function(theta = 0.1, beta = 0, gamma1 = 0, gamma2 = 0, phi = 1.5, m = 0,
           k = 5, days = 624, curstart = 400) {

  # Generate a single time series
  time = 1:days
  mu =  theta + beta*time
  season1 = ifelse(m > 0, 1, 0) * (gamma1*cos(2*pi*time/52) + gamma2*sin(2*pi*
     time/52))
  season2 = ifelse(m > 1, 1, 0) * (gamma1*cos(4*pi*time/52) + gamma2*sin(4*pi*
     time/52))

  # Combine individual components and add random variance
  # Use R's mu|size NBinom parameterization
  observed = exp(mu + season1 + season2)
  observed = rnbinom(days, mu = observed, size = observed/(phi-1))

  # Pick a day in the outbreak period uniformly at random
  t = floor(runif(1, curstart, days))

  # Pick the size of the outbreak using k and generate a set of events
  cases = table(round(rlnorm(rpois(1, k*sqrt(observed[t]*phi)), 0, 0.5)))
  for (c in names(cases)) {
    observed[t+min(strtoi(c), days)] =
      observed[min(t+strtoi(c), days)] + cases[[c]][1]
  }

  return(list(observed=observed, day=t))
}
```

References

[1] Wendy W Chapman, John N Dowling, and Michael M Wagner. Fever detection from free-text clinical records for biosurveillance. *Journal of biomedical informatics*, 37(2):120–7, apr 2004.

[2] K. E. Cheng, D. J. Crary, J. Ray, and C. Safta. Structural models used in real-time biosurveillance outbreak detection and outbreak curve isolation from noisy background morbidity levels. *Journal of the American Medical Informatics Association*, pages 435–440, 2012.

[3] Kimberly N. Gajewski, Amy E. Peterson, Rohit A. Chitale, Julie A. Pavlin, Kevin L. Russell, and Jean Paul Chretien. A review of evaluations of electronic event-based biosurveillance systems. *PLoS ONE*, 9(10):7–10, 2014.

[4] Anna Goldenberg, Galit Shmueli, Richard A. Caruana, and Stephen E. Fienberg. Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales on JSTOR, 2002.

[5] W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, aug 1927.

[6] Upmanu Lall and Ashish Sharma. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resources Research*, 32(3):679–693, mar 1996.

[7] Thomas Lotze, Galit Shmueli, and Inbal Yahav. Simulating Multivariate Syndromic Time Series and Outbreak Signatures. *SSRN Electronic Journal*, may 2007.

[8] N. C. Matalas. Mathematical assessment of synthetic hydrology. *Water Resources Research*, 3(4):937–945, dec 1967.

[9] Zaruhi R. Mnatsakanyan, Howard S. Burkom, Mohammad R. Hashemian, and Michael A. Coletta. Distributed information fusion models for regional public health surveillance. *Information Fusion*, 13(2):129–136, apr 2012.

[10] Mohamad Farhan Mohamad Mohsin. The Preliminary Design of Outbreak Detection Model Based on Inspired Immune System. *Third World Congress on Information and Communication Technologies*, pages 190–196, 2013.

[11] Angela Noufaily, Doyo G Enki, Paddy Farrington, Paul Garthwaite, Nick Andrews, and André Charlett. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in medicine*, 32(7):1206–22, mar 2013.

[12] Ben Y Reis, Marcello Pagano, and Kenneth D Mandl. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1961–5, feb 2003.

[13] C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190, feb 1981.

[14] A Shamshad, M Bawadi, W Wanhussin, T Majid, and S Sanusi. First and second order Markov chain models for synthetic generation of wind speed time series. *Energy*, 30(5):693–708, apr 2005.

[15] Galit Shmueli. Wavelet-Based Monitoring for Biosurveillance. *Axioms*, 2(3):345–370, 2013.

[16] Galit Shmueli and Stephen E Fienberg. Current and potential statistical methods for monitoring multiple data streams for biosurveillance. In *Statistical Methods in Counterterrorism*, pages 109–140. Springer, 2006.

[17] Steffen Unkel, Paddy Farrington, Paul Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of Royal Statistical Society: Series A*, 175(1):49–82., 2012.

[18] Weng-keen Wong, Andrew Moore, Gregory Cooper, and Michael M Wagner. Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning*, pages 808–815, Menlo Park, California, 2003. AAAI Press.

DISTRIBUTION:

1	MS 0359	D. Chavez, LDRD Office, 1911
1	MS 0899	Technical Library, 9536 (electronic copy)


Sandia National Laboratories