Risk Management, Signal Processing and Econometrics: A New Tool for Forecasting the Risk of Disease Outbreaks

A[a,*], B[b,*], C[c,*], D[d,*]

[a] *A address*
[b] *B address*
[c] *C address*
[d] *D address*

Abstract

This paper takes a novel approach for forecasting the risk of disease emergence by combining risk management, signal processing and econometrics to develop a new forecasting approach. We propose quantifying risk using the Value at Risk criterion and then propose a two staged model based on Multivariate Singular Spectrum Analysis and Quantile Regression (MSSA-QR model). The proposed risk measure (PLVaR) and forecasting model (MASS-QR) is used to forecast the worst cases of waterborne disease outbreaks in 22 European and North American countries based on socio-economic and environmental indicators. The results show that the proposed method perfectly forecasts the worst case scenario for less common waterborne diseases whilst the forecasting of more common diseases requires more socio-economic and environmental indicators.

Keywords: Value at Risk; Disease; Outbreaks; Forecasting; Quantile Regression; Multivariate Singular Spectrum Analysis.

1. Introduction

The accurate forecasting of disease outbreaks continue to challenge researchers, governments and policy makers (Graham et al. , 2018; Metcalf and Lessler , 2018). The task itself is challenging as an outbreak is a result

[*] Corresponding author
Email addresses: `A's Email` (A), `B's Email` (B), `C's Email` (C), `D's Email` (D)

of interactions between pathogens/parasites, hosts and other environmental variables (Alizon et al. , 2013; Griffiths et al. , 2011).

Accordingly, in the recent past, researchers have adopted a variety of tools from different parts of science to forecast disease outbreaks. For instance, Lowe et al. (2017) used precipitation, minimum temperature, and El Niño index forecasts to predict the dengue incidence in Ecuador. Their results show that using climatological forecasts could improve the accuracy of dengue outbreak forecast. Han and Drake (2016) proposed using statistical machine learning methods to forecast the outbreaks of a disease. They argued that applying machine learning methods to existing big data on environmental, epidemiological and molecular systems could help public health authorities to predict the flow or risks of disease emergence (including outbreak risks). Liao et al. (2017) used a Bayesian Belief Network (BBN) to predict the risk of further outbreaks. They suggest that the BBN technique can be used for early warnings of infectious diseases.

Although many of the methods considered in disease outbreak risk forecasting proved to be accurate and effective, most of the research forecasts the number of cases/incidence, ratios or the probability of occurrence as outbreak risks. On the other hand, in risk management, one is usually interested in worst case scenarios. For instance, in financial risk analysis, instead of forecasting the average value of an asset, it is common to forecast the value which is the lowest with 95% confidence. Such values are referred to as Value at Risk (Davino et al., 2014) and shows the value of the asset in in extremely negative conditions (the probability of extreme events taking place is 5%).

In this paper, we are concerned with forecasting the worst case scenarios for disease outbreaks. Relying on financial risk analysis, a new risk measure is proposed to present the worst case scenario. More specifically, a model based on the Multivariate Singular Spectrum Analysis (Sanei and Hassani, 2015) and the Quantile Regression (Koenker , 2005) is developed to forecast the disease outbreak worst case scenario. The proposed method is used to forecast annual outbreaks of 13 waterborne disease in 22 European and North American countries between 2011 and 2015. The data from 10 socio-economic and environmental indicators between 1998 and 2010 is used to estimate the coefficients of the model (train the model). Results show that with relatively small number of indicators and training data, the proposed model has the ability to forecast the worst cases of outbreaks for less common waterborne diseases. For more common waterborne disease like Diarrhoea, Pertussis and Malaria, however, more indicators are needed.

The remainder of the paper is organised as follows. The proposed forecasting method is presented in Section 2. Section 3 gives a complete description of the waterborne disease dataset and indicators used to forecast the disease outbreaks. The results from the forecasting exercise for waterborne disease outbreaks are presented in Section 3. Finally, Section 4 concludes the paper.

2. Methodology

2.1. Value at Risk and Population Loss Value at Risk

The Value at Risk (VaR) (Leavens , 1945) is one of the common risk measures in financial risk analysis. The VaR measure shows the minimum value of an asset (or its return) with $1-\alpha$ confidence level, i.e. the probability that the value of an asset goes under the VaR is α. In other words, the VaR shows the scenario which with confidence level $1-\alpha$ worst that that won't happen (the risk that cases worst than VaR happens in reality is α). Since in investment problems, the worst cases are always the lower values (e.g. lower returns, price, or income) the VaR in risk level α (confidence level $1-\alpha$) is defined as follows:

$$VaR_\alpha(Y) = \inf\{y \in \mathbb{R} : F_Y(y) = \alpha\}$$

where Y is the value (return, price, ...) of the financial asset. The VaR_α is the αth quantile of the value distribution ($F_Y(y)$), It shows the value of an asset in risk situations which means with $1-\alpha$ confidence the VaR_α is the worst case scenario (for more details on VaR see McNeil et al., 2005).

Adopting the VaR concept from finance, we define the Population Loss Value at Risk (PLVaR), as the worst case scenario in disease outbreak with risk level α:

$$PLVaR_\alpha(Y) = \inf\{y \in \mathbb{R} : F_Y(y) = 1-\alpha\}, \tag{1}$$

where Y is the number (or ratio) of losses in disease outbreak. Unlike $VaR_\alpha(Y)$, the $PLVaR_\alpha(Y)$ is the $(1-\alpha)$th quantile of the Y, since the worst case in disease outbreak is the case with largest number (ratio) of losses. In this manner, the $PLVaR_\alpha$ shows the worst case scenario in disease outbreak, with $1-\alpha$ confidence level.

The $PLVaR$ can be used as a risk measure in disease control and outbreak prevention planes. The $PLVaR$ has the ability to forecast the disease outbreaks along with the size of the break out. Non-zero Values of $PLVaR$ show the outbreak situations, while the larger values show the estimate the larger outbreaks. For instance, the $PLVaR_{0.01} = 0$ means in 0.99 confidence level, there is not a disease outbreak (in other words, it means the chance of disease outbreak is under 1%). Using $PLVaR$ as a risk measure, one may forecast the future values of $PLVaR_\alpha$ in order to forecast the size of the future outbreaks.

2.2. Multivariate Singular Spectrum Analysis

The Horizontal MSSA Recurrent (HMSSA-R) forecasting algorithm uses following steps to forecast multivariate time series. Those interested in an in-depth explanation of the theory underlying MSSA are directed to Sanei and Hassani (2015). In presenting this algorithm we mainly follow and rely on the notations in Sanei and Hassani (2015).

2.2.1. HMSSA-R Optimal Forecasting Algorithm

1. Consider M time series with identical series lengths of N_i, such that $Y_{N_i}^{(i)} = (y_1^{(i)}, \dots, y_{N_i}^{(i)})$ $(i = 1, \dots, M)$.
2. For forecasting exercises we would split each time series into three parts leaving $\frac{2}{3}^{rd}$ for model training and testing, and $\frac{1}{3}^{rd}$ for validation.
3. Beginning with a fixed value of $L = 2$ $(2 \leq L \leq \frac{N}{2})$ and in the process, evaluating all possible values of L for Y_{N_i}, using the training data construct the trajectory matrix $\mathbf{X}^{(i)} = [X_1^{(i)}, \dots, X_K^{(i)}] = (x_{mn})_{m,n=1}^{L,K_i}$ for each single series $Y_{N_i}^{(i)}$ $(i = 1, \dots, M)$ separately.
4. Then, construct the block trajectory matrix $\mathbf{X}_H$ as follows:

$$\mathbf{X}_H = \left[\ \mathbf{X}^{(1)} : \ \ \mathbf{X}^{(2)} : \ \ \cdots \ \ : \mathbf{X}^{(M)}\ \right].$$

5. Let vector $U_{H_j} = (u_{1j}, \dots, u_{Lj})^T$, with length L, be the j^{th} eigenvector of $\mathbf{X}_H\mathbf{X}_H^T$ which represents the SVD.

6. Evaluate all possible combinations of r $(1 \leq r \leq L-1)$ step by step for the selected L and construct $\widehat{\mathbf{X}}_H = \sum_{i=1}^{r} U_{H_i} U_{H_i}^T \mathbf{X}_H$ as the reconstructed matrix obtained using r eigentriples:

$$\mathbf{X}_H = \left[\ \widehat{\mathbf{X}}^{(1)} :\ \ \widehat{\mathbf{X}}^{(2)} :\ \ \cdots\ \ : \widehat{\mathbf{X}}^{(M)}\ \right].$$

7. Consider matrix $\widetilde{\mathbf{X}}^{(i)} = \mathcal{H}\widehat{\mathbf{X}}^{(i)}$ $(i = 1, \ldots, M)$ as the result of the Hankelization procedure of the matrix $\widehat{\mathbf{X}}^{(i)}$ obtained from the previous step for each possible combination of SSA choices.

8. Let $U_{H_j}^{\triangledown}$ denote the vector of the first $L-1$ coordinates of the eigenvectors U_{H_j}, and π_{H_j} indicate the last coordinate of the eigenvectors U_{H_j} $(j = 1, \ldots, r)$.

9. Define $\upsilon^2 = \sum\limits_{j=1}^{r} \pi_{H_j}^2$.

10. Denote the linear coefficients vector $\mathcal{R}$ as follows:

$$\mathcal{R} = \frac{1}{1-\upsilon^2} \sum_{j=1}^{r} \pi_{Hj} U_{Hj}^{\triangledown}. \tag{2}$$

11. If $\upsilon^2 < 1$, then the h-step ahead HMSSA forecasts exist and is calculated by the following formula:

$$\left[\hat{y}_{j_1}^{(1)}, \ldots, \hat{y}_{j_M}^{(M)}\right]^T = \begin{cases} \left[\tilde{y}_{j_1}^{(1)}, \ldots, \tilde{y}_{j_M}^{(M)}\right], & j_i = 1, \ldots, N_i, \\ \\ \mathcal{R}^T \mathbf{Z}_h, & j_i = N_i + 1, \ldots, N_i + h, \end{cases} \tag{3}$$

where, $\mathbf{Z}_h = \left[Z_h^{(1)}, \ldots, Z_h^{(M)}\right]^T$ and $Z_h^{(i)} = \left[\hat{y}_{N_i-L+h+1}^{(i)}, \ldots, \hat{y}_{N_i+h-1}^{(i)}\right]$ $(i = 1, \ldots, M)$.

12. Seek the combination of L and r which minimises a loss function, $\mathcal{L}$ and thus represents the optimal HMSSA-R choices for decomposing and reconstructing in a multivariate framework.

13. Finally use the selected optimal L to decompose the series comprising of the validation set, and then select r singular values for reconstructing the less noisy time series, and use this newly reconstructed series for forecasting the remaining $\frac{1}{3}^{rd}$ observations (or the test set as relevant to this study).

2.3. Quantile Regression

The Quantile Regression (QR) models the τth quantile of the response variable using a regression line:

$$Q_\tau = \beta_{0,\tau} + \sum_{i=1}^{p} \beta_{i,\tau} x_i + \varepsilon_\tau,$$

where $x_1 \ldots, x_p$ are independent variables and Q_τ is the τth quantile of response variable y with cumulative distribution function $F_Y(.)$:

$$Q_\tau = \inf\{y \in \mathbb{R} : F_Y(y) = \tau\}, \quad 0 < \tau < 1.$$

The coefficients of the model can be estimated by minimizing the loss function $L_\tau(e) = \left(\tau - I_{(e<0)}\right) e$ where $I_{(e<0)}$ is the Indicator function (for more details on QR see Davino et al., 2014):

$$I_{(e<0)} = \begin{cases} 1 & \text{if} e < 0 \\ 0 & \text{otherwise} \end{cases}$$

The QR model is a simple tool for risk analysis. For instance, one may use the QR model to estimate the VaR (or PLVaR) for response variable y based on given situation (indicators) $x_1, \ldots, x_p$. On the other hand, one may use the QR model to control the worst case scenario using the control variables $x_1, \ldots, x_p$.

2.4. MSSA-QR model for PLVaR forecasting

In order to forecast the PLVaR, we propose a two stage model. At the first stage, we use MSSA to forecast the indicators in the model. The second stage, uses forecasted values of indicators, to estimate the outbreak risk. It should be noted that in first stage, not all the variables need to be forecasted using MSSA. The future values of some indicators are already forecasted (for instance the population structure and population growth rates for different countries are forecasted using Birth/Death models and are available from `http://www.un.org/en/development/desa/population/`). Furthermore, some of the indicators are related to governments policies and can be forecasted based on governments announced policies. The MSSA-QR model for PLVaR h step ahead forecasting follows these steps:

First Stage: Forecasting the indicators

1. Use data available from the past ($t = 1, \dots, N$) for M countries/regions and the birth/death models to calculate h step ahead forecast for population indicators (e.g. population structure, growth etc.).
2. Assess the government's announced policies and use data available from the past ($t = 1, \dots, N$) to forecast the indicators related to government's policies (like infrastructural developments) for the desired time horizon.
3. Use the HMSSA-R algorithm and calculate the h step ahead forecasts for the rest of the indicators, based on historical data (each indicator is a M-variate time series where M is the number of countries/regions).

Second Stage: Forecasting the PLVaR for a given risk level α

1. Use the data available in time period $t = 1, \dots, N$ and countries/regions $i = 1, \dots, M$ to fit the QR model as:

$$PLVaR_{\alpha}(Y_{t,i}) = Q_{1-\alpha} = \beta_{0,1-\alpha} + \sum_{j=1}^{p} \beta_{j,1-\alpha} x_{j,t,i} + \varepsilon_{1-\alpha,t,i},$$

   where $Y_{t,i}$ is the number (or ratio) of deaths caused by disease outbreak at time t and country/region i. The $x_{j,t,i}$ is the jth indicator observed value at time t and country/region i. The $\varepsilon_{\alpha,t,i}$ is the innovation term with mean zero and constant variance σ^2_{α}.
2. Use the fitted QR model and forecasted values of indicators (from the First Stage) to forecast future PLVaRs:

$$\widehat{PLVaR}_{\alpha}(Y_{t+k,i}) = \widehat{\beta}_{0,1-\alpha} + \sum_{j=1}^{p} \widehat{\beta}_{j,1-\alpha} \widehat{x}_{j,t+k,i}, \quad k = 1, \dots, h$$

2.5. Model accuracy measures

Root mean squared error: The common accuracy measure in time series forecasting models, is the Root Mean Square Error (RMSE). For M-variate time series the RMSE is formulated as follows:

$$RMSE = \sqrt{\sum_{i=1}^{M} \sum_{t=1}^{N} (y_{t,i} - \widehat{y}_{t,i})^2},$$

where $\widehat{y}_{t,i}$ is the forecasted value of time series.

Exceedance rate: Suppose $\widehat{Q_\tau}$ is the estimated value of τth quantile based on observations $y_1, \ldots, y_N$. The exceedance rate of $\widehat{Q_\tau}$ is the relative frequency of the observations greater than $\widehat{Q_\tau}(Y)$. If the estimation of τ quantile is accurate, the exceedance rate should be close to $1 - \tau$. In risk assessment applications, the exceedance rate is used to evaluate the accuracy of estimated VaR. If the exceedance rate is less than $1-\tau$ the estimated VaR will present the worst case scenario accurately.

In this research, the exceedance rate is used to investigate the accuracy of QR in PLVaR forecasting (with risk level α).

$$ER_\alpha = \frac{1}{N}\sum_{i=1}^{M}\sum_{t=1}^{N} I_{(y_{t,i} > \widehat{PLVaR}_\alpha(Y_{t,i}))},$$

where $I_{(.)}$ is Indicator function. Exceedance rate lower than α means the risk of using $\widehat{PLVaR}_\alpha(Y_{t,i})$ as the worst case scenario is less than α.

3. Data Description and Results

In order to forecast the waterborne and disease outbreak risk, we use the input dataset, published by World Health Organization (WHO) and used to calculate the 2000-2016 Disease burden and mortality estimates. The dataset contains the annual number of deaths cussed by 13 waterborne diseases between 1998 and 2016, for 22 European and North American countries (WHO, 2018)[1]. The annual number of deaths per million, cussed by each disease, is a measure of disease outbreak for that disease.

Table 1 shows the list of waterborne disease considered in this study whilst Table 2 shows the list of countries involved. The $PLVaR_\alpha$ is considered as the $(1-\alpha)$th quantile of the annual number of deaths per million. The $PLVaR$ is forecasted using water related environmental and socio-economic indicators. The description of the indicators are as follows:

- **FSS:** This indicator is based on an assessment of the percentage of fish stocks caught within a countrys Exclusive Economic Zone (EEZ) that are overexploited or collapsed(Wendling et al., 2018; YCELP, 2018).

[1] The dataset is available from World Health Organization (`http://www.who.int/healthinfo/global_burden_disease/estimates/en/`). The original dataset contains 47 countries from Europe and North America. The countries with no records of water- or disease-related environmental indicators, in that period, are dropped from this study.

Table 1: Waterborne diseases in this study.

1	Chlamydia	8	Dengue
2	Diarrhoeal Diseases	9	Japanese Encephalitis
3	Pertussis	10	Trachoma
4	Poliomyelitis	11	Ascariasis
5	Malaria	12	Trichuriasis
6	Schistosomiasis	13	Hookworm Disease
7	Onchocerciasis		

Table 2: List of countries in this study.

1	Canada	9	Guatemala	17	Puerto Rico
2	Croatia	10	Iceland	18	Republic of Moldova
3	Denmark	11	Ireland	19	Sweden
4	Estonia	12	Italy	20	Switzerland
5	Finland	13	Latvia	21	United Kingdom
6	France	14	Netherlands	22	United States of America
7	Germany	15	Panama		
8	Greece	16	Poland		

- **FPRO:** Fisheries production (Total) (tonnes)[2](FAO, 2018)
- **FWP:** Freshwater KBAs completely covered by protected areas (SDG 15.1.2) (Percentage)(BirdLife Internationa, 2018)
- **POP14:** Child population 0-14 (% of total) (% of population)(UNPD, 2018)
- **POP65:** Elderly population 65 and above (% of total) (% of population)(UNPD, 2018)
- **POPG:** Population growth (Percentage)(UNPD, 2018)
- **IS_R:** Access to improved sanitation: rural (% of rural population) (UNMDG, 2018)
- **IS_U:** Access to improved sanitation: urban (% of urban population) (UNMDG, 2018)
- **IWS_R:** Access to improved water sources: rural (% of rural population) (UNMDG, 2018)
- **IWS_U:** Access to improved water sources: urban (% of urban population) (UNMDG, 2018)

The FSS, FPRO and FWP indicators, are the environmental indicators related to the freshwater disease risk. For instance, the countries with larger FSS (and relatively lower FPRO) has a higher risk of freshwater disease (Peeler and Feist , 2011). Indicators POP14, POP65 and POPG, indicate the structure of the population. These indicators are included in the study due to the fact that on one hand, child and elderly populations are more vulnerable in disease outbreaks. On the other hand, the larger child population increase the risk of break out since they usually are cureless while the elderly population are more cautious and usually more experienced. Indicators IS_R, IS_U, IWS_R and IWS_U are related to government policies and infrastructural developments related to clean water resources.

[2] The rest is downloaded from `http://environmentlive.unep.org/downloader`

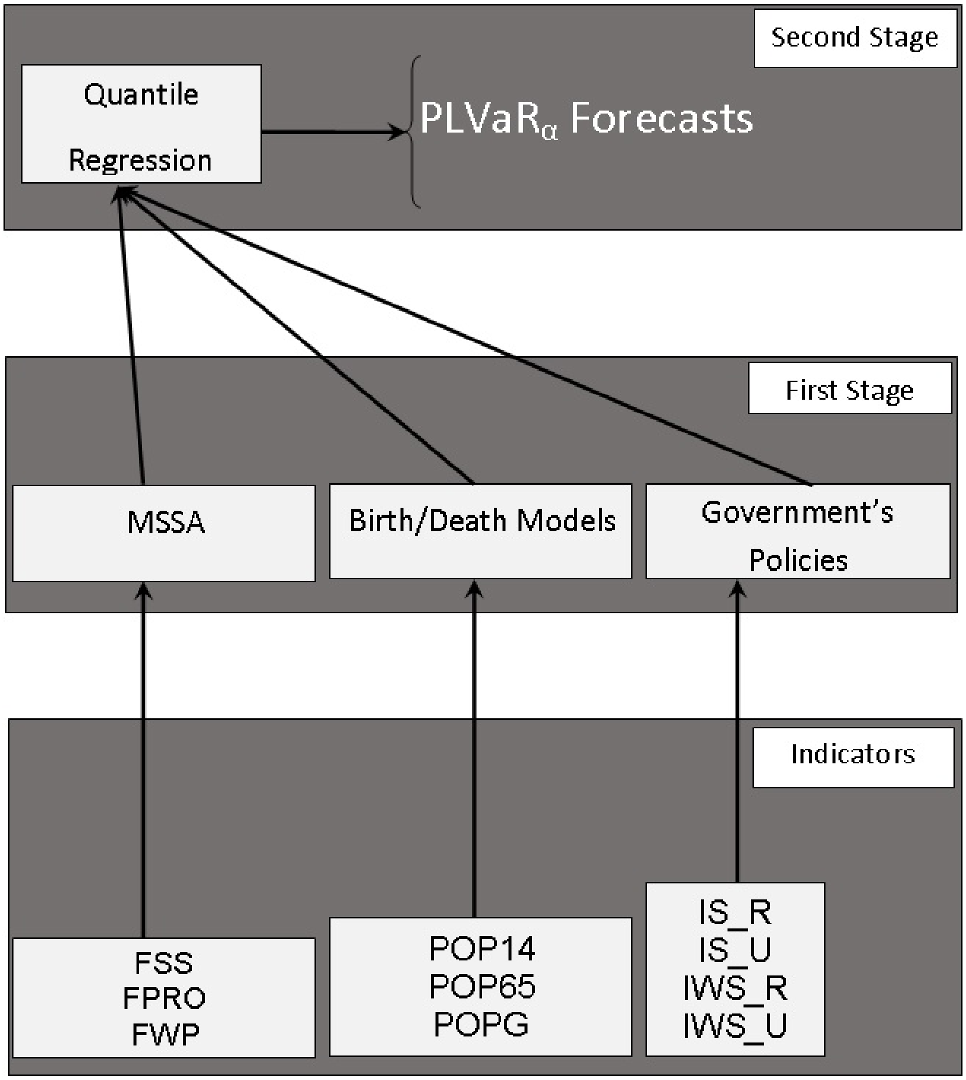

Figure 1: MSSA-QR model for waterborne disease PLVaR forecasting

The $PLVaR$ is forecasted using the MSSA-QR model for confidence levels 0.9, 0.95 and 0.99 (risk levels $\alpha = 0.1, 0.05, 0.01$). Figure 1 shows the diagram of the model.

In the first stage, MSSA is applied to FSS, FPRO and FWP as environmental indicators. The number of components in MSSA is selected based on minimum in-sample RMSE, using the data available before 2011. Since we do not have access to government policies on water and sanitation resources (i.e. IS_R, IS_U, IWS_R and IWS_U) in all of these 22 countries, MSSA is used to

Table 3: Out-of-sample RMSE produced by HMSSA-R, the number of components and window length in MSSA.

	RMSE						
Indicator	2011	2012	2013	2014	2015	r†	L‡
FSS	11.9396	16.1707	16.3747	16.587	.^{a}	2	31
FPRO	1.69E+05	1.78E+05	1.57E+05	1.92E+05	1.58E+05	1	10
FWP	11.1997	13.5262	19.1778	21.2637	23.5379	1	10
IS_R	2.0862	2.5818	3.0744	3.505	3.5473	1	7
IS_U	0.5605	0.5576	0.56	0.5676	0.9434	1	7
IWS_R	2.0018	2.3421	2.6736	2.9148	2.9185	1	11
IWS_U	0.6103	0.6915	0.787	0.8225	0.8248	1	11

.† Number of components selected based on minimum in-sample RMSE

.‡ Window length selected based on minimum in-sample RMSE

.^{a} The RMSE is not calculated since the 2015 observation is not available for any of the countries.

forecast these indicators too. The out-of-sample RMSE is calculated based on the forecasts for 2011 to 2015. Table 3 shows the out-of-sample RMSE for each year and indicator. As mentioned before, the POP14, POP65 and POPG indicator forecasts are available based on Berth/Death models from `http://www.un.org/en/development/desa/population/`.

In the second stage, the data from 1998 to 2010 are used to estimate the QR model coefficients in each confidence level. Table 4 shows the exceedance rate (ER_α) in each disease and confidence level for the estimated PLVaR. The out-of-sample ER_α for forecasted PLVaR (from 2011 to 2015) are given in Tables 5 and 6.

According to the Table 4, the in-sample ER_α is less than the risk level for most diseases. In more common diseases, (i.e. Diarrhoea, Pertussis and Malaria), however, the ER_α is slightly larger than the risk level. We record similar results during the out-of-sample forecasting exercise. Tables 5 and 6 show that in all time horizons (from 2011 to 2015), for less common diseases, the ER_α does not exceed the risk level.

Table 4: In-sample Exceedance rate (ER_α) for estimated PLVaR based on 1998-2010 data.

	Confidence Level[†]				Confidence Level[†]		
Disease	0.9	0.95	0.99	Disease	0.9	0.95	0.99
Chlamydia	0.0185	0.0185	0.0074	Dengue	0.0296	0.0185	0.0000
Diarrhoeal Diseases	0.1148	0.0704	0.0074	Japanese Encephalitis	0.0185	0.0185	0.0037
Pertussis	0.1000	0.0556	0.0333	Trachoma	0.0185	0.0185	0.0037
Poliomyelitis	0.0741	0.0667	0.0000	Ascariasis	0.0333	0.0222	0.0148
Malaria	0.0807	0.0526	0.0246	Trichuriasis	0.0037	0.0037	0.0037
Schistosomiasis	0.0741	0.0519	0.0185	Hookworm	0.0222	0.0148	0.0000
Onchocerciasis	0.0037	0.0037	0.0037				

.[†] Confidence Level is 1 - α where α is risk level.

Overall, according to these results, it is evident that the MSSA-QR model and the forecasted PLVaR values can be used as useful measures for forecasting the worst case scenario in waterborne disease control and prevention. The model is not without its weaknesses, as we notice that it struggles at forecasting the more common disease like Diarrhoea, Pertussis and Malaria. However, we believe the performance for these diseases could be improved using more indicators. This is because the more common diseases are usually affected by more socioeconomic and environmental variables. For instance, the climatological and economic-development variables could affect the risk of a Malaria outbreak.

4. Conclusion

In this paper, a new model for forecasting the disease outbreak risk is proposed. In order to quantify the risk, we adopt a risk measure from financial risk analysis and develop the Population Loss Value at Risk (PLVaR) as a measure of disease outbreak risk. The larger values of PLVaR show the bigger risk of disease outbreak. The PLVaR is forecasted using a two stage model based on Multivariate Singular Spectrum Analysis and Quantile Regression (MSSA-QR model). The proposed risk measure (PLVaR) and

Table 5: Out-of-sample Exceedance rate (ER_α) for estimated PLVaR.

	Confidance	ER_α				
Disease	Level[†]	2011	2012	2013	2014	2015
Chlamydia	0.9	0.0000	0.0000	0.0000	0.0000	0.0000
	0.95	0.0000	0.0000	0.0000	0.0000	0.0000
	0.99	0.0000	0.0000	0.0000	0.0000	0.0000
Diarrhoeal Diseases	0.9	0.4091	0.3636	0.3636	0.2857	0.3684
	0.95	0.3182	0.3182	0.2727	0.2857	0.2632
	0.99	0.2727	0.2273	0.2273	0.1905	0.2105
Pertussis	0.9	0.1364	0.1818	0.2727	0.2857	0.2105
	0.95	0.1364	0.1364	0.1818	0.2381	0.1053
	0.99	0.0909	0.1364	0.1364	0.1905	0.1053
Poliomyelitis	0.9	0.0455	0.0909	0.0909	0.0476	0.1053
	0.95	0.0000	0.0455	0.0455	0.0000	0.1053
	0.99	0.0000	0.0455	0.0455	0.0000	0.0000
Malaria	0.9	0.0455	0.1364	0.0455	0.0476	0.1053
	0.95	0.1364	0.1364	0.1818	0.1905	0.2632
	0.99	0.0909	0.0909	0.0455	0.0000	0.1053
Schistosomiasis	0.9	0.0000	0.1364	0.0000	0.0476	0.0000
	0.95	0.0000	0.1364	0.0000	0.0476	0.0000
	0.99	0.0000	0.0000	0.0000	0.0000	0.0000
Onchocerciasis	0.9	0.0000	0.0455	0.0455	0.0000	0.0000
	0.95	0.0000	0.0455	0.0455	0.0000	0.0000
	0.99	0.0000	0.0455	0.0455	0.0000	0.0000

.[†] Confidence Level is 1 - α where α is risk level.

Table 6: Out-of-sample Exceedance rate (ER_α) for estimated PLVaR.

	Confidance	ER_α				
Disease	Level[†]	2011	2012	2013	2014	2015
Dengue	0.9	0.0455	0.0455	0.1364	0.1429	0.1053
	0.95	0.0455	0.0455	0.0909	0.0952	0.0526
	0.99	0.0000	0.0000	0.0000	0.0000	0.0000
Japanese Encephalitis	0.9	0.0000	0.0455	0.0000	0.0476	0.0000
	0.95	0.0000	0.0455	0.0000	0.0476	0.0000
	0.99	0.0000	0.0455	0.0000	0.0000	0.0000
Trachoma	0.9	0.0000	0.0000	0.0000	0.0000	0.0526
	0.95	0.0000	0.0000	0.0000	0.0000	0.0526
	0.99	0.0000	0.0000	0.0000	0.0000	0.0526
Ascariasis	0.9	0.0000	0.0909	0.0000	0.0000	0.0000
	0.95	0.0000	0.0909	0.0000	0.0000	0.0000
	0.99	0.0000	0.0000	0.0000	0.0000	0.0000
Trichuriasis	0.9	0.0455	0.0000	0.0000	0.0000	0.0000
	0.95	0.0455	0.0000	0.0000	0.0000	0.0000
	0.99	0.0455	0.0000	0.0000	0.0000	0.0000
Hookworm	0.9	0.0455	0.0000	0.0000	0.0476	0.0526
	0.95	0.0455	0.0000	0.0000	0.0476	0.0526
	0.99	0.0000	0.0000	0.0000	0.0000	0.0526

.[†] Confidence Level is 1 - α where α is risk level.

forecasting model (MASS-QR) is used to forecast the worst cases of waterborne disease outbreaks in 22 European and North American countries based on socio-economic and environmental indicators. The results show that the proposed method perfectly forecasts the worst case scenario for less common waterborne diseases. According to our findings, the forecasting of more common diseases needs more socio-economic and environmental indicators.

We evidence that the proposed method has the ability to forecast the worst case scenarios in disease outbreak and provides a practical tool for policy makers and health institutions to control and prevent the outbreaks. Furthermore, introducing a PLVaR as a risk measure adopted from financial risk analysis opens a new door to epidemiological and environmental risk analysis using other risk analysis tools in finance. For instance, using PLVaR, one may adopt the copula method to investigate the relations between different outbreaks. Moreover, more research is required into developing and evaluating the accuracy of the proposed PLVar, MSSA-QR model at forecasting the risk of disease outbreaks in more common diseases.

Alizon, S., de Roode, J. C., and Michalakis, Y. (2013), Multiple infections and the evolution of virulence. *Ecology Letters*, 16(4): 556567.

BirdLife International, IUCN and UNEP(https://www.ibat-alliance.org/ibat)

Davino, C., Furno, M. and Vistocco, D. (2014). *Quantile Regression Theory and Application*, United Kingdom: John Wiley.

FAO, Fisheries and Aquaculture Department(http://www.fao.org/fishery/statistics/global)

Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

Graham, M., Suk, J. E., Takahashi, S., Metcalf, C. J., Jimenez, A. P., Prikazsky, V., Ferrari, M. J., and Lessler, J. (2018), Challenges and Opportunities in Disease Forecasting in Outbreak Settings: A Case Study of Measles in Lola Prefecture, Guinea, *The American Journal of Tropical Medicine and Hygeine*, 98(5): 14891497.

Griffiths, E., Pedersen, A. B., Fenton, A., and Petchey, O. L. (2011), The nature and consequences of coinfection in humans. *Journal of Infection*, 63(3): 200206.

Han, B.A. and Drake, J.M. (2016). Future directions in analytics for infectious disease intelligence, *EMBO Reports*, 17: 785-789.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Leavens, D. H. (1945). Diversification of investments, *Trusts and Estates*, 80(5), 469-473.

Liao, Y., Xu, B., Wang, J. and Liu, X. (2017). A new method for assessing the risk of infectious disease outbreak, *Scientific Reports*, 7:40084.

Lowe, R., Stewart-Ibarra A.M., Petrova, D., Garcia-Diez, M., Borbor-Cordova, M.J., Mejia, R., Regato, M. and Rodó, X. (2017). Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador, *The Lancet Planetary Health*, 4: e142-e151.

McNeil, A.J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management*, United States: Princeton University Press.

Metcalf, C. J., and Lessler, J. (2018), Opportunities and challenges in modeling emerging infectious diseases. *Science*, 357(6347): 149152.

MDG(http://mdgs.un.org/), MDG Indicators Database(http://mdgs.un.org/unsd/mdg/default.aspx)

Peeler, E.J. and Feist, S.W. (2011), Human intervention in freshwater ecosystems drives disease emergence, *Freshwater Biology*, 56: 705-716. doi:10.1111/j.1365-2427.2011.02572.x

Sanei, S. and Hassani, H. (2015). *Singular Spectrum Analysis of Biomedical Signals.* United States: CRC Press.

UNPD(http://www.un.org/en/development/desa/population/), World Population

Wendling, Z., D. Esty, J. Emerson, M. Levy, A. de Sherbinin, et al. 2018. The 2018 Environmental Performance Index Report. New Haven, CT: Yale Center for Environmental Law and Policy. https://epi.envirocenter.yale.edu/node/36476.

Yale Center for Environmental Law and Policy - YCELP - Yale University, Yale Data-Driven Environmental Solutions Group - Yale University, Center for International Earth Science Information Network - CIESIN - Columbia University, and World Economic Forum - WEF. 2018. 2018 Environmental Performance Index (EPI). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4X928CF.