Real-time prediction of severe influenza epidemics using Extreme Value Statistics

Maud Thomas, Holger Rootzén

Real-time prediction of severe influenza epidemics using Extreme Value Statistics

Maud Thomas
Sorbonne Université, CNRS, LPSM, 4 place Jussieu, F-75005 Paris, France,
maud.thomas@sorbonne-universite.fr
and
Holger Rootzén
Mathematical Sciences, Chalmers University and University of Gothenburg,
Gothenburg, Sweden

Summary: Each year, seasonal influenza epidemics cause hundreds of thousands of deaths worldwide and put high loads on health care systems. A main concern for resource planning is the risk of exceptionally severe epidemics. Taking advantage of the weekly influenza case reporting in France, we use recent results on multivariate GP models in Extreme Value Statistics to develop methods for real-time prediction of the risk that an ongoing epidemic will be exceptionally severe and for real-time detection of anomalous epidemics. Quality of predictions is assessed on observed and simulated data.

Keywords: Anomaly detection; Extreme Value Statistics; Generalized Pareto models; Influenza epidemics; Real-time prediction

1 Introduction

Every year, seasonal influenza epidemics cause 250,000–500,000 deaths worldwide [Rambaut et al., 2008] and put high strain on public health care systems in a short time frame, due essentially to increased doctor visits, and overcrowded emergency departments and intensive care units. Predicting the likelihood of an exceptionally severe epidemic in the future is of paramount importance for health resource planning [Bresee and Hayden, 2013, Khan and Lurie, 2014]. The three following questions are thus of central interest to public health policy makers.

(Q1) estimation of risks of occurrence of a very severe epidemic during the following years,

(Q2) real-time prediction of the risk that an ongoing epidemic will be exceptionally severe, and

(Q3) real-time detection of unusual, thus potentially dangerous, epidemics.

In France, countrywide data on seasonal influenza morbidity have been available since 1985 (see Figure 1 below). The Sentinelles network monitors cases of Influenza-like Illness (ILI) defined by the presence of fever in excess of 39°C, respiratory symptoms, and muscle pains [Réseau Sentinelles, 2019]. Though only a fraction of the ILI reported visits are in fact caused by influenza, their total number reflects the burden on the health care system.

Figure 1 suggests that influenza epidemics, or at least a proportion of them, may be conceptualised as extreme episodes that occur within the ILI time series. This led us to lean on Extreme Value Statistics (EVS) to address the above questions. EVS have been developed to handle extreme events, such as extreme floods, heat waves or episodes with huge financial losses [e.g. Katz et al., 2002, Embrechts et al., 1997]. EVS allow for prediction of risks of episodes outside of the observed range.

Question (Q1) may be answered by standard and well established EVS methods. For instance, they were applied by Chen et al. [2015] to avian influenza, and by Thomas et al. [2016] to influenza mortality, and emergency department visits.

Previous approaches to Question (Q2) include high dimensional times series prediction [Davis et al., 2016] and the method of analogues [Réseau Sentinelles, 2019]. The US Center for Disease Control initiated a data challenge to predict the 2013–2014 US influenza epidemic. Their conclusion was "Forecasting has become technically feasible, but further efforts are needed to improve forecast accuracy so that policy makers can reliably use these predictions" [Biggerstaff et al., 2016]. Building on recently developed EVS results and methods based on multivariate Generalized Pareto (GP) distributions [Rootzén et al., 2018, Kiriliouk et al., 2019], we extend the toolbox of available techniques and develop methods to tackle Question (Q2).

Question (Q3) is an anomaly detection problem. Recent publications have used EVS in this context [Guillou et al., 2014, Thomas et al., 2017, Goix, 2016, Chiapino et al., 2019]. In the present paper, the multivariate GP models estimated in the course of solving Question (Q2) are used to detect anomalous epidemics.

For predictions to be useful in practice, it is necessary to have an understanding of their reliability. Methods for evaluating the quality of prediction of extremes do not seem available in the literature. Here, we describe a strategy that uses standardized Brier

scores, Precision-Recall curves and Average Precision scores [Steyerberg et al., 2010, Brownlee, 2020, Saito and Rehmsmeier, 2015].

Section 2 presents the Sentinelles network data and the definition of epidemics. Section 3 describes the methods and in particular the multivariate GP distributions. In Section 4 the EVS methods are applied to the Sentinelles ILI data and Section 5 develops the strategy for assessing the accuracy of the real-time prediction of extremes. Finally, Section 6 contains the conclusion.

2 The Sentinelles network data

The French nationwide Sentinelles network consists of approximately 1,500 general practitioners in France who participate on a voluntary basis in the ILI surveillance, and report new cases of ILI observed in their practice. Based on these data, nationwide weekly ILI incidence rates—i.e. numbers of new cases in France per week per 100,000 individuals—are estimated. Epidemics are identified using the Serfling method [Serfling, 1963]. It consists in fitting a cyclic regression model to the weekly ILI rates and setting the start of the epidemic at the first of the first two consecutive weeks during which the ILI incidence rates exceed the upper bound of the 90% prediction interval [for further details see Réseau Sentinelles, 2019].

Weekly ILI incidence rates from January 1985 to February 2019 were downloaded for analysis. Figure 1 shows the time series of weekly ILI incidence rates, which includes 35 epidemics. The epidemic with the highest peak corresponds to the 1989-epidemic, with a value of 1,793. The lowest peak was observed in 2014, with a value of 325. The durations of the Serfling epidemics range from 5 to 16 weeks, in 1991 and 2010, respectively.

In this study, the 1985–2018 data were used for estimation and the 2019 data were kept as a test sample. Since we were interested in very high ILI incidence rates and relied on EVS methods, we focused on the most active part of the epidemics. The Serfling method was thus adapted as follows. The start of the epidemic was set at the first of the first two consecutive weeks during which the ILI incidence rates exceeded a given constant threshold. The end was set at the end of the Serfling epidemic. The threshold was chosen as the 0.88-quantile (=272) of the 1985-2018 data to synchronise the peaks of the epidemics as much as possible, see Figure 2. This definition detected 34 epidemics between 1985 and 2018, and was thus in accordance with the Sentinelles network. The durations of the 34 epidemics ranged from 3 to 12 weeks in 2014 and 1985/2010, respectively.

The size of an epidemic was defined as the sum of the weekly ILI incidence rates from the start to the end of the epidemic. The smallest epidemic size was 847 in 2014 and the largest was 8,062 in 1989.

The size of an influenza epidemic depends on multiple factors including immunity and vaccination prevalence in the population. Some of these factors may likely be impacted by the sizes of past epidemics. Nevertheless, as there is little indication that this translates into statistical dependence between epidemics—shown by the correlation and distance correlation plots in the Supplementary Material (Section 1)—we thus assumed that the ILI incidence rates of different epidemics were mutually independent.

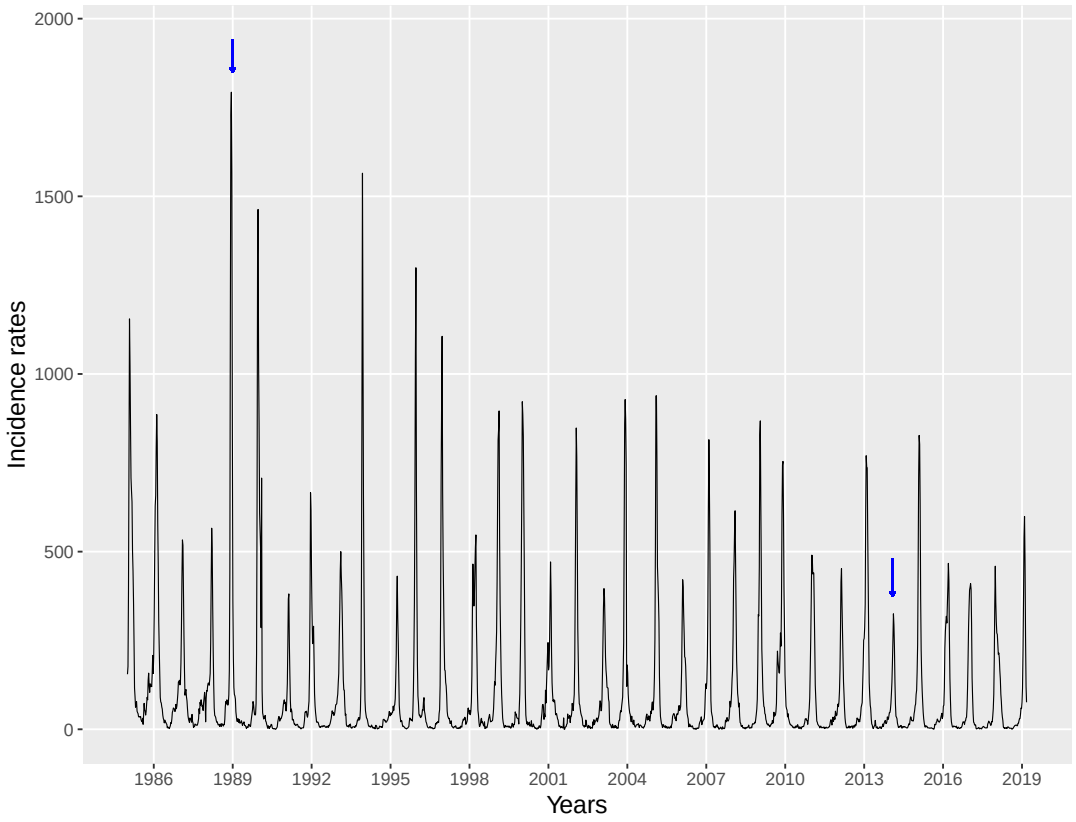
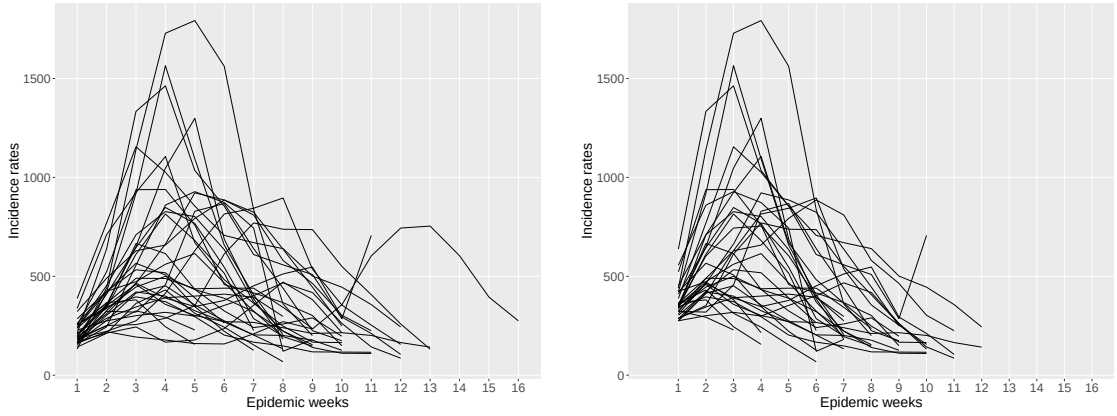


Figure 1: Weekly ILI incidence rates in metropolitan France from January 1985 to February 2019. Arrows indicate epidemics with the highest and the lowest peaks.



a) b)

Figure 2: a) Weekly ILI incidence rates for epidemics identified with the Serfling method, and b) Weekly ILI incidence rates for epidemics obtained from the definition in this paper.

3 Methods: EVS modelling

The main interest of EVS methods is to allow prediction outside of the range of the observations. In this section, we describe the EVS tools required to answer the questions raised in this paper.

Section 3.1 presents the univariate Peaks over Thresholds (PoT) method which is used for Question(Q1). Our approach to Question (Q2) is described in Section 3.2: it is based on the multivariate PoT method, dwelling on recent results on multivariate EVS models [Rootzén et al., 2018, Kiriliouk et al., 2019]. Section 3.3 deals with Question (Q3) and combines a general framework for anomaly detection with Generalized Pareto (GP) modelling.

3.1 Univariate PoT method

The one-dimensional PoT method, introduced in the hydrological literature by Smith [1984] consists in selecting a suitably high threshold u and defining excesses above u as the differences between the observations and u. Under general conditions, the conditional distribution of the excesses, given that they are positive, is asymptotically as $u \to \infty$ a GP distribution with cumulative distribution function (cdf)

$$H(x) = \begin{cases} 1 - (1 + \frac{\gamma}{\sigma} x)_+^{-1/\gamma} & \text{if } \gamma \neq 0,\ x \geq 0 \\ 1 - \exp(-\frac{x}{\sigma}) & \text{if } \gamma = 0,\ x \geq 0. \end{cases} \tag{1}$$

Here $\sigma > 0$ is a scale parameter, and $\gamma \in \mathbb{R}$ is a shape parameter. For $a \in \mathbb{R}$, $(a)_+ = a$ if $a \geq 0$ and 0 if $a < 0$. The parametrisation is chosen so that the cdf for $\gamma = 0$ is the limit as $\gamma \to 0$ of the cdf with non-zero γ. A useful account of this method is given in [Coles, 2001].

Assuming the observations are independent and identically distributed with common cdf F, the parameters of the cdf (1) are estimated from the observed excesses. For $y > u$, $F(y)$ is estimated, by

$$\widehat{F}(y) = 1 - \widehat{p}_u(1 - \widehat{H}(y - u)),$$

where $\widehat{H}$ is the GP cdf (1) with parameters replaced by their estimated values, and $\widehat{p}_u$ is the empirical frequency of observations which exceed the threshold u [Coles, 2001].

The level that the maximum of n observations will exceed with probability $1 - \alpha$ is estimated by the α-th quantile $\widehat{y}_\alpha$ of $\widehat{F}(y)^n$. For example, if $\gamma = 0$, H is an exponential distribution and

$$\widehat{y}_\alpha = u + \widehat{\sigma}\{\log \widehat{p}_u - \log(1 - \alpha^{1/n})\}, \tag{2}$$

for α such that $\widehat{p}_u \geq 1 - \alpha^{1/n}$.

3.2 Multivariate PoT method

Our approach to real-time prediction is to predict characteristics of interest of an ongoing epidemic, conditionally on the latest observed ILI incidence rates. Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ be a random vector where Y_1 is next to last and Y_2 last observed ILI incidence rates

and Y_3 is a characteristic of interest. For prediction, we need to specify the multivariate distribution of $\boldsymbol{Y}$.

The multivariate PoT method was introduced by [Michel, 2009, Brodin and Rootzén, 2009]. A suitably high threshold is chosen for each component of $\boldsymbol{Y}$. Excess vectors are defined as the component-wise differences between observations and thresholds and considered as positive if at least one of the components exceed its threshold. Under general conditions, the joint distribution of the positive excess vectors is asymptotically a multivariate GP distribution as the thresholds tend to ∞. For the sake of completeness, we briefly recall the definitions and properties of multivariate GP distributions that will be needed in this paper [for further details, see Rootzén et al., 2018, Kiriliouk et al., 2019].

Unlike in the univariate case, the family of multivariate GP distributions cannot be described as a parametric family. Rootzén et al. [2018] have developed four representations of multivariate GP distributions for which closed formulas of densities are available. In this paper, we use the U-representation since it presents nicer properties across the dimension, which are useful for prediction. Moreover, at the price of standardization, the marginals of the distributions can be assumed to be standard exponential distributions [see Section 3 in Kiriliouk et al., 2019].

For later use, we describe the U-representation in dimension 3. Let $\boldsymbol{U} = (U_1, U_2, U_3)$ be a 3-dimensional random vector such that $\mathbb{E}[\mathrm{e}^{\max \boldsymbol{U}}] < \infty$, where $\max \boldsymbol{U} = \max\{U_1, U_2, U_3\}$, and let $f_{\boldsymbol{U}}$ be the probability density function of $\boldsymbol{U}$. For $i = 1, 2, 3$, let f_i and F_i denote the density and distribution functions of U_i, respectively. The vector $\boldsymbol{U}$ or its distribution is referred to as the generator.

According to Equation (3.4) in [Kiriliouk et al., 2019], the density function $h_{\boldsymbol{U}}$ of the 3-dimensional GP distribution with standard exponential margins generated by $\boldsymbol{U}$ is

$$h_{\boldsymbol{U}}(\boldsymbol{x}) = \frac{1_{\{\boldsymbol{x} \not\leq \boldsymbol{0}\}}}{\mathbb{E}[\mathrm{e}^{\max \boldsymbol{U}}]} \int_0^\infty f_{\boldsymbol{U}}(\boldsymbol{x} + \log t) \mathrm{d}t, \tag{3}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$ and $\boldsymbol{x} + \log t = (x_1 + \log t, x_2 + \log t, x_3 + \log t)$. The indicator function $1_{\{\boldsymbol{x} \not\leq \boldsymbol{0}\}}$ equals one if at least one of the components of $\boldsymbol{x}$ is positive, and is zero otherwise. Different choices of distributions for $\mathbf{U}$ yield different GP models.

Question (Q2) was formulated as the conditional prediction that Y_3 will exceed some level ℓ given Y_1 and Y_2. Assuming that the corresponding excess vector $\boldsymbol{X} = (X_1, X_2, X_3)$ is positive and follows a multivariate GP distribution with generator $\mathbf{U}$, then

$$P[X_3 \geq \ell | X_2 = x_2, X_1 = x_1] = \frac{\int_{x_3=\ell}^\infty 1_{\{\boldsymbol{x} \not\leq \boldsymbol{0}\}} \int_0^\infty f_{\boldsymbol{U}}(\boldsymbol{x} + \log t) \mathrm{d}t \mathrm{d}x_3}{\int_{x_3=-\infty}^\infty 1_{\{\boldsymbol{x} \not\leq \boldsymbol{0}\}} \int_0^\infty f_{\boldsymbol{U}}(\boldsymbol{x} + \log t) \mathrm{d}t \mathrm{d}x_3}. \tag{4}$$

Formulas for general $\mathbf{U}$ distributions are given in the Supplementary Material (Proposition 1, Section 3).

3.3 Anomaly detection

Question (Q3) belongs to the field of anomaly detection. In the present context, a natural approach is to use the estimated GP model from Section 3.2 to detect epidemics that exhibit a significantly different pattern from the data used to fit the model. A statistical

test for anomalous epidemics may be based on the GP negative log-likelihood, with a very large value suggesting that the new observation might be anomalous. The quantiles of the GP negative log-likelihood distribution were estimated by simulation in order to define the decision region of the test [e.g. Section 2 of Root et al., 2015]. However, it must be stressed that "anomalous" has to be understood with respect to the fitted GP model.

4 Results: Prediction of very high ILI loads on the health care system

In this section, the methods described in Section 3 are applied to the Sentinelles ILI data. Recall that the 1985–2018 data were used for estimation and the 2019 data were kept as a test sample.

For each epidemic, we shall refer to the first week of the epidemic as Week 1, the second week as Week 2 and so on until the end of the epidemic. For $j = 1, 2, 3$ and $k = 1985, \ldots, 2019$, we let Y_j^k denote the ILI incidence rate of Week j in year k, and S^k denote the size of the epidemic of year k. We omit the index k when we refer to a generic epidemic.

To address Questions (Q1), (Q2) and (Q3), we focus on predictions for the Week 3 ILI incidence rate Y_3 and the epidemic size S.

4.1 Question (Q1): Risk of very high ILI incidence rates and epidemic sizes over the following years

The univariate PoT method was applied to the Week 3 ILI incidence rates $(Y_3^{1985}, \ldots, Y_3^{2018})$ and to the epidemic sizes $(S^{1985}, \ldots, S^{2018})$.

As explained in Section 3.1, the first step is to choose a suitably high threshold for each series of observations. The threshold must be high enough to ensure that the asymptotic model is valid, but low enough to yield a sufficient number of positive excesses. For the ILI rates, the threshold u_I was chosen as the 0.9-quantile (=339) of the whole 1985-2018 series; for epidemic sizes, the threshold u_S as the 0.6-quantile (=4,144) of the series of epidemic sizes from 1985 to 2018. These thresholds yield 30 positive excesses for Week 3 ILI rates and 14 for epidemic sizes.

One-dimensional GP distributions (Equation (1)) were fitted to the ILI rate and size positive excesses. Likelihood ratio tests showed that the hypothesis $\gamma = 0$ was not rejected ($p = 0.64$, and $p = 0.98$ respectively), so that cdf's were assumed exponential. QQ-plots and estimates of scale parameters are shown in the Supplementary Material (Section 2, Figure 2 and Table 1).

Table 1 presents estimates of the levels $\widehat{y}_\alpha$ that during the following year and the following 10 years the Week 3 ILI incidence rate and the epidemic size will exceed with given probability $1 - \alpha$. These estimates were computed using Equation (2). For Week 3 ILI incidence rates, $u_I = 339$ yielded $\widehat{p}_{u_I} = 0.88$ and for epidemic sizes $u_S = 4,441$ yielded $\widehat{p}_{u_S} = 0.41$. For example, it was estimated that there is a 10% probability that the epidemic size will exceed 9,385 at least once during the next 10 years.

Table 1: Estimated levels that the Week 3 ILI incidence rate and the epidemic size will exceed with either probability 10% or 1% during the following year and the following 10 years.

$1-\alpha$	one year 10%	one year 1%	10 years 10%	10 years 1%
Week 3 ILI incidence rates	1,192	2,094	2,076	2,994
Epidemic sizes	6,165	9,452	9,385	12,733

4.2 Question (Q2): Real-time prediction of very high ILI incidence rates and epidemic sizes

In this section, the multivariate PoT method is applied to $(\mathbf{Y}^{1985}, \ldots, \mathbf{Y}^{2018})$ and $(\mathbf{S}^{1985}, \ldots, \mathbf{S}^{2018})$ where $\mathbf{Y}^k = (Y_1^k, Y_2^k, Y_3^k)$ and $\mathbf{S}^k = (Y_1^k, Y_2^k, S^k)$ for $k = 1985, \ldots, 2018$.

The thresholds u_I of 339 for the ILI incidence rates Y_1, Y_2 and Y_3 and u_S of $4,144$ for the epidemic size S were as in the previous section. For both $(\mathbf{Y}^{1985}, \ldots, \mathbf{Y}^{2018})$ and $(\mathbf{S}^{1985}, \ldots, \mathbf{S}^{2018})$, there were 32 positive excess vectors as defined in Section 3.2. In order to meet the assumption of standard exponential marginals, positive excess vectors were standardized by dividing each component by the corresponding scale parameter estimates (Table 1, Supplementary Material).

To define a multivariate GP model, the distribution of the generator $\mathbf{U} = (U_1, U_2, U_3)$ must be specified. Assuming that the three components U_1, U_2, U_3 of $\mathbf{U}$ were mutually independent, three families of marginal distributions were considered: Gumbel, reverse exponential, and reverse Gumbel distributions. The corresponding models were fitted to the 32 positive excess vectors. Formulas for the densities $h_{\boldsymbol{U}}$ for the three GP models are given in the Supplementary Material (Section 4).

Table 2 shows that both in terms of AIC and BIC, the best fit was that of the GP family with Gumbel generator, for both $\mathbf{Y}$ and $\mathbf{S}$. According to Equation (3), the corresponding density is

$$h_{\mathbf{U}}(\mathbf{x}) = \frac{\int_0^\infty \prod_{i=1}^3 \alpha_i (t\mathrm{e}^{x_i-\beta_i})^{-\alpha_i} \mathrm{e}^{-(t\mathrm{e}^{x_i-\beta_i})^{-\alpha_i}} \mathrm{d}t}{\int_0^\infty \left(1 - \prod_{i=1}^3 \mathrm{e}^{-(t/\mathrm{e}^{\beta_j})^{-\alpha_j}}\right) \mathrm{d}t}, \tag{5}$$

with $\alpha_1, \alpha_2, \alpha_3 > 1$ and $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$ (for further details see Supplementary Material, Section 4 and [Kiriliouk et al., 2019]). In the sequel, we refer to this GP model as the Gumbel model. The estimated parameters of the Gumbel model are given in the Supplementary Material (Table 2). More parsimonious sub-models were considered and consistently rejected on the basis of AIC, BIC and log-likelihood ratio tests (Supplementary Material, Table 3).

The estimated model was used to provide estimates of the probability that Y_3 and S will exceed a specified level given that Y_1 and Y_2 have been observed. Since Equation (4) is valid for positive excess vectors only, two situations must be considered

i) If at least one of Y_1 or Y_2 exceeds its threshold, Equation (4) is valid whatever Y_3.

Table 2: AIC and BIC of GP models for Week 3 ILI incidence rates and for epidemic sizes

Generator	Gumbel	Reverse exponential	Reverse Gumbel
AIC	194	2189	208
BIC	202	2196	215

a) Week 3 ILI incidence rates

Generator	Gumbel	Reverse exponential	Reverse Gumbel
AIC	227	2240	268
BIC	234	2248	275

b) Epidemic sizes

ii) If neither Y_1 nor Y_2 exceeds its threshold, then whether the excess vector will be positive or not is unknown. In this case, the right-hand side of Equation (4) must be multiplied by the probability that Y_3 (or S) exceeds its threshold given that Y_1 and Y_2 do not exceed theirs. The corresponding empirical probability was 0.33 for both Y_3 and S.

The procedure is illustrated in Table 3 which presents predictions for the 2019 epidemic. The largest observed Week 3 ILI incidence rate between 1985 and 2018 was 1,729. The table shows the estimated probabilities that the 2019 Week 3 ILI incidence rate exceeds a fraction κ of the 1985-2018 maximum ILI incidence rate 1,729 for $\kappa = 0.5, 0.75, 0.95, 1$. Estimates for epidemic sizes are presented similarly with a largest observed epidemic size of 7,241. The prediction probabilities, even for the lowest level were quite small, and in effect this level was not exceeded in 2019. For the 2019 epidemic the Weeks 1, 2 and 3 ILI incidence rates were 336, 540, and 500, respectively, and the epidemic size was 1,192.

Table 3: Prediction probabilities for the 2019 epidemic of exceedances of levels 1,729$\times\kappa$ for Week 3 ILI incidence rates and 7,241$\times\kappa$ for epidemic sizes.

	Week 3 ILI incidence rates				Epidemic sizes			
κ	0.5	0.75	0.95	1	0.5	0.75	0.95	1
Level	864	1,297	1,643	1,729	4,031	6,046	7,659	8,062
Probability	0.185	0.012	0.001	0.0007	0.026	0.008	0.003	0.002

4.3 Question (Q3): Real-time prediction of anomalous epidemics

The quantiles of the GP negative log-likelihood were obtained as follows: the estimated Gumbel model was used to generate 1,500 datasets, each consisting of 33 three-dimensional positive excess vectors. For each simulated dataset, a Gumbel model was fitted to the 32 first vectors and the estimated negative log-likelihood was computed at the 33rd vector. The significance levels obtained as the empirical quantiles of these negative log-likelihoods are shown in Table 4.

Table 4: Quantiles of the estimated negative log-likelihood.

Significance level	10%	5%	1%	0.1%
Quantile	4.72	5.60	7.79	14.50

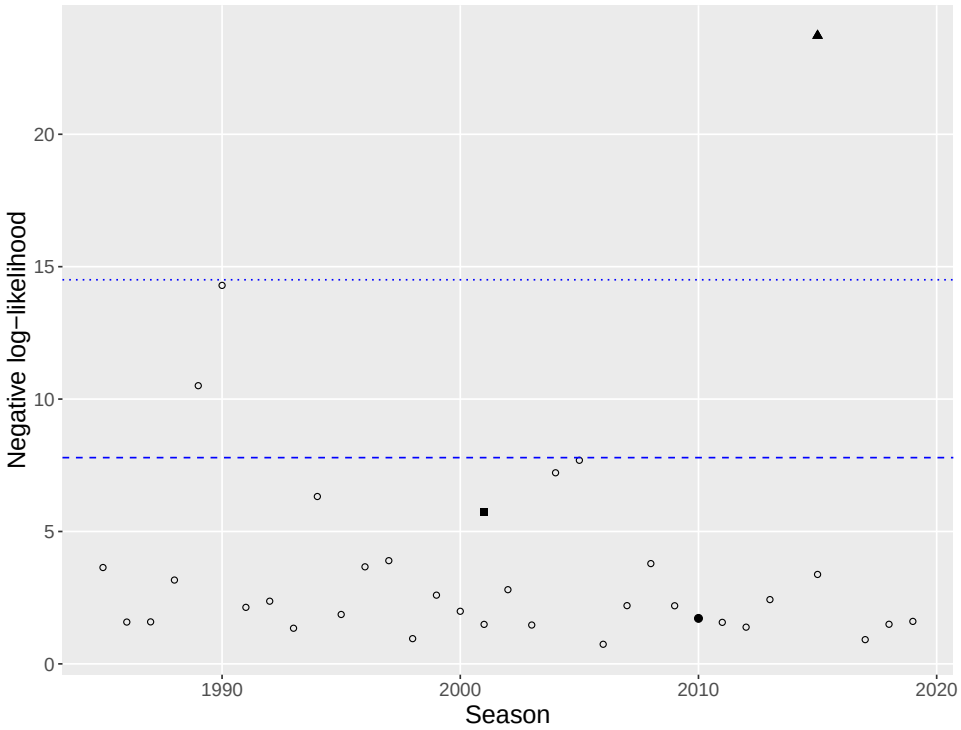


Figure 3: Leave-one-out negative Gumbel log-likelihoods for the ILI incidence rates of Weeks 1-3 of the 33 epidemics with positive excesss vectors observed between 1985 and 2019 (open circles). Swine flu pandemic (closed circle), simulated point with very large third component (closed square), and simulated anomalous point (closed triangle). Dashed and dotted blue lines are quantiles for the 1% and 0.1% significance levels.

To illustrate the meaning of "anomalous", we simulated two extra positive excess vectors: (i) a positive excess vector with a very high third component equal to the

0.99-quantile of the third component of simulated positive excess vectors and (ii) an anomalous positive excess vector obtained from (i) by multiplying the first and the third components by 1.5 and the second by 0.5. Figure 3 shows that only the anomalous point (ii) exceeds the 0.999-quantile. Interestingly, the 2009-10 swine flu pandemic was not unusual compared to the others epidemics.

5 Assessment of real-time predictions

This section presents a strategy for assessing real-time prediction of exceedances of very high levels, and its application to the 1985–2019 ILI data.

5.1 Methods

There is a substantial literature on assessment of forecasting [see Lerch et al., 2017, and references therein]. To our knowledge, apart from [Renard et al., 2013], the literature provides metrics aimed at comparing predictions with observations. However, standard assessment metrics are not appropriate when the frequency of exceedances in the data is small.

To overcome this drawback, we use (i) standardized Brier scores; (ii) Precision-Recall Curves; and (iii) Average Precision scores, together with simulations from estimated models.

(i) The standardized Brier score is defined as

$$1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{p}_i - o_i\right)^2}{p(1-p)},$$

where N is the number of predictions, $\widehat{p}_i$ is the prediction probability of exceedance, $o_i = 1$ if an exceedance was observed, and 0 otherwise, and $p = \frac{1}{N}\sum_{i=1}^{N} o_i$, see e.g. [Steyerberg et al., 2010]. The score is bounded by 1, with larger values indicating better prediction. Using the predictor $\widehat{p}_i = p$ gives the value 0.

(ii) A question such as "will the Week 3 ILI incidence rate or the epidemic size be higher than a given high level, say the largest rate or size observed up to now?" may be formulated as a binary classification problem. The data are divided into two classes: Positives (exceedances) and Negatives (no exceedances). The strategy consists first in computing $\widehat{p}_i$ and then in comparing this estimate to some cut-off probability value p_c. If $\widehat{p}_i \geq p_c$ then the observation is assigned to the Positives class, and to the Negatives class otherwise. The "true positives" ("false positives") correspond to the observations that are correctly (incorrectly) assigned to the Positives class, and the "true negatives" ("false negatives") to the observations that are correctly (incorrectly) assigned to the Negatives class.

Common methods to assess the performance of binary classifiers include true positive and true negative rates, and ROC (Receiver Operating Characteristics) curves. These methods, however, are uninformative when the classes are severely imbalanced, which is the case when predicting very high level exceedances which are rare by nature. In this context, Saito and Rehmsmeier [2015] have argued that Precision-Recall curves are more

informative. These curves display the values of

$$\text{Precision}(p_c) = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

against the values of

$$\text{Recall}(p_c) = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}\,.$$

as the cut-off probability p_c varies from 0 to 1. Precision quantifies the number of correct positive predictions out all positive predictions made; and Recall quantifies the number of correct positive predictions out of all positive predictions that could have been made. Both focus on the Positives class (the minority class) and are unconcerned with the Negatives (the majority class). The Precision-Recall curve of a skillful model bows towards the point with coordinates $(1, 1)$. The curve of a no-skill classifier will be a horizontal line on the plot with a y-coordinate proportional to the number of Positives in the dataset. For a balanced dataset this will be 0.5 [Brownlee, 2020].

(iii) The Average Precision score is an approximation to the area under the Precision-Recall curve [Su et al., 2015]. A perfect prediction model would have an Average Precision score equal to 1, and the closer the score is to 1, the better the prediction performance of the model.

5.2 Quality of real-time predictions

Quality of real-time predictions was assessed on both the Sentinelles 1985-2019 series and on simulated data. For purpose of comparison, prediction probabilities were also estimated using a standard logistic regression model, with Y_1 and Y_2 as covariates and response variable coded as 1 in the presence of an exceedance and 0 otherwise. Levels used in this section are those defined in Table 3 (Section 4.2).

Assessment on the 1985-2019 ILI data A leave-one-out procedure was performed on the 1985–2019 epidemics yielding 35 estimates of prediction probabilities for the two lower levels ($\kappa = 0.5, 0.75$). Prediction probabilities for the two higher levels could not be estimated since there was only one exceedance for $\kappa = 0.95$ and none for $\kappa = 1$.

Figure 4 shows the prediction probabilities of level exceedances stratified according to whether an exceedance was observed or not. Contrary to GP prediction, logistic regression was never able to discriminate between the two outcomes. Table 5 presents the corresponding standardized Brier scores and confirms that GP prediction performs much better than the logistic regression.

Precision-Recall curves and Average Precision scores are not shown since they were uninformative due to insufficient data.

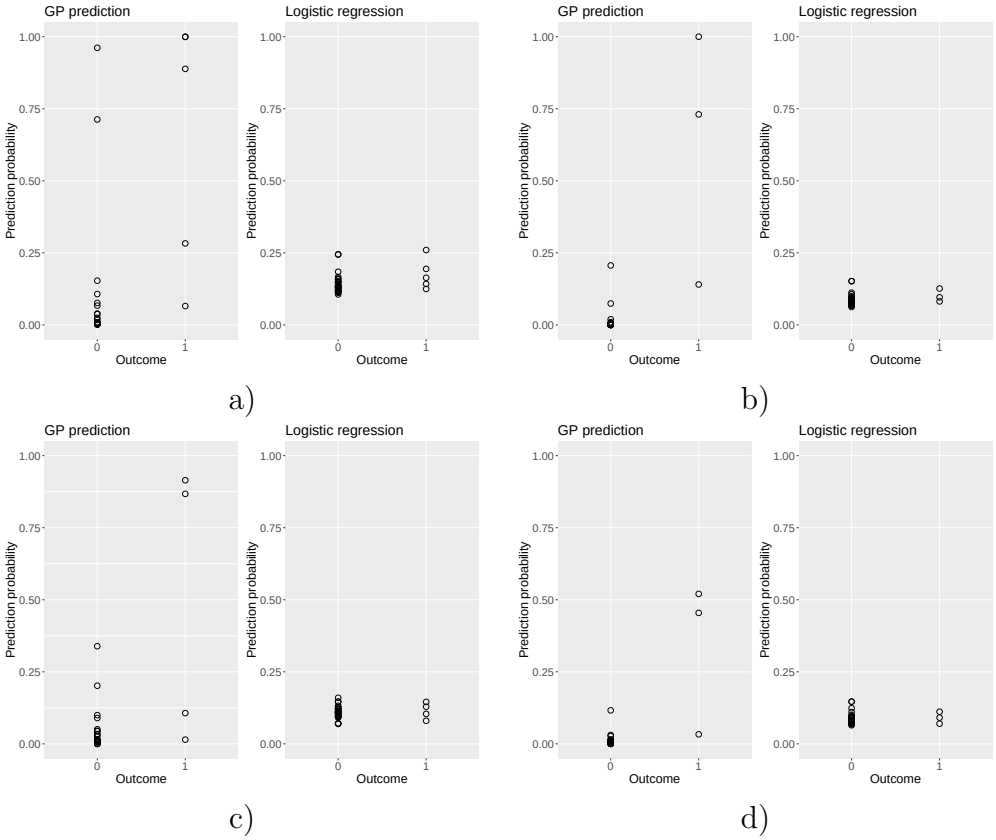


Figure 4: Prediction probabilities of level exceedances for the 1985–2019 ILI incidence rates obtained from the Gumbel model and from logistic regression using a leave-one-out procedure. Outcome is 0 if there was no exceedance and 1 otherwise. a) Week 3 ILI incidence rates, Level = 816 ($\kappa = 0.5$) b) Week 3 ILI incidence rates, Level = 1,224 ($\kappa = 0.75$) c) Epidemic sizes, Level = 4,031 ($\kappa = 0.5$) d) Epidemic sizes, Level = 6,046 ($\kappa = 0.75$)

Table 5: Standardized Brier scores derived from a leave-one-out procedure on observed data for the GP prediction and the logistic regression for predictions of exceedances of 1,729$\times\kappa$ for Week 3 ILI incidence rates and 7,241$\times\kappa$ for epidemic sizes.

	Week 3 ILI incidence rates		Epidemic sizes	
κ	0.5	0.75	0.5	0.75
Level	816	1,224	4,031	6,046
GP prediction	0.33	0.69	0.44	0.46
Logistic	0.06	0.02	0.005	0.002

Assessment on simulated data 1,500 datasets consisting of 33 three-dimensional vectors were simulated from the estimated Gumbel models for Week 3 ILI incidence rates

and epidemic sizes, respectively. The simulations were carried out following Section 7 of [Rootzén et al., 2018]. A Gumbel model was fitted to the first 32 vectors of each dataset and the estimated model was then used to predict the third component of the 33rd vector, conditionally on the first two components.

Figure 5 shows boxplots of the prediction probabilities for the GP prediction for Week 3 ILI incidence rates and epidemic sizes for the four levels ($\kappa = 0.5, 0.75, 0.95, 1$). The widths of the boxes indicate that the performance of the GP prediction is better for incidence rates than for epidemic sizes.

The boxplots for predictions obtained from the logistic regression are shown in Figure 6 for the two lower levels. Predictions could not be made for the higher levels since, for these levels, there were too few exceedances in the simulated data to allow estimation of the parameters. The quality of the GP prediction is much better than that of the logistic regression.

Standardized Brier scores and Average Precision scores for predictions with GP and logistic predictions are presented in Table 6, and Precision-Recall curves are shown in Figure 7.

For the simulated data, the true parameters for the Gumbel model are known. The prediction probabilities obtained from the true model give results similar to the fitted Gumbel model (Figure 3, Supplementary Material).

Table 6: Standardized Brier scores and Average Precision scores for the GP prediction, the logistic regression and the true model for predictions of exceedances of $1{,}729\times\kappa$ for Week 3 ILI incidence rates and $7{,}241\times\kappa$ for epidemic sizes for the simulated data.

	κ	0.5	0.75	0.95	1
	Level	816	1,224	1,551	1,632
Brier scores	GP prediction	0.72	0.75	0.80	0.84
	Logistic	0.19	-0.18	-	-
Average Precision scores	GP prediction	0.92	0.91	0.93	0.96
	Logistic	0.72	0.51	-	-

a) Week 3 ILI incidence rates

	κ	0.5	0.75	0.95	1
	Level	3,620	5,431	6,879	7,241
Brier scores	GP prediction	0.40	0.51	0.47	0.44
	Logistic	-0.03	-0.50	-	-
Average Precision scores	GP prediction	0.64	0.71	0.64	0.60
	Logistic	0.52	0.40	-	-

b) Epidemic sizes

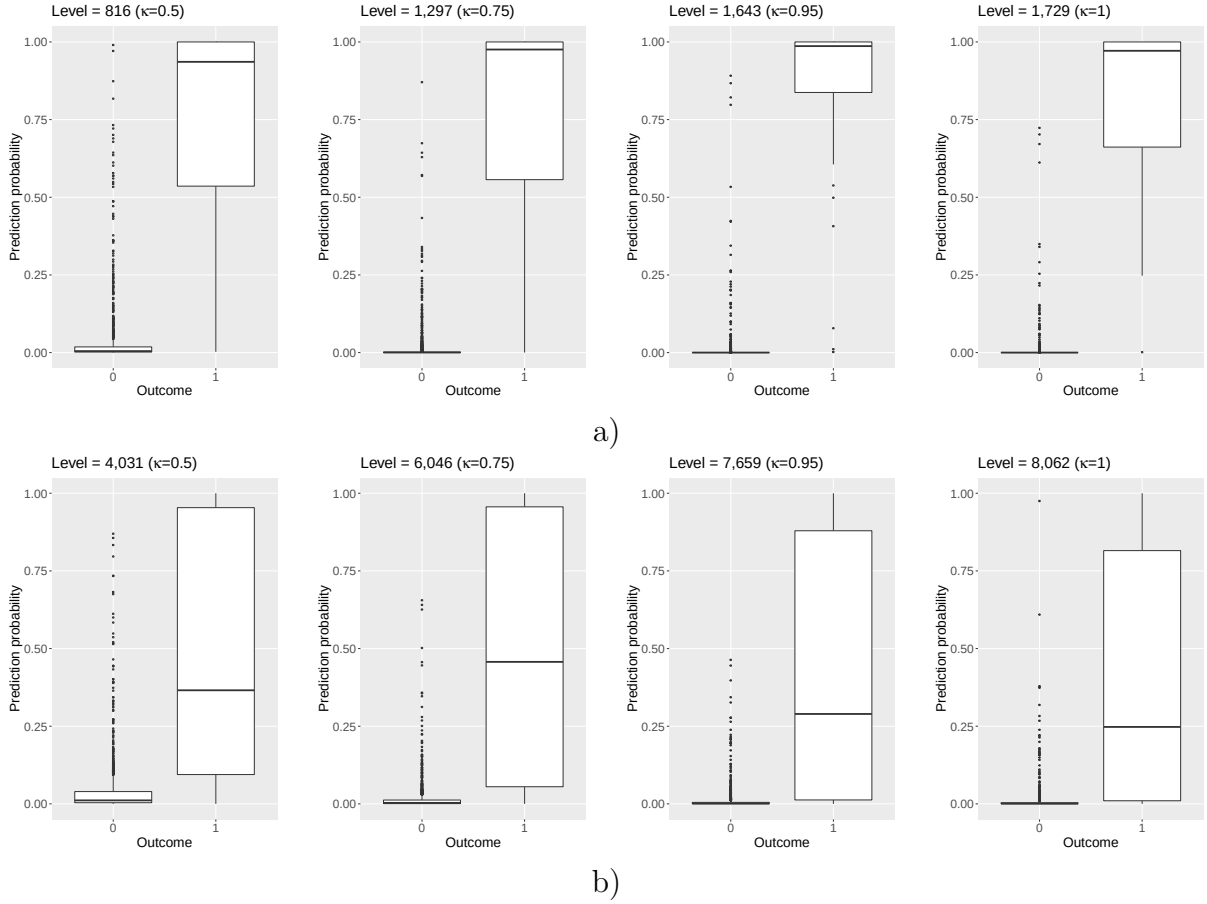


Figure 5: Boxplots of prediction probabilities of level exceedances for simulated data obtained from the fitted Gumbel model for a) Week 3 ILI incidence rates and b) epidemic sizes. Outcome is equal to 0 if there was no exceedance and 1 otherwise.

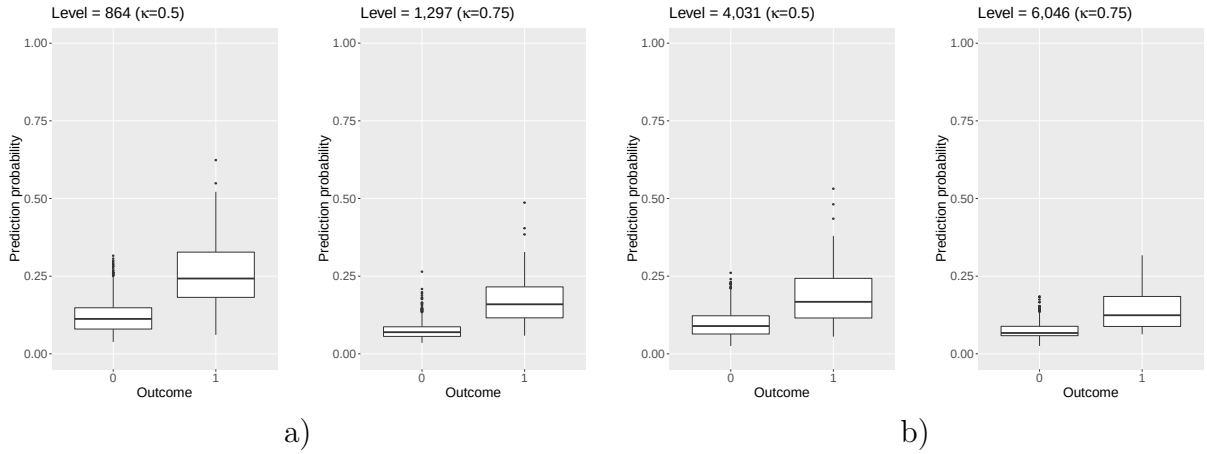
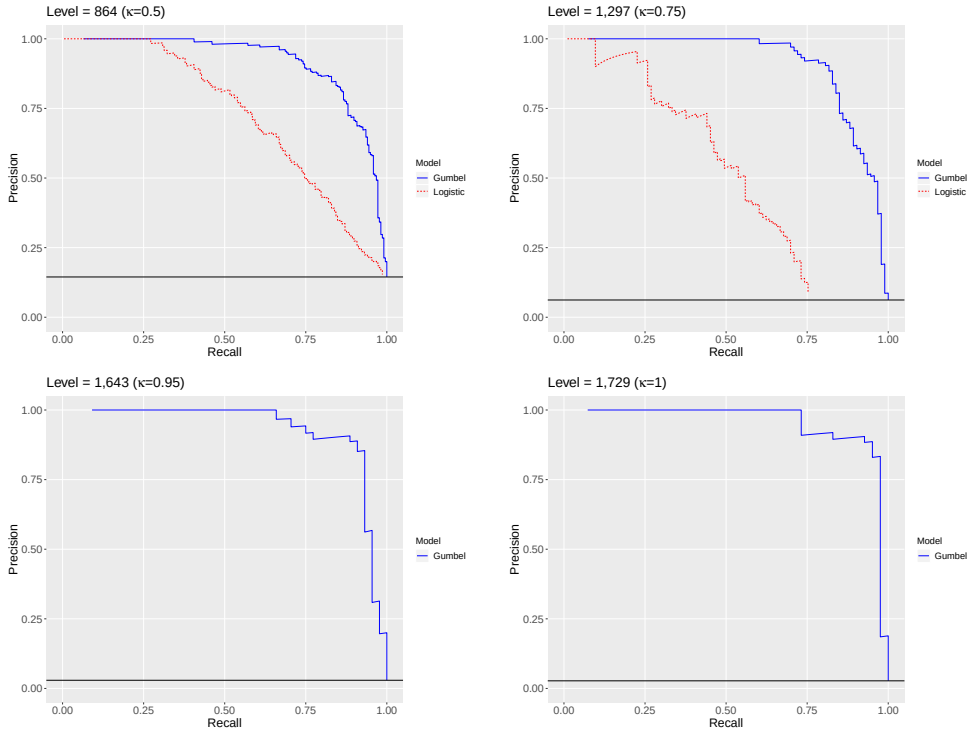
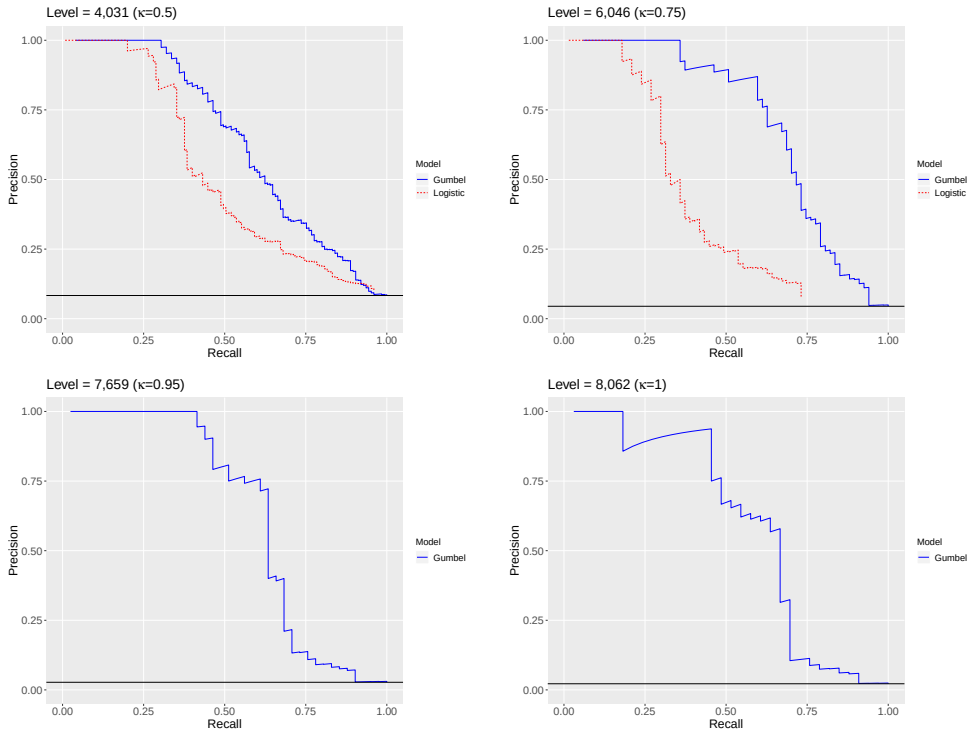


Figure 6: Boxplots of prediction probabilities of level exceedances for simulated data obtained from the logistic regression for a) Week 3 ILI incidence rates and b) epidemic sizes. Outcome is equal to 0 if there was no exceedance and 1 otherwise.



a)



b)

Figure 7: Precision-Recall curves for prediction of a) the Week 3 ILI incidence rates and epidemic sizes for the Gumbel model (plain blue line), for the logistic regression (dashed red line).

6 Conclusion

Using standard univariate EVS, we derived estimates of the risk of very high ILI incidence rates and of epidemic sizes for the following years. Such estimates provide input for long term resource planning, standard in areas such as environmental science, but have so far been little used in epidemiology and health care planning.

The main contribution of this paper was to develop methods for real-time prediction of short term risks as early as the start of an ongoing epidemic. These methods build on recent results based on multivariate GP distributions [Rootzén et al., 2018, Kiriliouk et al., 2019].

The choice of 3-dimensional GP models and the assumption of standard exponential margins were imposed by the small number of available data. However, these restrictions may be relaxed in other contexts if enough data is available.

To a large extent, assessment of predictions of extreme events remains an open issue. Here, our predictions were assessed using standardized Brier scores, Precision-Recall curves and Average Precision scores [Steyerberg et al., 2010, Brownlee, 2020, Saito and Rehmsmeier, 2015].

Acknowledgment: We thank Tom Britton, Anna Kiriliouk, Andreas Pettersson, Thordis Torainsdottir and Jenny Wadsworth for help and comments.

R codes: The data and the `R` codes are publicly available at `github.com/maudmhthomas/predict_extremeinfluenza`. Numerical optimizations used the R-function `optim` with the 1-dimensional integrals in the likelihoods calculated by the `R`-package `pracma`.

References

M. Biggerstaff, D. Alper, M. Dredze, S. Fox, I. C.-H. Fung, K. S. Hickmann, B. Lewis, R. Rosenfeld, J. Shaman, M.-H. Tsou, et al. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC infectious diseases*, 16(1):357, 2016.

J. Bresee and F. Hayden. Epidemic influenza-responding to the expected but unpredictable. *The New England Journal of Medecine*, 368(7):589–92, 2013.

E. Brodin and H. Rootzén. Univariate and bivariate GPD methods for predicting extreme wind storm losses. *Insurance: Mathematics and Economics*, 44(3):345–356, 2009.

J. Brownlee. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.

J. Chen, X. Lei, L. Zhang, and B. Peng. Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang, China. *PloS one*, 10(2):e0118521, 2015.

M. Chiapino, S. Clémençon, V. Feuillard, and A. Sabourin. A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, pages 1–22, 2019.

S. Coles. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

R. A. Davis, P. Zang, and T. Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.

P. Embrechts, C. Klüpperberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer Verlag, Berlin, 1997.

N. Goix. *Machine learning and extremes for anomaly detection*. PhD thesis, Paris, ENST, 2016.

A. Guillou, M. Kratz, and Y. L. Strat. An extreme value theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of Salmonella. *Statistics in medicine*, 33(28):5015–5027, 2014.

R. W. Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304, 2002.

A. S. Khan and N. Lurie. Health security in 2014: building on preparedness knowledge for emerging health threats. *The Lancet*, 384(9937):93–97, 2014.

A. Kiriliouk, H. Rootzén, J. Segers, and J. L. Wadsworth. Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135, 2019.

S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106–127, 2017.

R. Michel. Parametric estimation procedures in multivariate generalized Pareto models. *Scandinavian journal of statistics*, 36(1):60–75, 2009.

A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, 2008.

B. Renard, K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, L. Neppel, K. Najib, J. Carreau, P. Arnaud, Y. Aubert, et al. Data-based comparison of frequency analysis methods: A general framework. *Water Resources Research*, 49(2):825–843, 2013.

Réseau Sentinelles. Inserm/Sorbonne Université. `https://www.sentiweb.fr`, 2019.

J. Root, J. Qian, and V. Saligrama. Learning efficient anomaly detectors from k-nn graphs. In *Artificial Intelligence and Statistics*, pages 790–799, 2015.

H. Rootzén, J. Segers, and J. L. Wadsworth. Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145, 2018.

T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3): e0118432, 2015.

R. E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494, 1963.

R. L. Smith. Threshold methods for sample extremes. In *Statistical extremes and applications*, pages 621–638. Springer, 1984.

E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

W. Su, Y. Yuan, and M. Zhu. A relationship between the average precision and the area under the roc curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 349–352, 2015.

A. Thomas, S. Clemençon, A. Gramfort, and A. Sabourin. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In *AISTATS*, pages 1011–1019, 2017.

M. Thomas, M. Lemaitre, M. L. Wilson, C. Viboud, Y. Yordanov, H. Wackernagel, and F. Carrat. Applications of extreme value theory in public health. *PloS one*, 11(7): e0159312, 2016.