A statistical assessment of influenza intensity thresholds from the moving epidemic and WHO methods

Johannes Bracher[1,2], Jonas M. Littek[1]

9 June 2023

[1]Karlsruhe Institute of Technology (KIT), Chair of Statistics and Econometrics
[2]Heidelberg Institute for Theoretical Studies (HITS), Computational Statistics Group

This manuscript is is a preprint and has not been subject to peer review.

Abstract

Intensity thresholds serve to make influenza activity comparable across countries and surveillance systems. The moving epidemic method (MEM) and the WHO method are widely used to classify season peaks as low, medium, high or very high. In both methods, thresholds correspond to quantiles of a normal distribution fitted to a set of potentially transformed historical observations. However, there are differences in how this set is constructed. We assess the impact of these choices analytically and in simulation studies. We find that under the default settings of both the MEM and WHO method, on average more season peaks than intended are classified as high or very high. Combining characteristics of both, better-calibrated thresholds can be achieved. Even these, however, must be expected to have rather modest positive predictive values and should be interpreted with care.

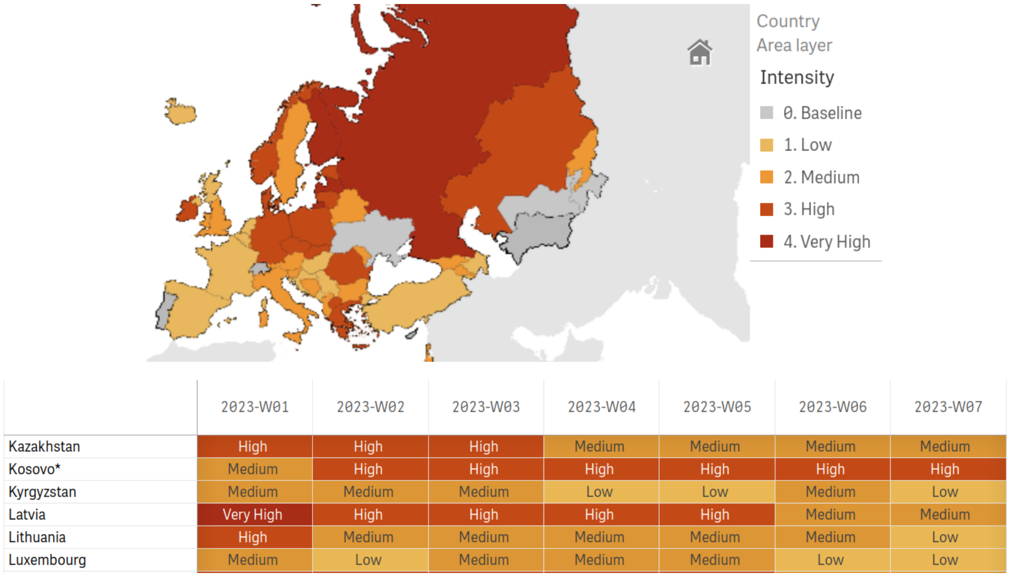
Author's summary

Intensity thresholds are a common tool to make influenza activity comparable across different countries and surveillance systems. The moving epidemic method (MEM) and the WHO method are widely used to this end and enable classification of season peak intensity as low, medium, high or very high. The two approaches are similar in that they base thresholds on quantiles of a normal distribution fitted to a reference set of historical observations. However, three methodological differences exist. Firstly, in the MEM the normal distribution is fitted to log-transformed incidence data, while the WHO method by default operates on the original scale. Secondly, the MEM uses more than one observation from each past season, fixing the total number to include into the reference set. The WHO method uses only the highest value from each season. Lastly, in the WHO method, but not the MEM, historical data are by default smoothed prior to the computation of thresholds. We assess the impact of these choices on thresholds both analytically and in a simulation study. The latter is based on re-sampling of influenza-like illness (ILI) data from France and the US, thus reflecting temperate climate settings. We find that when the normal distribution is fitted to untransformed observations, a rather large proportion of new season peaks are classified as high or very high intensity. This can be mitigated by a logarithmic transformation. When fixing the total number of past observations included in the reference set as in the MEM, thresholds increase in expectation the more seasons are available. When only few are available, there is a high chance of classifying new season peaks as high or very high intensity. Smoothing incidence time series prior to computing thresholds results in somewhat less variable estimators, but also a lowering of thresholds. If these are applied to unsmoothed new season peaks, there will again be a large proportion classified as high or very high intensity. If they are applied to smoothed new data, this problem is avoided. In terms of sensitivity and positive predictive values of thresholds, we find that these cannot be expected to exceed rather modest levels if the number of available historical seasons is low; thresholds for very high intensity are particularly affected by this problem. Our practical recommendation is to include one observation per season into the reference set and employ a log transformation in the computation of thresholds. Smoothing can be applied to somewhat reduce the variability of thresholds. However, threholds then need to be applied to smoothed rather than raw new peak values, which slightly modifies their interpretation.

Keywords: calibration, influenza, intensity threshold, moving epidemic method, re-sampling, sensitivity, WHO method.

1 Introduction

Influenza intensity thresholds are a common tool to make influenza activity comparable across different countries and surveillance systems. Based on data from past influenza seasons, they provide a classification into *low*, *medium*, *high* or *very high* intensity. While thresholds are designed to assess the season peak intensity, they are also used to monitor spatial and temporal patterns over the course of a season. To this end, intensity classifications are commonly summarized in heat charts and maps; see Figure 1 for an example from the *Flu News Europe* platform run by the World Health Organization (WHO) and the European Center for Disease Prevention and Control (ECDC). Intensity thresholds thus play an important role in assessing the influenza situation at an international level.



	2023-W01	2023-W02	2023-W03	2023-W04	2023-W05	2023-W06	2023-W07
Kazakhstan	High	High	High	Medium	Medium	Medium	Medium
Kosovo*	Medium	High	High	High	High	High	High
Kyrgyzstan	Medium	Medium	Medium	Low	Low	Medium	Low
Latvia	Very High	High	High	High	High	Medium	Medium
Lithuania	High	Medium	Medium	Medium	Medium	Medium	Low
Luxembourg	Medium	Low	Medium	Medium	Medium	Low	Low

Figure 1: Influenza intensity map (top, week 01/2023) and heat chart (bottom, weeks 01–07/2023) as published by *Flu News Europe* (`https://flunewseurope.org/SeasonOverview`). Copyright: World Health Organization 2023 and European Centre for Disease Prevention and Control 2023. Reproduction is authorised, provided the source is acknowledged

Following the 2009 influenza H1N1 pandemic, the *Review Committee on the Functioning of the International Health Regulations and on Pandemic Influenza (H1N1)* recommended that member states apply and evaluate their methods for severity assessment every year, thus improving their pandemic preparedness (WHO, 2011, p.118). In the subsequent *WHO Pandemic Infuenza Severity Assessment (PISA)* guideline (WHO, 2017), two statistical methods were recommended to determine influenza intensity thresholds. These are the so-called *WHO method* (WHO, 2014) and the *moving epidemic method* (MEM; Vega et al. 2015) which has also been recommended by the European Centre for Disease Prevention and Control (e.g., ECDC 2017). As will be detailed in Section 3, the WHO and MEM approaches bear important similarities and can be seen as variations of the same general approach.

Especially the moving epidemic method has been adopted by many national public health agencies, and a large number of articles documenting practical applications exist (see Section 4). The theoretical statistical properties of the MEM and WHO methods, however, have not yet been studied in detail. In the present work, we aim to close this gap and derive some recommendations for the computation of influena intensity thresholds. We obtain some analytical results, complemented by simulation experiments based on re-sampling of real-world data from France and the US. These indicate that using the recommended default settings, both the moving epidemic and the WHO methods tend to produce too low thresholds, leading to a higher number of season peaks classified as high or very high intensity than intended. The best behaviour of thresholds is achieved by combining characteristics of the MEM and WHO methods. Even the recommended configuration, however, is characterized by rather modest sensitivity and positive predictive values unless data on many historical seasons are available. This concerns in particular the threshold for very high intensity and should be kept in mind when using both the MEM and WHO method.

The remainder of the article is structured as follows. Section 2 lists several desirable properties of a threshold setting procedure. Section 3 provides definitions of the MEM and WHO methods, highlighting three differences between these otherwise similar approaches. Section 4 consists of an overview of published applications of the MEM and WHO methods. This is followed by an examination of statistical properties of the two methods in Section 5. In Section 6 we conduct the simulation study before concluding with a discussion in Section 7. Here we also provide practical recommendations on the different implementation choices.

2 Desirable properties of a thresholding method

We start by stating properties of a thresholding method for influenza activity which we consider desirable. These will guide the analyses in the remainder of the paper.

Calibration. As stated in the *WHO Pandemic Influenza Severity Assessment* guideline, the rationale for both the MEM and the WHO method is that

> *"about 50–60% of the season peaks should be above the moderate threshold,* ±10% *above the high threshold and* ±2.5% *above the extraordinary threshold"* (WHO, 2017, p.10).

This is a purely statistical definition, and threshold setting corresponds to estimating quantiles of an underlying distribution of season peak values. A major requirement for the thresholds is thus *calibration.* We call a method calibrated if in the long run, the right fractions of seasons are classified as low, medium, high and very high. If, for instance, a method flags new season peaks as very high intensity in considerably more than 2.5% of the cases, this will hamper the usefulness of thresholds.

Sensitivity and specificity. In addition to classifying the right number of seasons into the various categories, the individual classifications should be correct. A season peak exceeding the very high/high/medium threshold should indeed be among the 2.5%/10%/60% most extreme season peaks according to the underlying true distribution. Conversely, a season peak which is not actually among the 2.5%/10%/60% most extreme ones should not be flagged as such. As a complementary perspective to sensitivity and specificity, confusion matrices can be considered to assess how peaks which are truly very high, high, medium or low intensity are classified.

Stability. Estimated thresholds should not be overly variable. Ideally they would not fluctuate strongly around the corresponding quantiles of the true distribution of season peaks.

Simplicity. To ensure broad practical applicability, understanding the thresholding method should not require advanced statistical training. The method should be simple enough that a quantitatively literate user can develop a good intuition of its functioning and parameters.

Ease of practical application. Besides conceptual simplicity, ease of application is central. Methods should be straightforward to apply using well-documented and ideally open-source software packages.

We note that while the overarching goal of the *WHO PISA* guidelines is to enhance preparedness for *pandemic influenza*, the intended statistical properties of the MEM and WHO methods refer to *seasonal influenza waves.* The rationale is that a good understanding of the range of peak intensities of seasonal influenza is the basis for a meaningful assessment of pandemic influenza outbreaks. Our reasoning in the following therefore focuses on this setting of subsequent waves of seasonal influenza. Note that we only address the question of peak intensity thresholds, but not season onset / baseline thresholds (Vega et al., 2013). Also, we do not address the question of determining peak intensity in real time, i.e., before the season peak has been reached or before it is clear whether a peak has been reached. Rather than simple thresholds, this would require control charts (see e.g., Liu et al. 2019), which we consider outside the scope of the present article.

3 Definition of the moving epidemic and WHO methods

While framed slightly differently in the respective documentations, the MEM and WHO methods can be seen as two special cases of the same general approach. We assume that thresholds are based on weekly data (typically incidences) from m past seasons and applied to the $(m + 1)$-th season. Implicitly it is further assumed that each influenza season consists of just one wave and does not feature multiple peaks separated by longer time spans. This is typically the case in temperate, but not necessarily in tropical regions. Thresholds are are then obtained as follows.

1. **Smoothing of historical data** (optional): apply an l-week moving average to all historical seasons. If data are smoothed, we denote by $x_{j,k}^{\text{raw}}$ the raw observation from season $j = 1, \dots, m$, week $k = 1, \dots, 52$, and by
$$x_{j,k}^{\text{smo}} = \frac{1}{l} \sum_{d=0}^{l-1} x_{j,k-d}^{\text{raw}}, \quad k = l, \dots, 52$$
the smoothed version. In the remainder of this description we denote whichever of $x_{j,k}^{\text{raw}}$ and $x_{j,k}^{\text{smo}}$ is used to compute thresholds by $x_{j,k}$.
2. **Sorting:** Within each historical season $j = 1, \dots, m$ order all observations $x_{j,k}$ in decreasing order, denoting the i-th largest observation from season j by $x_j^{(i)}$.
3. **Selection of reference set:** Select the n largest observations from each of the m past seasons to construct a reference set $\mathcal{X} = \{x_j^{(i)} : j = 1, \dots, m; i = 1, \dots, n\}$.
4. **Data transformation:** Apply a monotonically increasing transformation $y_j^{(i)} = f(x_j^{(i)})$ to all members of the reference set $\mathcal{X}$ to obtain a reference set $\mathcal{Y}$ of transformed historical observations.
5. **Fitting a normal distribution:** Assume that the transformed values in $\mathcal{Y}$ come from a normal distribution and compute estimates of its mean and standard deviation,
$$\begin{aligned} \bar{y} &= \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} y_j^{(i)}}{nm} \\ s &= \sqrt{\frac{\sum_{j=1}^{m} \sum_{i=1}^{n} (y_j^{(i)} - \bar{y})}{nm - 1}}. \end{aligned} \tag{1}$$
6. **Computation of thresholds:** Define intensity thresholds on the transformed scale as quantiles of the normal distribution $\text{N}(\bar{y}, s^2)$, i.e. compute
$$\hat{q}_{Y,\alpha} = \bar{y} + z_\alpha s, \tag{2}$$
where z_α is the α quantile of the standard normal distribution. These can be interpreted as estimates of quantiles $q_{Y,\alpha}$ of an underlying theoretical distribution. A common choice is
    - the 40th percentile $\hat{q}_{Y,0.4} = \bar{y} - 0.25s$ as the threshold for medium intensity;
    - the 90th percentile $\hat{q}_{Y,0.9} = \bar{y} + 1.28s$ as the threshold for high intensity;
    - the 97.5th percentile $\hat{q}_{Y,0.975} = \bar{y} + 1.96s$ as the threshold for very high intensity.
7. **Transformation of thresholds to original scale:** Obtain corresponding thresholds on the original scale by applying the inverse transformation, i.e. setting
$$\hat{q}_{X,\alpha} = f^{-1}(\hat{q}_{Y,\alpha})$$
for $\alpha = 0.4, 0.9, 0.975$.
8. **Application of thresholds:** The obtained thresholds are applied to classify the season peak value of the next season. Depending on the exact specification, thresholds can be applied either to the raw peak value $x_{m+1}^{\text{raw},(1)}$ or the smoothed version $x_{m+1}^{\text{smo},(1)}$.

The MEM and WHO methods are special cases of this procedure, see also overview in Table 1. In the MEM method, no smoothing is applied and the default data transformation f is the natural logarithm. Vega et al. (2015) recommend to use $5 \leq m \leq 10$ seasons to ensure a recent data basis. The number of observations included per season is set to $n = 30/m$, rounded to the nearest integer (with a minimum value of $n = 1$). The total number of historical observations is thus kept approximately fixed. We note that this description refers to the specification from Vega et al. (2015), which is also reflected in the default settings of the accompanying R package `mem` (Lozano, 2020). The package, however, permits the user to choose n, m, and f (`i.type.intensity = 5` for no transformation, `i.type.intensity = 6` for the log transformation), thus offering considerable flexibility. We note that while `mem` does not currently allow for data smoothing using a moving average, alternatives like LOESS (locally estimated scatterplot smoothing) are available.

The implementation of the WHO method in the publicly available *WHO Average Curves* application (WHO, 2023) likewise offers a lot of flexibility; our description is based on the default settings as of May 2023, as well as the description in WHO (2014). Smoothing of data prior to the computation of thresholds is recommended, with a default of $l = 3$ (adapted from a recommendation of $l = 4$ in WHO 2014, p68). Subsequently, $n = 1$ observation per season is used and by default no transformation is applied to the reference set. If peak incidences vary strongly across seasons, a log transformation is recommended. At least three historical seasons are required to compute thresholds, but it is noted that the "accuracy of these thresholds should be expected to increase with the number of seasons of good quality data available" (WHO, 2023, p22). Thresholds are by default applied to unsmoothed new season peaks.

Both the inclusion of multiple observations per season in the MEM ($n > 1$) and the smoothing of data in the WHO method ($l > 1$) can be seen as attempts to base estimation of thresholds on more data than just one peak value per historical season. This is intended to make estimation more stable. The impact of these strategies on the resulting thresholds will be discussed in Section 5.

Table 1: Default settings of the moving epidemic and WHO methods.

	moving epidemic method	WHO Method
smoothing of historical data	none	moving average, $l = 3$
number n of observations used per season	$n = \max[\text{round}(30/m), 1]$	$n = 1$
default transformation for reference set	natural logarithm	none
recommended number m of historical seasons	$5 \leq m \leq 10$	$m \geq 3$, more recommended
smoothing of new season peak	none	none

4 Review of recent use cases

To improve our understanding of the different settings in which the MEM and WHO methods are applied in practice we performed a literature search of articles published in English language and citing the papers Vega et al. (2015), WHO (2014) and WHO (2017) until December 2020 (identified via *CrossRef* and *Google Scholar*). The results are summarized in Table 2. As can be seen from the large number of entries from the years 2019 and 2020, the MEM has quickly become a standard approach in the determination of intensity thresholds for influenza and other respiratory diseases. The articles come from numerous countries and in many cases have been co-authored by representatives of national or regional public health agencies. In most analyses, threshold levels at the 40th, 90th and 97.5th percentile as in Section 3 are used. Variability with respect to the number m of historical seasons included is considerable, with a range from 3 to 16 seasons. Consequently, the number n of observations included per season ranges from two to ten (none of the above papers indicated a modification of the default setting $n = 30/m$ of the moving epidemic method). We only found three published applications of the WHO method, one of them providing a comparison to the thresholds from the MEM method (Rguig et al., 2020).

5 Analytical results on statistical properties

Both the MEM and WHO method essentially serve to estimate quantiles of the distribution of season peak values $(q_{X,0.4}, q_{X,0.9}, q_{X,0.975})$ from historical data. In the following we obtain some analytical results on how the estimators $(\hat{q}_{X,0.4}, \hat{q}_{X,0.9}, \hat{q}_{X,0.975})$ behave under different variations of the procedure outlined in Section 3. In all cases we will simplifyingly assume that the available historical seasons are independent realizations of the same random process, i.e., we ignore time trends, correlations between neighbouring seasons and other changes in seasonal patterns. To highlight that we treat all incidence values as random rather than observed variables in this section, we will denote them by capital rather than lowecase letters (e.g., $X_j^{(i)}$ rather than $x_j^{(i)}$). All derivations are provided in Supplement A.

5.1 Choice of number n of observations used per season

We first consider the impact of the number n of observations used per historical season. We assume that no smoothing is applied in Step 1 of the algorithm described in Section 3. Denote by $\mathbf{Y}_j$ the random

Table 2: Applications of the MEM and WHO method to determine intensity thresholds for respiratory diseases. We did not include works where only baseline thresholds are computed. The number of seasons included to compute thresholds is denoted by m, the number of observations used per season by n. The "Percentiles" column indicates which percentiles were used for the medium, high and very high thresholds, with "?" indicating that no explicit information was found. Abbreviations: SARI = severe acute respiratory infection; ILI = influenza-like illness; RSV = respiratory syncytial virus.

(a) Moving epidemic method

Region	Disease	Years covered	m	n	Percentiles	Authors
Australia	ILI/influenza	2012–2017	5	6	40, 90, 99	Vette et al. (2018)
Australia, Chile, New Zealand, South Africa	ILI/SARI	2013–2019	6	5	40, 90, 97.5	Sullivan et al. (2019)
Catalonia	ILI	2010–2016	5	6	?	Basile et al. (2018)
Catalonia	ILI	2005–2018	12	3	?	Basile et al. (2019)
Catalonia	ILI/influenza	2010–2017	7	4	?	Torner et al. (2019)
Egypt	SARI/ILI	2010–2017	6	5	40, 90, 97.5	AbdElGawad et al. (2020)
Egypt	SARI	2013–2015	3	10	?	Elhakim et al. (2019)
England	ILI	2010–2016	6	5	?	Wagner et al. (2018)
Finland	influenza	2011–2016	5	6	?	Pesälä et al. (2019)
Montenegro	ILI	2010–2018	7	4	40, 90, 97.5	Rakocevic et al. (2019)
Morocco	ILI	2005–2017	11	3	40, 90, 97.5	Rguig et al. (2020)
Netherlands	RSV	2005–2017	12	3	40, 90, 97.5	Vos et al. (2019)
Norway	influenza	2006–2015	9	3	?	Benedetti et al. (2019)
Pakistan	ILI, SARI	2008–2017	10	3	40, 90, 97.5	Nisar et al. (2020)
Portugal	ILI	2012–2017	5	6	40, 90, 97.5	Páscoa et al. (2018)
Scotland	influenza	2010–2018	7	4	?	Murray et al. (2018)
Scotland	influenza	2010–2019	7–8	4	40, 90, 97.5	Dickson et al. (2020)
Slovenia	RSV	2008–2018	10	3	40, 90, 97.5	Grilc et al. (2021)
Spain (17 regions)	ILI	2003–2015	4–10	3–8	40, 90, 97.5	Bangert et al. (2017)
Spain	ILI	2001–2018	16	2	40, 90, 97.5	Redondo-Bravo et al. (2020)
Tunisia	ILI	2009-2018	9	3	50, 90, 95	Bouguerra et al. (2020)
United Kingdom	ILI	2000–2013	10	3	40, 90, 97.5	Green et al. (2015)
United Kingdom	ILI/RSV	2011–2018	4–6	5–8	40, 90, 97.5	Harcourt et al. (2019)
USA	ILI/influenza	2003–2015	11	3	50, 90, 98	Biggerstaff et al. (2017)
USA	ILI	2010–2015	5	6	50, 90, 98	Dahlgren et al. (2018)
USA	influenza	2010–2016	6	5	50, 90, 98	Dahlgren et al. (2019)

(b) WHO method

Region	Disease	years covered	m	n	percentiles	authors
Cambodia	ILI	2009–2015	7	1	mean, 90, 95	Ly et al. (2017)
Morocco	ILI	2005–2017	11	1	40, 90, 97.5	Rguig et al. (2020)
Philippines	ILI	2006–2012	7	1	90	Lucero et al. (2016)
Victoria/Australia	ILI	2002–2011	6–10	1	90, 95	Tay et al. (2013)

vector of the n largest transformed incidence values from season j in decreasing order, i.e.

$$\mathbf{Y}_j = (Y_j^{(1)}, \ldots, Y_j^{(n)})^\top.$$

We assume that $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are identically and independently distributed, denoting their theoretical mean and covariance matrix of by

$$\mathbb{E}\left(\mathbf{Y}_j\right) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \quad \text{and} \quad \operatorname{Cov}\left(\mathbf{Y}_j\right) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{1,1} & \cdots & \sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \sigma_{n,1} & \cdots & \sigma_{n,n} \end{pmatrix}. \tag{3}$$

To make notation more intuitive we also write $\sigma_i^2 = \sigma_{i,i}$. Construction of thresholds is based on the reference set $\mathcal{Y}$, which pools the elements of the vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. It can be shown that the expectations of $\bar{Y}$ and S^2 as defined in equation (1) are given by

$$\mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mu_i, \tag{4}$$

$$\mathbb{E}(S^2) = \frac{m}{mn-1}\sum_{i=1}^{n}(\sigma_i^2 + \mu_i^2) \; - \; \frac{1}{n(mn-1)}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sigma_{i,i'} \; - \; \frac{m}{n(mn-1)}\left(\sum_{i=1}^{n}\mu_i\right)^2, \tag{5}$$

respectively. Moving to the resulting thresholds,

$$\mathbb{E}(\hat{q}_{Y,\alpha}) \approx \mathbb{E}(\bar{Y}) + z_\alpha\sqrt{\mathbb{E}(S^2)} \tag{6}$$

$$\mathbb{E}(\hat{q}_{X,\alpha}) \approx f^{-1}\left\{\mathbb{E}(\bar{Y}) + z_\alpha\sqrt{\mathbb{E}(S^2)}\right\} \tag{7}$$

usually holds in good approximation in our applied setting if the transformation function f is the identity or the natural logarithm; see Appendix A.1 for details. It can further be shown that for $n = 1$ the approximation becomes

$$\mathbb{E}(\hat{q}_{Y,\alpha}) \approx \mu_1 + z_\alpha\sigma_1 = q_{Y,\alpha},$$

where the second equality holds only if the transformed peak values $Y_j^{(1)}$ indeed come from a normal distribution. This is a desirable property as our estimator of the relevant quantile is approximately unbiased. In expectation, thresholds on the transformed scale are thus such that the nominal threshold exceedance probabilities (60%/10%/2.5%) are achieved. If a different value $n > 1$ is chosen, this will generally no longer be the case. Equations (4)–(6) tell us by how much $\hat{q}_{Y,\alpha}$ can then be expected to differ from $q_{Y,\alpha}$. By the definition of μ_i as the expectation of the i-th largest observation in a given season, $\mathbb{E}(\bar{Y})$ decreases in n. While the expected thresholds also depend on $\mathbf{\Sigma}$, this downward bias will usually translate to the $\hat{q}_{Y,\alpha}$. As a consequence, when choosing $n > 1$, one must expect to classify a larger number of season peaks as high or very high intensity. Intuitively speaking, extending the reference by observations which are close to peaks, but not actually peaks, dilutes the reference set and pulls thresholds downward. When choosing $n = 30/m$ as suggested for the MEM, thresholds will tend to increase the more years of data are used to compute thresholds, and the probability of exceeding thresholds will decrease (approaching the nominal values from above).

5.2 Smoothing of time series prior to computing of thresholds

Next we assess the role of smoothing historical data prior to computing thresholds. As derivations get tedious otherwise, we only consider the case where f is the identity function, i.e., no transformation is applied to the reference set and thus $Y_j^{(i)} = X_j^{(i)}$. Also, as smoothing and multiple observations per season are generally not used in parallel we assume $n = 1$.

Now denote by p_j the peak week of the smoothed incidence in season j such that $X_j^{(1)} = X_{j,p_j}^{\text{smo}}$; and by

$$\mathbf{X}_j^{\text{raw}} = (X_j^{\text{raw},(1)}, \dots, X_j^{\text{raw},(l)})^\top$$

the random vector containing the raw observations $X_{j,p_j-l+1}^{\text{raw}}, X_{j,p_j-l+2}^{\text{raw}}, \dots, X_{j,p_j}^{\text{raw}}$ in decreasing order. The vector $\mathbf{X}_j^{\text{raw}}$ is thus a re-ordering of the raw observations making up the smoothed peak value. The theoretical mean and covariance matrix of $\mathbf{X}_j^{\text{raw}}$ are denoted by

$$\mathbb{E}\left(\mathbf{X}_j^{\text{raw}}\right) = \boldsymbol{\mu}^{\text{raw}} = \begin{pmatrix} \mu_1^{\text{raw}} \\ \vdots \\ \mu_n^{\text{raw}} \end{pmatrix} \quad \text{and} \quad \text{Cov}\left(\mathbf{X}_j^{\text{raw}}\right) = \mathbf{\Sigma}^{\text{raw}} = \begin{pmatrix} \sigma_{1,1}^{\text{raw}} & \cdots & \sigma_{1,n}^{\text{raw}} \\ \vdots & \ddots & \vdots \\ \sigma_{n,1}^{\text{raw}} & \cdots & \sigma_{n,n}^{\text{raw}} \end{pmatrix}, \tag{8}$$

respectively. Note that by construction we have $\mu_1^{\text{raw}} \geq \mu_2^{\text{raw}} \geq \cdots > \mu_l^{\text{raw}}$. As we assumed f to be the identity and $n = 1$, the reference set $\mathcal{Y}$ consists simply of $Y_1^{(1)}, \dots, Y_m^{(1)}$ with

$$Y_j^{(1)} = X_{j,p_j}^{\text{smo}} = \frac{1}{l}\sum_{d=1}^{l} X_j^{\text{raw},(i)},$$

i.e., the average values of $\mathbf{X}_1^{\text{raw}}, \dots, \mathbf{X}_m^{\text{raw}}$. It is straightforward to show that in this case the expectations of $\bar{Y}$ and S^2 are

$$\mathbb{E}(\bar{Y}) = \frac{1}{l}\sum_{i=1}^{l}\mu_i^{\text{raw}}, \quad \mathbb{E}(S^2) = \frac{1}{l^2}\sum_{i=1}^{l}\sum_{i'=1}^{l}\sigma_{i',i}^{\text{raw}}. \tag{9}$$

These can be plugged into equation (6) to approximate the expected thresholds. Similarly to the previous section, for $l = 1$ we obtain

$$\mathbb{E}(\hat{q}_{Y,\alpha}) \approx \mathbb{E}(\bar{Y}) + z_\alpha \sqrt{\mathbb{E}(S^2)} \approx \mu_1^{\text{raw}} + z_\alpha \sigma_1^{\text{raw}}. \tag{10}$$

Now assume that the peak of the raw time series is always part of the vector $\mathbf{X}_j^{\text{raw}}$ and thus has mean μ_1^{raw} and standard deviation $\sigma_1^{2,\text{raw}}$. This will typically be the case if a season presents one well-shaped peak. If in addition the normality assumption holds, then equation (10) implies that the threshold $\hat{q}_\alpha$ is again approximately unbiased. For $l > 1$ we get by construction $\mathbb{E}(\bar{Y}) \leq \mu_1^{\text{raw}}$ and usually also $\mathbb{E}(\sqrt{S^2}) \leq \sigma_1^{\text{raw}}$. In practice both are likely to hold in strict inequality. The high and very high thresholds will thus be lower than for $l = 1$ and exceeded more frequently by unsmoothed new peaks. If thresholds are applied to new smoothed peak values, though, the nominal exceedance probabilities will be achieved in expectation as the mean and variance of the smoothed peak values correspond to the expressions from equation (9).

As in Section 5.1, the key aspect is that the reference set need to be comparable to new peaks; both smoothed peak values and values which are close to, but not actually peaks are systematically different from new unsmoothed peak values. They are thus not suitable as reference values, and the resulting thresholds will be biased. While we did not adapt our derivation to the case where data are log-transformed prior to computing thresholds, our simulation results in Section 6.3.2 confirm that the described qualitative patterns likewise hold.

5.3 Sensitivity and specificity under normality and $n = 1$

While in the previous sections we focused on the expected thresholds, we now turn to the resulting sensitivity and specificity. An implicit assumption of the MEM and WHO methods is that the observations in the reference set $\mathcal{Y}$ as well as the new transformed season peak $Y_{m+1}^{(1)}$ are identically and independently normally distributed; as discussed in Section 5.1, this assumption is implausible if $n > 1$, so we focus on the case where $n = 1$ observation is used per historical season and thus $\mathcal{Y} = \{Y_1^{(1)}, Y_2^{(1)}, \dots, Y_m^{(1)}\}$. The normality assumption is then

$$Y_1^{(1)}, \dots, Y_m^{(1)}, Y_{m+1}^{(1)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_1, \sigma_1^2).$$

Defining the threshold $\hat{q}_{Y,\alpha}$ via equations (1) and (2), it can be shown that

$$\hat{q}_{Y,\alpha} \overset{\text{approx}}{\sim} \mathrm{N}\left[\mu_1 + z_\alpha \sigma_1,\ \ \sigma_1^2 \times \left(\frac{1}{m} + \frac{z_\alpha^2}{2(m-1)}\right)\right]. \tag{11}$$

Based on this we can compute the sensitivity and specificity of our thresholding procedure at level α. The sensitivity describes the probability that given

$$Y_{m+1}^{(1)} \geq q_{Y,\alpha} = \mu_1 + z_\alpha \sigma_1,$$

i.e., the new transformed peak is in theory among the $(1-\alpha) \times 100\%$ highest peaks, it will also be classified as such, i.e.,

$$Y_{m+1}^{(1)} \geq \hat{q}_{Y,\alpha}.$$

Note that we here treat both $Y_{m+1}^{(1)}$ and $\hat{q}_{Y,\alpha}$ as random. It can be shown that the probability in question can be approximated by

$$\text{sens}_\alpha = \Pr(Y_{m+1}^{(1)} > \hat{q}_{Y,\alpha} \ \mid\ Y_{m+1}^{(1)} > q_{Y,\alpha}) \approx \int_{z_\alpha}^{\infty} \phi(x) \times \Phi\left(\frac{x - z_\alpha}{\sqrt{\frac{1}{m} + \frac{z_\alpha^2}{2(m-1)}}}\right) \mathrm{d}x, \tag{12}$$

where ϕ and Φ are the standard normal density function and cumulative density function, respectively. We note that this expression only depends on α, but not μ_1 and σ_1^2. Similarly, the specificity can be approximated by

$$\text{spec}_\alpha = \Pr(Y_{m+1}^{(1)} < \hat{q}_\alpha \ \mid\ Y_{m+1}^{(1)} < q_{Y,\alpha}) \approx \int_{-\infty}^{z_\alpha} \phi(x) \times \left[1 - \Phi\left(\frac{x - z_\alpha}{\sqrt{\frac{1}{m} + \frac{z_\alpha^2}{2(m-1)}}}\right)\right] \mathrm{d}x. \tag{13}$$

The positive predictive value can be computed using the common formula

$$\text{PPV}_\alpha = \Pr(Y_{m+1}^{(1)} > q_{Y,\alpha} \ \mid Y_{m+1}^{(1)} > \hat{q}_\alpha) = \frac{(1-\alpha) \times \text{sens}_\alpha}{(1-\alpha) \times \text{sens}_\alpha + \alpha \times (1 - \text{spec}_\alpha)}. \tag{14}$$

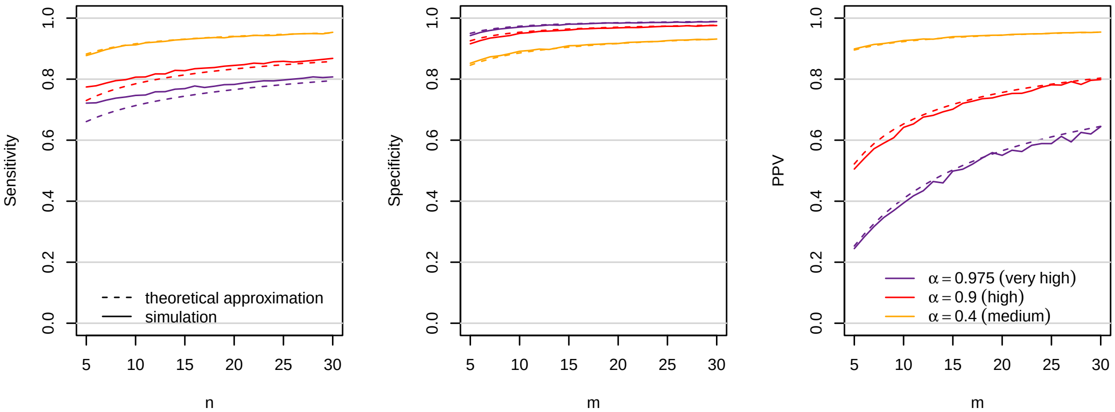


Figure 2: Sensitivity, specificity and positive predictive values under normality assumption, by number m of historical seasons used. Solid lines show a theoretical approximation, dashed lines simulation results.

Neither of these expressions has any simpler closed form, but they are straightforward to evaluate numerically. We visualize the resulting curves in Figure 2. As the approximations are derived under the assumption of large m while in our setting m is typically small, we also show simulation-based versions. The simulations indicate that the theoretical sensitivity for small m is slightly higher than in the approximation, but otherwise agreement is quite close. Little surprisingly, the sensitivity, specificity and positive predictive values increase in m. The sensitivity and PPV are lowest for the very high threshold at $\alpha = 0.975$. For the most practically relevant values of $5 \leq m \leq 10$, the PPV is only between 0.25 and 0.4 in this case. More than half of the seasons flagged as very high intensity will thus be false positives. For the high threshold ($\alpha = 0.9$), the respective PPVs are between 0.5 and 0.6. Even if the underlying normality assumptions are fulfilled, there are thus natural limits to the classification performance, which for small m and high α is at relatively modest levels. We note that this general theoretical result is the same irrespective of whether smoothing has been applied to all peaks in question or not. While smoothing will usually make thresholds less variable, any effect on sensitivity and specificity is canceled out by the fact that smoothed new peaks are likewise less variable.

5.4 Basing thresholds on confidence intervals

In previous works (WHO 2014; Vega et al. 2015), the thresholds as defined in Section 3 have been described as upper ends of one-sided confidence intervals for the arithmetic (WHO method) or geometric mean (MEM) of the reference observations. This, however, is imprecise terminology as in the computations the standard deviation s rather than the standard error $s/\sqrt{nm}$ is used (see documentation of the `mem` package and WHO 2014, p.69). The intervals thus correspond to prediction rather than confidence intervals. The use of actual confidence intervals is also possible in the `mem` package (`i.type.intensity = 1` for the geometric, `i.type.intensity = 2` for the arithmetic mean). Under this specification, equation (6) becomes

$$\mathbb{E}(q_{Y,\alpha}) \approx \mathbb{E}(\bar{Y}) + \frac{z_\alpha}{\sqrt{nm}}\sqrt{\mathbb{E}(S^2)}.$$

This is not a desirable property as given enough historical observations (large mn), thresholds for all levels α will converge to $\mathbb{E}(\bar{Y})$. Each new season will then be classified either as low or as very high.

6 Simulation study

6.1 Simulation setup

The analytical results from Section 5 involve some simplifying approximations and the assumption that the observations in the reference set $\mathcal{Y}$ are indeed normally distributed. In the following we study the respective aspects empirically in a simulation study. In order to realistically mimic the seasonal patterns of influenza, we re-sample historical surveillance data rather than generating fully synthetic data. Assume M seasons of historical data on a measure of influenza activity are available. We then repeat the following steps 500 times:

- Sample a sequence of 15 seasons from the M available seasons. This is done with equal probability for each season and *with replacement*, meaning that the same season can appear more than once. This approach is called the *seasonal block bootstrap* (Politis, 2001).
- For each value $m = 5, \dots, 15$:
  - Restrict the generated sequence to the first m seasons.
  - Compute thresholds for medium ($\alpha = 0.4$), high ($\alpha = 0.9$) and very high intensity ($\alpha = 0.975$). This is done using a number of variations of the thresholding procedure, see below.
  - Evaluate which of the M historical season peaks would be classified as low, moderate, high and very high.
  - Compute relevant summary statistics including mean thresholds, exceedance probabilities, sensitivities, specificities and positive predictive values.

The range $m = 5, \dots, 15$ is motivated by the values found in real-world applications, see overview in Section 4. We apply a total of eight variations of the general approach described in Section 3 to the generated data.

(a) no smoothing, no transformation of the reference set, $n = 1$. This corresponds to the WHO method, but without the optional smoothing step.

(c) no smoothing, logarithmic transformation of the reference set, $n = 1$.

(b) no smoothing, no transformation of the reference set, $n = 30/m$.

(d) no smoothing, logarithmic transformation of the reference set, $n = 30/m$. This corresponds to the default MEM approach.

To study the impact of smoothing we apply the following settings:

(e) smoothing with $l = 3$ (alternatively $l = 7$), log transformation of the reference set, $n = 1$, thresholds applied to unsmoothed new peaks. This corresponds to the WHO method with the optional log transformation applied.

(f) same as (g), but thresholds applied to smoothed new peaks.

Finally, to assess the behaviour if confidence rather than prediction intervals are used we apply

(g) no smoothing, log transformation of the reference set, $n = 1$, using confidence rather than prediction intervals.

(h) same as (g), but with $n = 30/m$.

All analyses were performed using the R language for statistical computing (R Core Team, 2020) and the package `mem` (Lozano, 2020).

6.2 Data

For our re-sampling scheme we use publicly available data on the weekly incidence of influenza-like illness per 100,000 inhabitants in France, 1986–2019, as published by Réseau Sentinelles (INSERM/Sorbonne Université, `https://www.sentiweb.fr`, Flahault et al. 2006). These are available both at the national and at a regional level. To make statements about the sensitivity and specificity of the "very high" categorization, we require a larger set of historical data than the available 35 seasons. We therefore pool curves from the 12 continental French administrative regions. As the overall level of ILI incidences varies considerably across regions, we scale all data such that the average season peak per region is 100. As the data from the Corse island region differ substantially from the other regions, with overall considerably lower reported incidences, we exclude them. Moreover, all curves from the pandemic 2009/2010 season were removed. In total we then dispose of 396 historical seasons.

Figure 3 shows an illustration of the re-scaled data from two regions (Grand Est and Nouvelle Aquitaine), along with four descriptive plots. In the bottom left panel we show boxplots of the first through sixth largest observation per season. Not surprisingly, values on average get smaller for increasing ranks. It is noteworthy that they also get less dispersed, meaning that variability among e.g. the sixth largest observations per season is smaller than among the peak values. The next panel shows the distribution of peak values without smoothing ($l = 1$) and with smoothing windows of $l = 3$ and $l = 7$. It is clearly visible that peak values get lower and less dispersed when smoothing is applied. The remaining panels show normal QQ plots of untransformed and log-transformed peak values. It can be seen that after transformation the distribution is roughly normal.

To assess the sensitivity of our simulation results to the choice of data set we re-ran all simulations using weekly weighted ILI (wILI) data from the US. These data stem from the from CDC *FluView* project (Charbonneau and James, 2019), cover the years 1998–2018, and were obtained via the CDC *FluSight* influenza forecasting platform (`https://github.com/FluSightNetwork/cdc-flusight-ensemble/`). Reported values are on the interval 0 to 1 and correspond to the fraction of general practitioner visits due to influenza-like symptoms. To increase the number of available seasons we pooled national-level data and data from the ten Health and Human Services (HHS) regions, re-scaling data to mean peak values of 100 as described for the French example above. In total we thus obtained 209 historical seasons. Results based on the US data have been moved to Supplement C and are briefly discussed in Section 6.4.

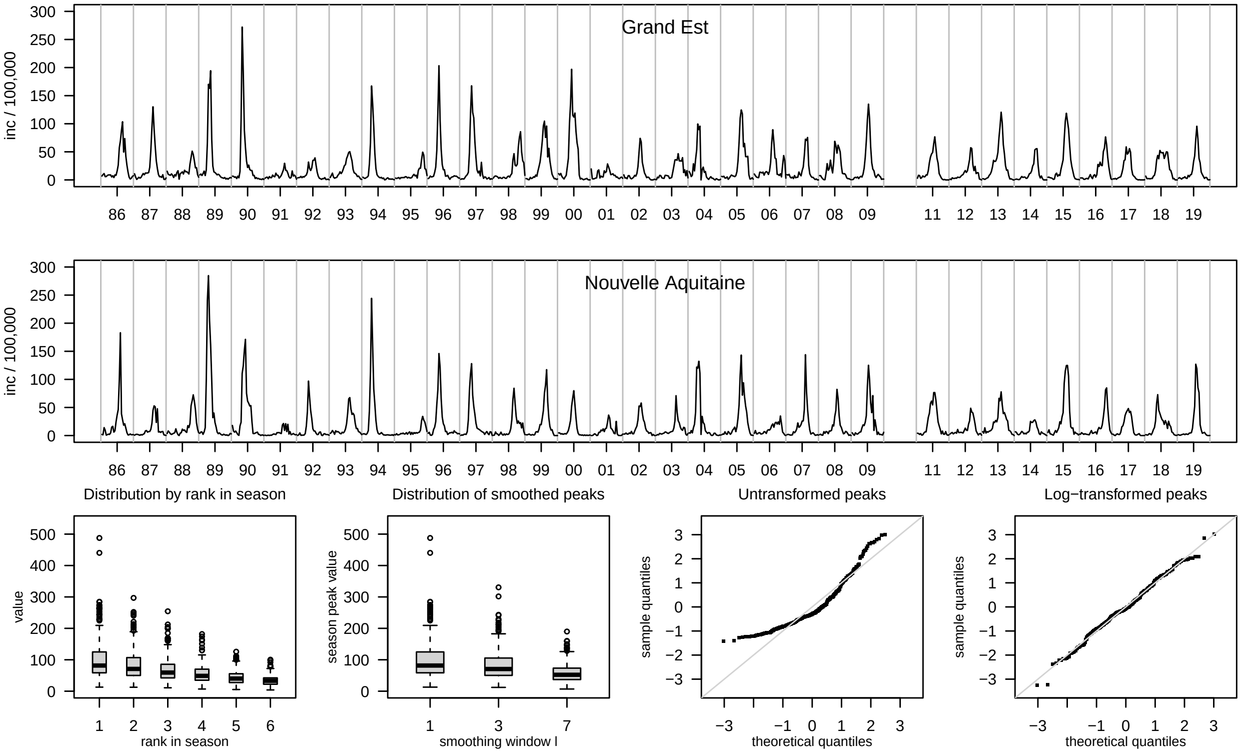


Figure 3: Time series of weekly ILI cases per 100,000 population in the French regions of Grand Est and Nouvelle Aquitaine, 1985–2019. Off-season weeks are omitted in the plot, with grey lines delimiting the different seasons. The bottom row shows descriptive plots of the distribution of season peaks. First: Boxplots of incidence values by rank within season. Second: Boxplot of smoothed peak values as a function of the smoothing window width l. Third: Normal QQ plot of untransformed peak values. Fourth: Normal QQ plot of log-transformed peak values.

6.3 Results based on French data

6.3.1 Choice of transformation f and number n of observations used per season

Figure 4 summarizes thresholds resulting from different combinations of transformation function f and number n of observations used per season (specifications a–d from Section 6.1). For each case we show mean thresholds, along with the empirical 5% and 95% quantiles, and the shares of new seasons classified into the different categories. This is complemented with summaries of the sensitivity, specificity and positive predictive value. All results are shown as a function of the number m of historical seasons used. Where applicable, analytical approximations are shown as lines. These have been computed using equations (7) (with empirical means and covariances plugged in) and (12)–(14).

Thresholds and threshold exceedance. Thresholds for high and very high intensity are higher when a log transformation is employed. This leads to better calibration, i.e., the shares of seasons exceeding the high or very high thresholds are closer to the intended levels of 10% and 2.5%. Indeed, when using $n = 1$ observation per season and $m \geq 10$ historical seasons, the thresholds based on log-transformed data have very close to nominal exceedance rates. Without this transformation, new season peaks are classified as very high in roughly 10% rather than 2.5% of the cases. This indicates that the normal assumption is more appropriate after log tansformation, an aspect that already became visible from the normal QQ plot in Figure 3.

As implied by the reasoning from Section 5.1, letting the number of observations used per season depend on the number of available seasons via $n = 30/m$ leads to average thresholds which increase in m. When using a log transformation, they increase from 180 for $m = 5, n = 6$ to 218 for $m = 10, n = 3$ and 140 for $m = 15, n = 2$. For $n = 1$, in which case thresholds can be interpreted as unbiased (Section 5.1), the average is around 270 and largely independent of m. Including historical observations which are not actual peaks thus leads to a considerable lowering of alarm thresholds and increases the number of alerts for high and very high influenza activity. For $m = 5, n = 6$ the proportion of seasons classified as very high is 15% if a log transformation is used and 24% otherwise. As can be seen from the fourth column, this is due to poor specificity, and as shown in the fifth column, leads to low positive predictive values. In the most extreme case where $m = 5$ seasons are used without a transformation, only one in ten seasons classified as very high is actually from the 2.5% most extreme seasons. We note, however, that even the most well-behaved specification with $n = 1$ and a log transformation only yields a PPV slightly above 20% for $m = 5$. For $m = 15$ this value roughly doubles. Our theoretical approximation is well aligned with the simulation results here (but much less so if no log transformation is used, as in this case the normality assumption is poorly fulfilled).

It can be noted that the thresholds generally have high variability (see the shaded bands in the first column of Figure 4). This is particularly pronounced for the very high threshold, and even more so if a logarithmic transformation is used. This reflects the general difficulty of estimating extreme quantiles from a small number of observations.

Confusion matrices. To complement these results and provide some more intuition on the intensity classifications, Figure 5 shows confusion matrices for thresholds computed with a log transformation and different values of m. This represents a more detailed breakup of the results from the second column of Figure 4. Note that the perspective differs somewhat from the computation of sensitivities and specificities. There, we focused on *threshold exceedance*, considering for instance whether the medium threshold was exceeded or not by a peak which was truly at least medium intensity. For the sensitivity of the medium threshold we thus ignored whether a peak which was truly medium intensity also exceeded the high threshold (which was in turn reflected in the specificity of the high threshold). Now we consider not just threshold exceedance, but the exact categorization. This is shown stratified by the true category of a peak, determined via the empirical distribution of all peaks (the highest 2.5% are considered very high, the next 7.5% high, the next 50% medium and the remaining 40% low).

For $n = 1$ (top two rows) it can be seen that in all categories misclassifications occur, but become somewhat less frequent for larger m. The log transformation yields better classification of truly medium and high peaks, while for truly low or very high peaks results are better without a transformation. This essentially reflects the fact that thresholds are more spaced out when using a log transformation (see Figure 4), giving medium and high peaks a better chance of correct classification. The rather modest positive predictive values from Figure 4 are here reflected by the fact that in most panels, a large part of the peaks classified as very high (purple rectangles) are actually high or medium.

For $n = 30/m$, the miscalibration issues identified previously are again visible. For $m = 5$ and with a log transformation, more than 60% of truly low peaks are classified as medium; roughly 40% of medium peaks are classified as high or very high; and close to 75% of high peaks are classified as very high. Similar pictures arise for thresholds computed on the natural scale, while as implied by theory, the problem is diminished with increasing m.

6.3.2 Smoothing of time series prior to computing thresholds

Thresholds and threshold exceedance. We next assess the impact of smoothing historical data prior to computing thresholds (specifications e–f from Section 6.1). Results for a window size of $l = 3$ and

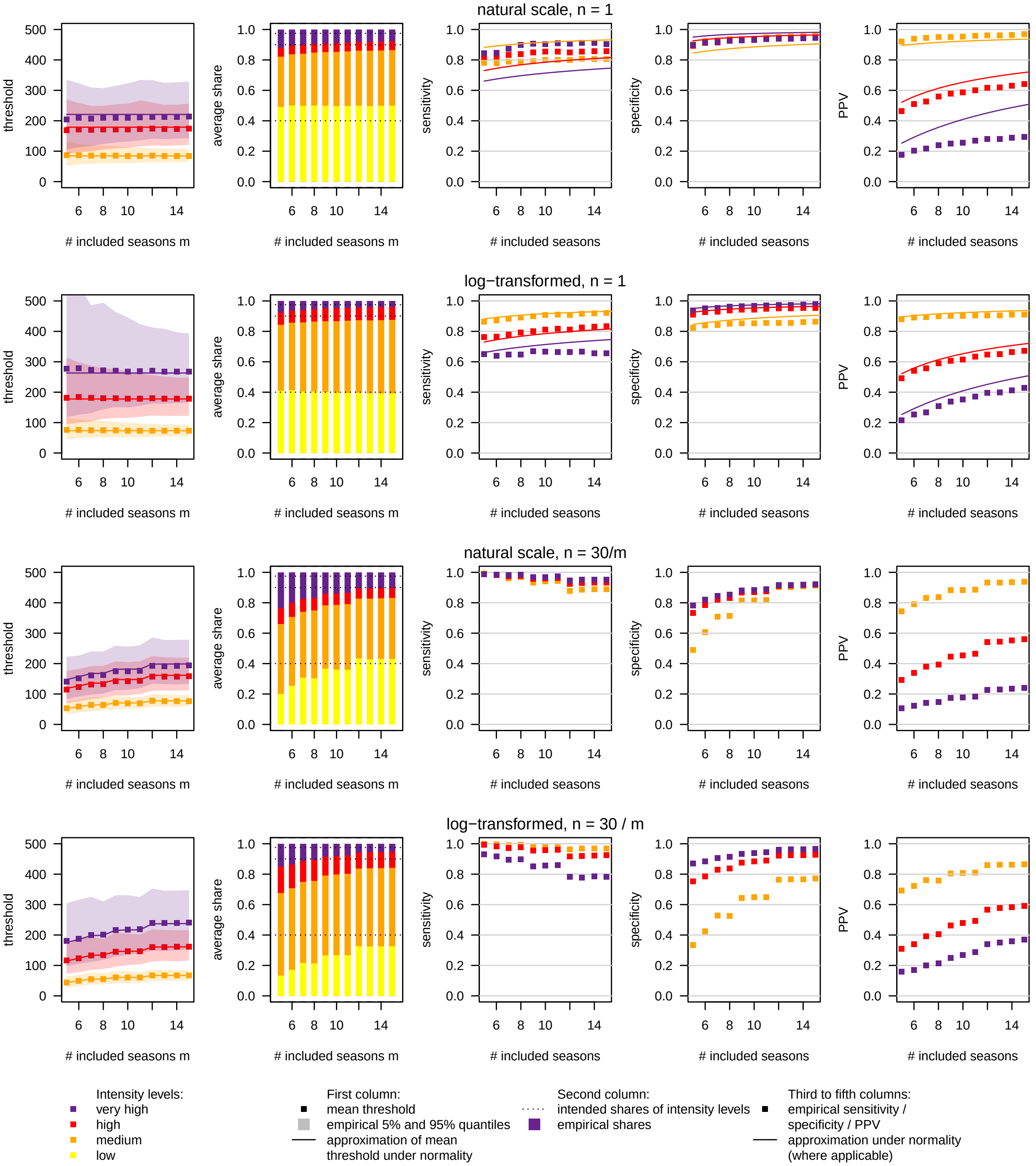


Figure 4: Impact of the choice of n and transformation function f. First column: simulation-based mean intensity thresholds (squares) along with bands delimited by the empirical 5% and 95% quantiles. Analytical approximations of mean threshold values (computed from empirical means and covariances) are displayed as lines. Second column: resulting average shares of season peaks classified as low, medium, high and very high intensity. Third to fifth columns: sensitivity, specificity and PPVs of the different thresholds. Simulation results are shown as squares. Where available analytical approximations are shown as lines.

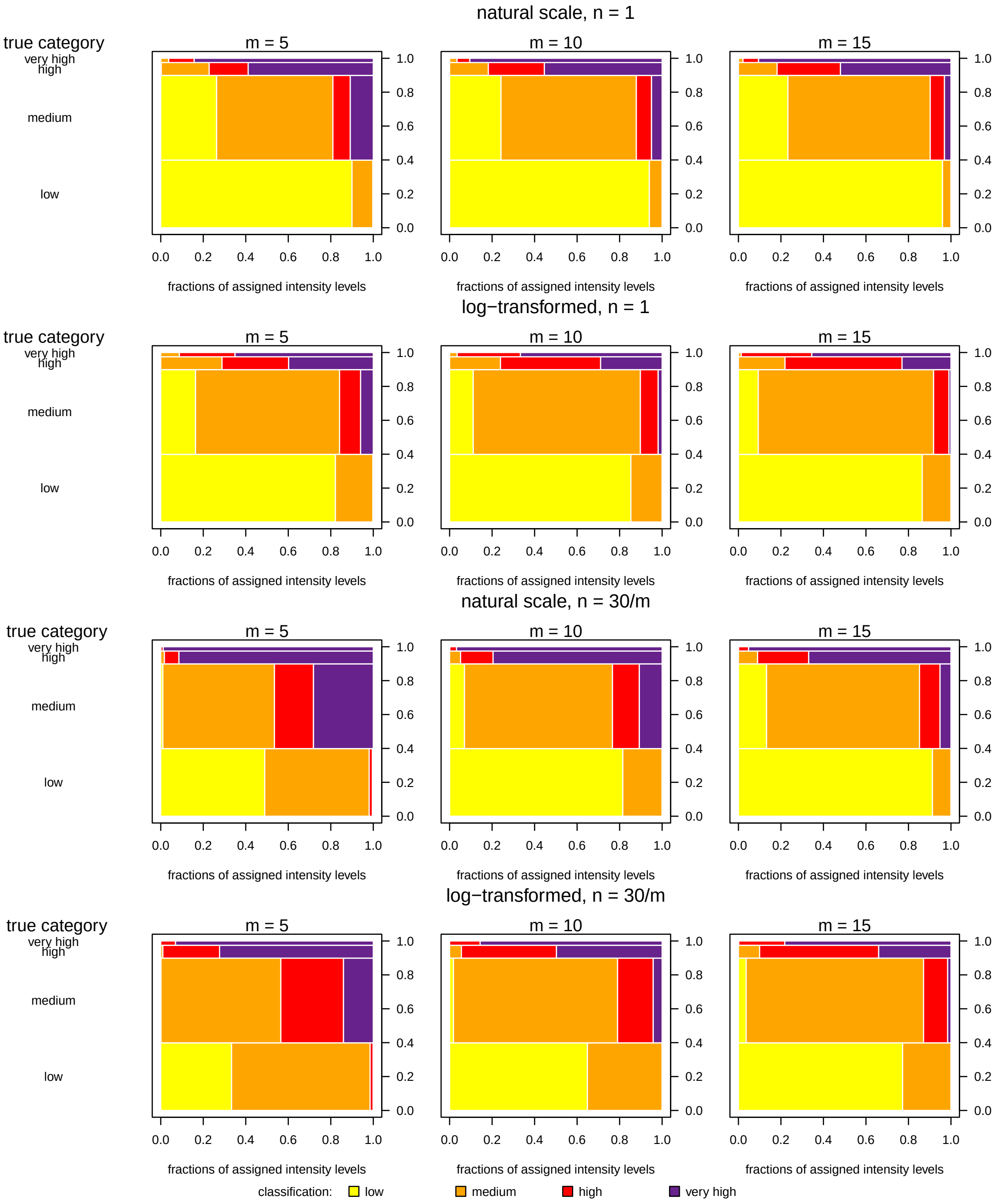


Figure 5: Confusion matrices for intensity classifications obtained with different choices of n and transformation function f. Mosaic plots show which fractions of season peaks which are truly very high, high, medium or low are classified into the four categories. The true class is determined with respect to the empirical quantiles of the distribution of peaks: very high (highest 2.5% of all peaks), high (next 7.5%), medium (next 50%), low (lowest 40% of all peaks).

including a log transformation are shown in Figure 6. The top row shows the case where thresholds are applied to unsmoothed new peak values. In accordance with the theoretical arguments from Section 5.2, the high and very high thresholds are exceeded more frequently than intended. When applying thresholds to smoothed new observations, the empirical and nominal exceedance levels are in good agreement. We note that thresholds are somewhat less variable than without smoothing, which we consider a desirable feature. As predicted by theory, sensitivity, specificity and positive predictive values are largely the same as without smoothing (compare the second rows of Figures 4 and 6). While they are slightly higher in the present case we are unsure whether this will be the case in general. We provide a display for a stronger smoothing with $l = 7$ in Supplementary Figure 9. If the resulting thresholds are applied to new unsmoothed peaks, more than a fifth of them is classified as very high intensity.

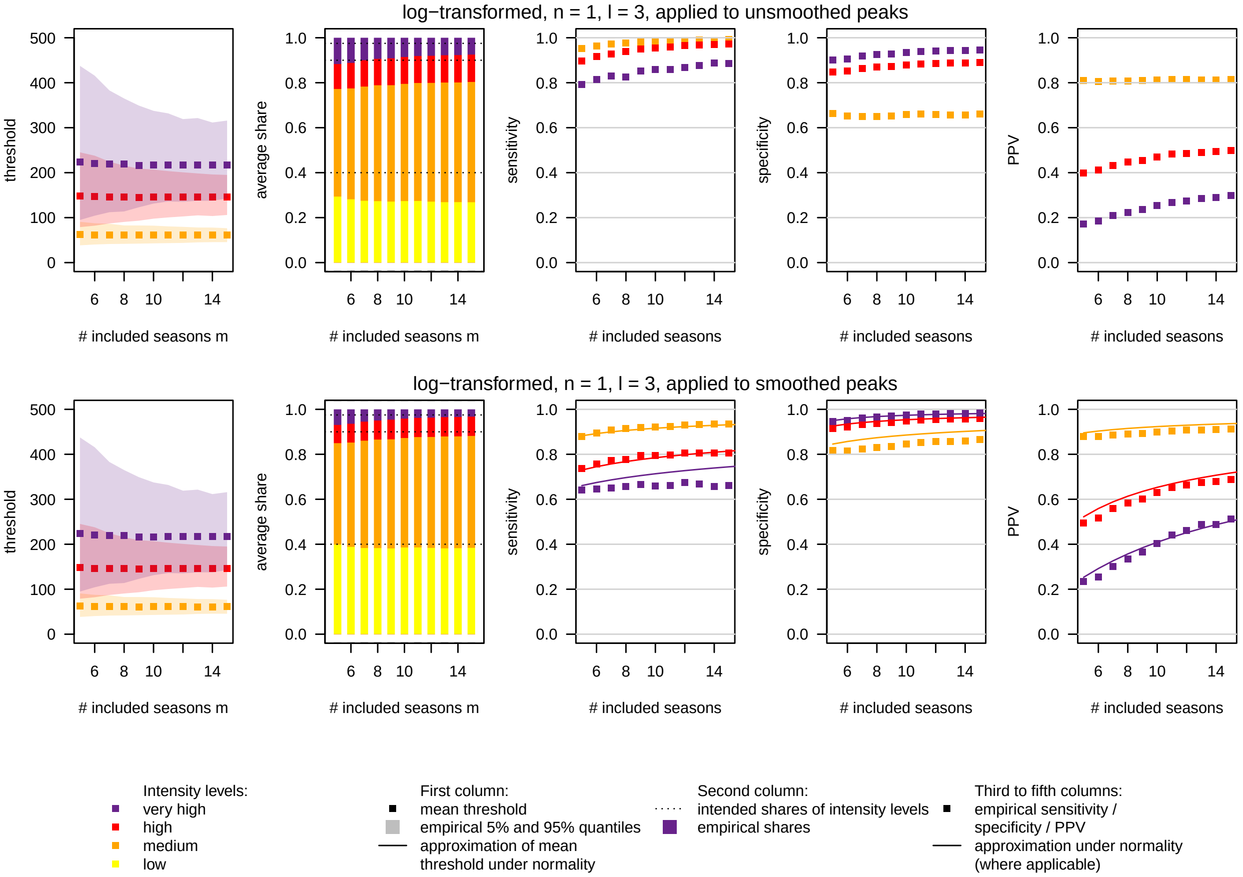


Figure 6: Impact of smoothing of historical data on thresholds. We applied a moving average with $l = 3$ to the historical time series prior to computing thresholds and subsequently applied them to either unsmoothed or smoothed new peak values. Results are shown for thresholds computed with a log transformation. See the caption of Figure 4 for details on the plot elements.

Confusion matrices. We complement this again with confusion matrices, shown in Figure 6.3.2. When unsmoothed new peaks are classified based on thresholds computed with smoothing, peaks tend to be assigned to too high categories. When thresholds are applied to smoothed new peaks, the results strongly resemble those obtained without smoothing of either historical or new peaks (compare second row of Figure 5).

6.3.3 Basing thresholds on confidence intervals

Lastly we address thresholds based on confidence intervals (specifications g–h from Section 6.1). As mentioned in Section 5.4 these are not the default in any of the considered methods. However, they are

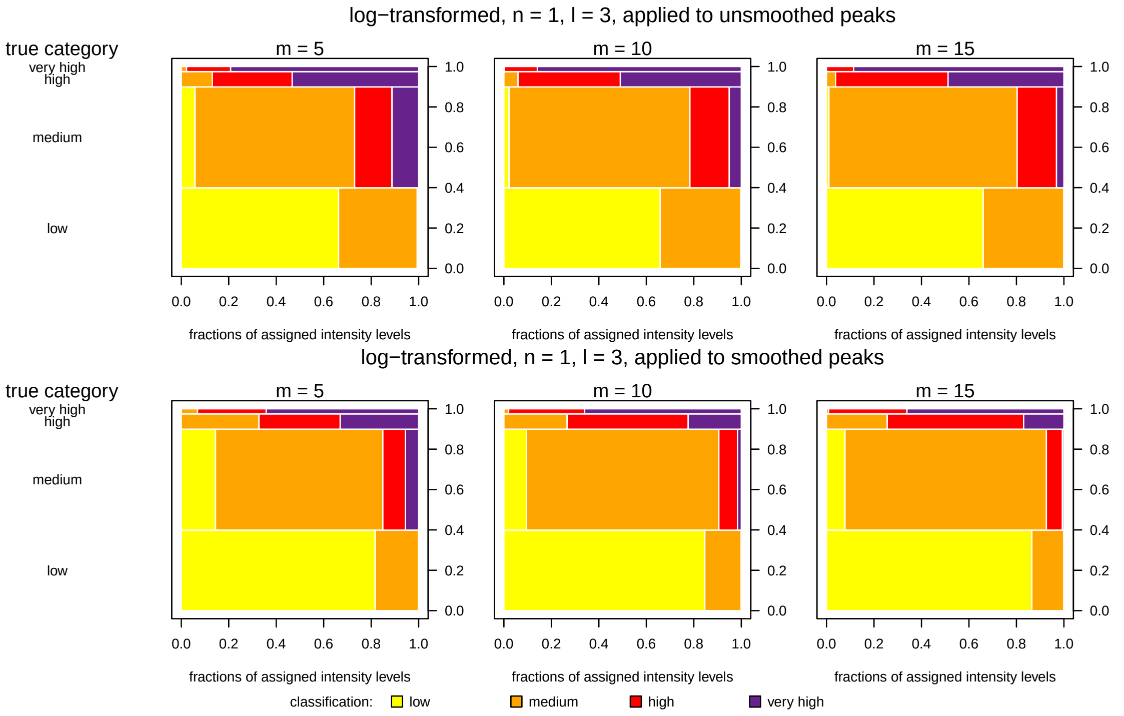


Figure 7: Confusion matrices for intensity classifications obtained with a smoothing window of width $l = 3$ and a log transformation. Top: thresholds applied to unsmoothed new peaks; bottom: thresholds applied to smoothed new peaks. Mosaic plots show which fractions of season peaks which are truly very high, high, medium or low are classified into the four categories, see camption of Figure 5 for details.

implemented in the `mem` R package and due to terminological imprecisions in previous methods descriptions may be used by analysts programming their own routines. Figure 8 shows that these thresholds are not well-behaved. As our theoretical reasoning suggested, the mean thresholds at different levels approach each other as m grows. Overall, a considerably too large number of seasons is classified as very high intensity.

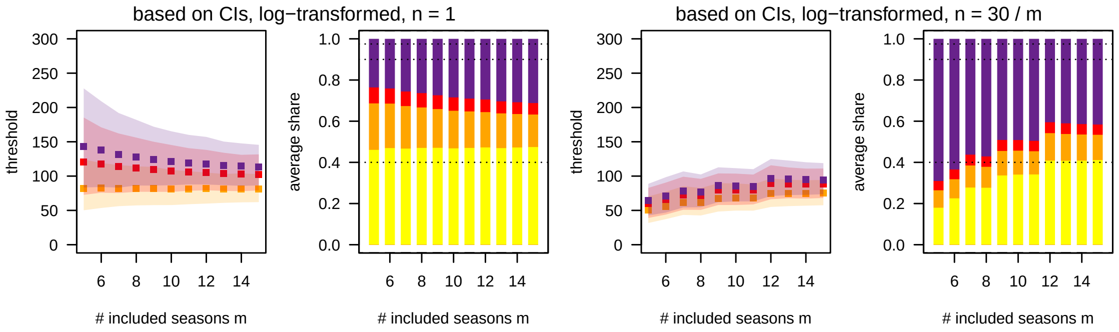


Figure 8: Average thresholds and exceedance shares when thresholds are based on confidence intervals rather than prediction intervals. See the caption and legend of Figure 4 for details on the plot elements.

6.4 Short summary of results based on US data

The results obtained using the US data are overall very well aligned with those presented for the France in the previous sections. While there are of course some differences in the exact values of exceedance probabilities, sensitivities etc., the overall patterns and conclusions are identical. The agreement with the analytical approximations even tends to be better for the US data.

7 Discussion

7.1 Practical recommendation

We provided a statistical assessment of implementation choices in a widely used framework for the computation of influenza intensity thresholds. Our practical recommendation in light of the theoretical and empirical findings is as follows. We suggest including just one observation per season into the reference set and employing a log transformation. To make thresholds somewhat less variable, data can be smoothed using a moving average with $l = 3$. If smoothing is applied the resulting thresholds should be applied to new data smoothed using the same procedure.

7.2 Further conclusions and outlook

For a more detailed conclusion we return to the criteria evoked in Section 2.

Calibration. The key to well-calibrated thresholds is that the observations in the reference set need to be comparable to new season peaks. This implies that no historical non-peak values should be included (i.e., one should choose $n = 1$). Otherwise thresholds will be pulled downwards and exceeded too frequently by new peaks. If historical data are smoothed ($l > 1$), comparability should be ensured by applying the same smoothing to new season peaks. Otherwise thresholds will again tend to be exceeded too often. Empirically we found that thresholds computed on log-transformed observations were better calibrated; while further study is needed we expect this to translate to other settings as distributions of peak values are typically skewed.

Sensitivity and specificity. Our theoretical and empirical results show that even if all assumptions are fulfilled, sensitivity and positive predictive values cannot be expected to exceed rather modest levels if m is small. This aspect is unaffected by whether smoothing is applied or not and reflects the general difficulty of estimating extreme quantiles from few historical observations. Especially for the very high threshold at $\alpha = 0.975$, sensitivity and the PPV must be expected to be limited and it can be asked whether these thresholds are practically meaningful. In any case it does not seem advisable to add even more extreme thresholds to the procedure. This problem is in theory reduced if many historical seasons are available, but in practice there is a tradeoff with the recency and comparability of these data. If there are relelvant temporal trends or changes in surveillance systems, historical data may get outdated and their inclusion may not be helpful.

Stability. The same difficulties limiting the sensitivity and PPV imply that thresholds must be expected to be rather variable if m is small. Empirically we found that this is even more the case if a log transformation is employed. Smoothing of historical data helps to somewhat reduce the variability of thresholds (but without improving sensitivity or specificity).

Simplicity. From a theoretical statistical standpoint the MEM and WHO methods are rather straightforward. However, informal exchange with users from public health indicates that given the large number of possible configurations, both methods are perceived as complex by practitioners. Our subjective opinion is that the methods overall strike a good balance here and that the demand for and practical benefits of more sophisticated methods may be limited. More specific guidance on how to choose the parameters of the methods and sensible default values are important to enable successful use of the methods.

Ease of practical application. Both methods can be applied using graphical user interfaces and thus do not require programming knowledge. For the MEM this is set up on top of an open-source R package with good documentation. For more technically versed users we recommend using this package directly. Combined with minimal data pre-processing it covers all aspects discussed in the present paper and enhances the automation and reproducibility of computations (which are the main challenge for graphical user interfaces).

In the present work we aimed to assess properties of the MEM and WHO methods in their current forms. However, a number of additional questions arise. Firstly, we focused purely on the assessment of season peak values, which is a useful perspective for temperate regions with one clear influenza wave per season. It could be discussed whether this should be complemented with other indicators and how more complex seasonal patterns as occuring in tropical regions should be handled. One possible alternative indicator is the season total, which conveys a different perspective on the severity of an influenza season. In this context we note that by smoothing an incidence time series prior to extracting peak values, we actually assess longer "peak periods", thus also changing the considered indicator. Which indicators are most relevant is a public health question and difficult to answer from a purely statistical standpoint; further exchange between the different subjects is thus needed.

Another general question is whether a categorization into four categories is the most practical way of conveying influenza activity. In principle it would also be possible to just report the percentile of the fitted distribution corresponding to a new peak. This would result in a continuous scale from 0 to 100, with higher values indicating higher intensity. However, given the previously discussed difficulties with estimating extreme quantiles, a percentile-based display may also convey a false sense of exactness.

A more technical question concerns how extreme seasons in the historical data should be handled. These can have a substantial effect on thresholds. In our simulation study we pragmatically removed observations from the pandemic season 2009/2010. However, a more principled approach to determining which historical seasons should be excluded would be desirable. As the same problem arises in outbreak detection, this strand of literature may be a helpful starting point (see e.g., Noufaily et al. 2013).

A limitation of our simulation study is that the re-sampled data do not actually stem from one and the same time series, but from several different regions. This made some re-scaling necessary. After log transformation, the resulting reference set is well-described by a normal distribution, but it is unclear how generalizable this is. A strength of our re-sampling scheme is that it ensures a realistic correlation structure between season peaks and the surrounding values, which would be challenging to achieve with fully synthetic data. A weakness is that dependence structures across seasons are not preserved. For certain respiratory diseases, alternating patterns between intense and mild seasons are common, likely due to gradual waning of immunity. Such patterns may affect the behaviour of thresholds, but are removed from our simulation study. We re-ran all simulation studies using a smaller data set from the United States and obtained very similar results. We take this as a sign of robustness at least for temperate settings. Intensity thresholds in tropical regions pose specific challenges like seasons with multiple peaks, which we did not address in the present manuscript.

To conclude, we re-emphasize the importance of a simple and interpretable thresholding method with a thorough open source software implementation like `mem`. The use of a standard approach will improve comparability of results and facilitate further methodological advances. With this work we hope to contribute to the development of best practices with a statistical perspective, complementing public health practitioners' applied experience.

Data and code

Materials to reproduce the presented results are available at `https://github.com/jbracher/mem_who`.

Ethics statement

No ethics approval was necessary as this study uses exclusively publicly available data.

Acknowledgements

We would like to thank Michael Höhle, Rob Moss, Laura Werlen and Daniel Wolffram for helpful discussions. Johannes Bracher was supported by the Helmholtz Foundation via the SIMCARD Information and Data Science Pilot Project as well as Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 512483310.

References

AbdElGawad, B., Vega, T., El Houssinie, M., Mohsen, A., Fahim, M., Abu ElSood, H., Jabbour, J., Eid, A., and Refaey, S. (2020). Evaluating tools to define influenza baseline and threshold values using surveillance data, Egypt, season 2016/17. *Journal of Infection and Public Health*, 13(3):430 – 437.

Bangert, M., Gil, H., Oliva, J., Delgado, C., Vega, T., De Mateo, S., and Larrauri, A. (2017). Pilot study to harmonize the reported influenza intensity levels within the Spanish influenza sentinel surveillance system (SISSS) using the Moving Epidemic Method (MEM). *Epidemiology and Infection*, 145(4):715–722.

Basile, L., Oviedo de la Fuente, M., Torner, N., Martínez, A., and Jané, M. (2018). Real-time predictive seasonal influenza model in Catalonia, Spain. *PLOS ONE*, 13(3):1–15.

Basile, L., Torner, N., Martínez, A., Mosquera, M., Marcos, M., and Jane, M. (2019). Seasonal influenza surveillance: Observational study on the 2017–2018 season with predominant B influenza virus circulation. *Vacunas*, 20(2):53 – 59.

Benedetti, G., White, R. A., Pasquale, H. A., Stassijns, J., van den Boogaard, W., Owiti, P., and Van den Bergh, R. (2019). Identifying exceptional malaria occurrences in the absence of historical data in South Sudan: a method validation. *Public Health Action*.

Biggerstaff, M., Kniss, K., Jernigan, D. B., Brammer, L., Bresee, J., Garg, S., Burns, E., and Reed, C. (2017). Systematic assessment of multiple routine and near real-time indicators to classify the severity of influenza seasons and pandemics in the United States, 2003–2004 through 2015–2016. *American Journal of Epidemiology*, 187(5):1040–1050.

Bouguerra, H., Boutouria, E., Zorraga, M., Cherif, A., Yazidi, R., Abdeddaiem, N., Maazaoui, L., ElMoussi, A., Abid, S., Amine, S., Bouabid, L., Bougatef, S., Kouni Chahed, M., Ben Salah, A., Bettaieb, J., and Bouafif Ben Alaya, N. (2020). Applying the moving epidemic method to determine influenza epidemic and intensity thresholds using influenza-like illness surveillance data 2009-2018 in Tunisia. *Influenza and Other Respiratory Viruses*, 14(5):507–514.

Dahlgren, F. S., Shay, D. K., Izurieta, H. S., Forshee, R. A., Wernecke, M., Chillarige, Y., Lu, Y., Kelman, J. A., and Reed, C. (2018). Evaluating oseltamivir prescriptions in centers for Medicare and Medicaid services medical claims records as an indicator of seasonal influenza in the United States. *Influenza and Other Respiratory Viruses*, 12(4):465–474.

Dahlgren, F. S., Shay, D. K., Izurieta, H. S., Forshee, R. A., Wernecke, M., Chillarige, Y., Lu, Y., Kelman, J. A., and Reed, C. (2019). Patterns of seasonal influenza activity in U.S. core-based statistical areas, described using prescriptions of oseltamivir in Medicare claims data. *Epidemics*, 26:23 – 31.

Dickson, E. M., Marques, D. F., Currie, S., Little, A., Mangin, K., Coyne, M., Reynolds, A., McMenamin, J., and Yirrell, D. (2020). The experience of point-of-care testing for influenza in Scotland in 2017/18 and 2018/19 – no gain without pain. *Eurosurveillance*, 25(44).

ECDC (2017). Risk assessment for seasonal influenza, EU/EEA, 2017–2018. Available online at `https://www.ecdc.europa.eu/sites/default/files/documents/RRA%20seasonal%20influenza%20EU%20EEA%202017-2018-rev_0.pdf`. Last accessed 27 December 2020.

Elhakim, M. M., Kandil, S. K., Abd Elaziz, K. M., and Anwar, W. A. (2019). Epidemiology of severe acute respiratory infection (SARI) cases at a sentinel site in Egypt, 2013–15. *Journal of Public Health*, 42(3):525–533.

Flahault, A., Blanchon, T., Dorléans, Y., Toubiana, L., Vibert, J. F., and Valleron, A. J. (2006). Virtual surveillance of communicable diseases: a 20-year experience in France. *Statistical Methods in Medical Research*, 15(5):413–421. PMID: 17089946.

Green, H., Charlett, A., Moran-Gilad, J., Fleming, D., Durnall, H., Thomas, D., Cottrell, S., Smyth, B., Kearns, C., Reynolds, A., Smith, G., Elliot, A., Ellis, J., Zambon, M., JM, W., McMenamin, J., and Pebody, R. (2015). Harmonizing influenza primary-care surveillance in the United Kingdom: piloting two methods to assess the timing and intensity of the seasonal epidemic across several general practice-based surveillance schemes. *Epidemiology and Infection*, 143(1):1–12.

Grilc, E., Prosenc Trilar, K., Lajovic, J., and Sočan, M. (2021). Determining the seasonality of respiratory syncytial virus in Slovenia. *Influenza and Other Respiratory Viruses*, 15(1):56–63.

Harcourt, S. E., Morbey, R. A., Smith, G. E., Loveridge, P., Green, H. K., Pebody, R., Rutter, J., Yeates, F. A., Stuttard, G., and Elliot, A. J. (2019). Developing influenza and respiratory syncytial virus activity thresholds for syndromic surveillance in England. *Epidemiology and Infection*, 147:e163.

HELM (2008). Workbook 40: Sampling distributions and estimation. Technical report, Loughbrough University. available online, https://www.lboro.ac.uk/departments/mlsc/student-resources/helm-workbooks/.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions. Vol. 1.* John Wiley and Sons.

Liu, L., Yue, J., Lai, X., Huang, J., and Zhang, J. (2019). Multivariate nonparametric chart for influenza epidemic monitoring. *Scientific Reports*, article number 17472.

Lozano, J. (2020). mem: The moving epidemic method. R package, version 2.16 available via CRAN, https://cran.r-project.org/web/packages/mem/index.html.

Lucero, M. G., Inobaya, M. T., Nillos, L. T., Tan, A. G., Arguelles, V. L. F., Dureza, C. J. C., Mercado, E. S., Bautista, A. N., Tallo, V. L., Barrientos, A. V., Rodriguez, T., and Olveda, R. M. (2016). National influenza surveillance in the Philippines from 2006 to 2012: seasonality and circulating strains. *BMC Infectious Diseases*, 16(1):762.

Ly, S., Arashiro, T., Ieng, V., Tsuyuoka, R., Parry, A., Horwood, P., Heng, S., Hamid, S., Vandemaele, K., Chin, S., Sar, B., and Arima, Y. (2017). Establishing seasonal and alert influenza thresholds in Cambodia using the WHO method: implications for effective utilization of influenza surveillance in the tropics and subtropics. *Western Pacific Surveillance and Response Journal*, 8:22–32.

Murray, J. L. K., Marques, D. F. P., Cameron, R. L., Potts, A., Bishop, J., von Wissmann, B., William, N., Reynolds, A. J., Robertson, C., and McMenamin, J. (2018). Moving Epidemic Method (MEM) applied to virology data as a novel real time tool to predict peak in seasonal influenza healthcare utilisation. The Scottish experience of the 2017/18 season to date. *Eurosurveillance*, 23(11):18–00079.

Nisar, N., Aamir, U. B., Badar, N., Mahmood, M. R., Yaqoob, A., Tripathy, J. P., Laxmeshwar, C., Tenzin, K., Zaidi, S. S. Z., Salman, M., and Ikram, A. (2020). Epidemiology of influenza among patients with influenza-like illness and severe acute respiratory illness in Pakistan: A 10-year surveillance study 2008-17. *Journal of Medical Virology*, 92(12):3028–3037.

Noufaily, A., Enki, D. G., Farrington, P., Garthwaite, P., Andrews, N., and Charlett, A. (2013). An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine*, 32(7):1206–1222.

Pesälä, S., Virtanen, M. J., Mukka, M., Ylilammi, K., Mustonen, P., Kaila, M., and Helve, O. (2019). Healthcare professionals' queries on oseltamivir and influenza in Finland 2011-2016 – can we detect influenza epidemics with specific online searches? *Influenza and Other Respiratory Viruses*, 13(4):364–371.

Politis, D. (2001). Resampling time series with seasonal components. *Frontiers in Data Mining and Bioinformatics: Proceedings of the 33rd Symposium on the Interface of Computing Science and Statistics: Orange County, California, June 13-17*, page 619–621.

Páscoa, R., Rodrigues, A. P., Silva, S., Nunes, B., and Martins, C. (2018). Comparison between influenza coded primary care consultations and national influenza incidence obtained by the General Practitioners Sentinel Network in Portugal from 2012 to 2017. *PLOS ONE*, 13(2):1–10.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rakocevic, B., Grgurevic, A., Trajkovic, G., Mugosa, B., Grujicic, S. S., Medenica, S., Bojovic, O., Alonso, J. E. L., and Vega, T. (2019). Influenza surveillance: determining the epidemic threshold for influenza by using the Moving Epidemic Method (MEM), Montenegro, 2010/11 to 2017/18 influenza seasons. *Eurosurveillance*, 24(12):1800042.

Redondo-Bravo, L., Delgado-Sanz, C., Oliva, J., Vega, T., Lozano, J., Larrauri, A., and the Spanish Influenza Sentinel Surveillance System (2020). Transmissibility of influenza during the 21st-century epidemics, Spain, influenza seasons 2001/02 to 2017/18. *Eurosurveillance*, 25(21).

Rguig, A., Cherkaoui, I., McCarron, M., Oumzil, H., Triki, S., Elmbarki, H., Bimouhen, A., Falaki, F. E., Regragui, Z., Ihazmad, H., Nejjari, C., and Youbi, M. (2020). Establishing seasonal and alert influenza thresholds in Morocco. *BMC Public Health.*

Sullivan, S. G., Arriola, C. S., Bocacao, J., Burgos, P., Bustos, P., Carville, K. S., Cheng, A. C., Chilver, M. B., Cohen, C., Deng, Y.-M., El Omeiri, N., Fasce, R. A., Hellferscee, O., Huang, Q. S., Gonzalez, C., Jelley, L., Leung, V. K., Lopez, L., McAnerney, J. M., McNeill, A., Olivares, M. F., Peck, H., Sotomayor, V., Tempia, S., Vergara, N., von Gottberg, A., Walaza, S., and Wood, T. (2019). Heterogeneity in influenza seasonality and vaccine effectiveness in Australia, Chile, New Zealand and South Africa: early estimates of the 2019 influenza season. *Eurosurveillance*, 24(45).

Tay, E. L., Grant, K., Kirk, M., Mounts, A., and Kelly, H. (2013). Exploring a proposed WHO method to determine thresholds for seasonal influenza surveillance. *PLOS ONE*, 8(10).

Torner, N., Basile, L., Martínez, A., Rius, C., Godoy, P., Jané, M., Domínguez, A., Aizpurua, J., Alonso, J., Azemar, J., Aizpurua, P., Ardaya, P. M., Basas, M. D., Batalla, J., Biendicho, P., Bonet, M., Callado, M., Campos, S., Casanovas, J. M., Ciurana, E., Claples, M., Cots, J. M., De la Rica, D., Domingo, I., Elizalde, G., Escapa, P., Fajardo, S., Fau, E., Fernandez, O., Fernandez, M., Ferrer, C., Forcada, A., Fos, E., Gadea, G., Garcia, J., Garcia, R., Gatius, C., Gelado, M. J., Grau, M., Grivé, M., Guzman, M. C., Hernández, R., Jimenez, G., Juscafresa, A., LLussa, A. M., López, C., Kristensen, L., Macià, E., Mainou, A., Marco, E., Martínez, M., Martínez, J. G., Marulanda, K. V., Masa, R., Moncosí, X., Naranjo, M. A., Navarro, D., Ortolà, E., París, F., Pérez, M. M., Pozo, C., Pujol, R., Ribatallada, A., Ruiz, G., Sabaté, S., Sanchez, R., Sarrà, N., Tarragó, E., Teixidó, A. M., Torres, A., Valén, E., Van Esso, D., Van Tarjcwick, C., Vink Schoenholzer, R., Zabala, E., Marcos, M. A., Mosquera, M. D. M., de Molina, P., Rubio, E., Isanta, R., Anton, A., Pumarola, T., Vilella, A., Gorrindo, P., Espejo, E., Andrés, M., Barcenilla, F., Navarro, G., Barrabeig, I., Pou, J., Alvarez, P., Plasencia, E., Rebull, J., Sala, M. R., Riera, M., Camps, N., Follia, N., Oller, A., Godoy, P., Bach, P., Rius, C., Hernández, R., Perez, R., Torra, R., Carol, M., Minguell, S., Marce, R., Garcia-Pardo, G., Olona, M., Alvarez, A., Ramon, J. M., Mòdol, J. M., Mena, G., Campins, M., Massuet, C., Tora, G., Ferràs, J., Ferrús, G., and The Working Group on PIDIRAC Sentinel Surveillance of Catalonia (2019). Assessment of two complementary influenza surveillance systems: sentinel primary care influenza-like illness versus severe hospitalized laboratory-confirmed influenza using the moving epidemic method. *BMC Public Health*, 19(1089).

Vega, T., Lozano, J. E., Meerhoff, T., Snacken, R., Beauté, J., Jorgensen, P., Ortiz de Lejarazu, R., Domegan, L., Mossong, J., Nielsen, J., Born, R., Larrauri, A., and Brown, C. (2015). Influenza surveillance in Europe: comparing intensity levels calculated using the Moving Epidemic Method. *Influenza and Other Respiratory Viruses*, 9(5):234–246.

Vega, T., Lozano, J. E., Meerhoff, T., Snacken, R., Mott, J., Ortiz de Lejarazu, R., and Nunes, B. (2013). Influenza surveillance in Europe: establishing epidemic thresholds by the Moving Epidemic Method. *Influenza and Other Respiratory Viruses*, 7(4):546–558.

Vette, K., Bareja, C., Clark, R., and Lal, A. (2018). Establishing thresholds and parameters for pandemic influenza severity assessment, Australia. *Bulletin of the World Health Organization*, 96:558–567.

Vos, L. M., Teirlinck, A. C., Lozano, J. E., Vega, T., Donker, G. A., Hoepelman, A. I., Bont, L. J., Oosterheert, J. J., and Meijer, A. (2019). Use of the moving epidemic method (MEM) to assess national surveillance data for respiratory syncytial virus (RSV) in the Netherlands, 2005 to 2017. *Eurosurveillance*, 24(20):1800469.

Wagner, M., Lampos, V., Cox, I. J., and Pebody, R. (2018). The added value of online user-generated content in traditional methods for influenza surveillance. *Scientific Reports*, 8(1):13963.

WHO (2011). Strengthening response to pandemics and other public-health emergencies: Report of the review committee on the functioning of the international health regulations (2005) and on pandemic influenza (H1N1) 2009. Available online at `https://www.who.int/ihr/publications/RC_report/en/`. Last accessed 26 December 2020.

WHO (2014). WHO global epidemiological surveillance standards for influenza. Available online at `https://www.who.int/influenza/resources/documents/influenza_surveillance_manual/en/` (last accessed 27 December 2020).

WHO (2017). Pandemic influenza severity assessment (PISA). Available online at `https://apps.who.int/iris/handle/10665/259392`. Last accessed 27 December 2021.

WHO (2023). WHO average curves app guidance and documentation, v.0.3. available online at `https://worldhealthorg.shinyapps.io/averagecurves/`.

A Derivations of analytical approximations

A.1 Derivations for Section 5.1

We start by addressing the expectations of empirical mean $\bar{Y}$ and variance S^2 of the reference observations, where

$$\bar{Y} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} Y_i^{(j)}$$

$$S^2 = \frac{1}{mn-1}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(Y_i^{(j)} - \bar{Y}\right)^2.$$

It is straightforward to see that

$$\mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mu_i. \tag{15}$$

For the variance S^2, we first note that it can be re-written as

$$S^2 = \frac{mn}{mn-1}\left\{\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(Y_i^{(j)} - \bar{Y}\right)^2\right\} \tag{16}$$

$$= \frac{mn}{mn-1}\left\{\underbrace{\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}Y_i^{(j)2}}_{\text{denote this by } a} \quad - \quad \underbrace{\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}Y_i^{(j)}\right)^2}_{\text{denote this by } b}\right\} \tag{17}$$

We consider the two terms a and b separately, starting by

$$\begin{aligned}\mathbb{E}(a) &= \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left(Y_i^{(j)2}\right)\\ &= \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\mathrm{Var}\left(Y_i^{(j)}\right) + \mathbb{E}\left(Y_i^{(j)}\right)^2\right\}\\ &= \frac{m}{mn}\sum_{i=1}^{n}(\sigma_i^2 + \mu_i^2).\end{aligned}$$

Then we note that

$$\begin{aligned}\mathbb{E}(b) &= \mathbb{E}\left\{\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}Y_i^{(j)}\right)^2\right\}\\ &= \mathrm{Var}\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}Y_i^{(j)}\right) \quad + \quad \mathbb{E}\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}Y_i^{(j)}\right)^2\\ &= \frac{1}{(mn)^2}\sum_{j=1}^{m}\mathrm{Var}\left(\sum_{i=1}^{n}Y_i^{(j)}\right) \quad + \quad \left(\frac{m}{mn}\sum_{i=1}^{n}\mu_i\right)^2\\ &= \frac{m}{(mn)^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sigma_{i,i'} \quad + \quad \frac{m^2}{(mn)^2}\left(\sum_{i=1}^{n}\mu_i\right)^2.\end{aligned}$$

Plugging these results back into equation (17) we obtain

$$\mathbb{E}(S^2) = \frac{m}{mn-1}\sum_{i=1}^{n}(\sigma_i^2 + \mu_i^2) \quad - \quad \frac{1}{n(mn-1)}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sigma_{i,i'} \quad - \quad \frac{m}{n(mn-1)}\left(\sum_{i=1}^{n}\mu_i\right)^2. \tag{18}$$

It is straightforward to see that for $n = 1$ the expressions (4) and (18) simplify to

$$\mathbb{E}(\bar{Y}) = \mu_1 \quad \text{and} \quad \mathbb{E}(S^2) = \sigma_1^2,$$

as given in equation (5).

There is no general way of computing the expectation $\mathbb{E}(S)$ from $\mathbb{E}(S^2)$, but unless the true distribution of S^2 has strong excess curtosis,

$$\mathbb{E}(S) \approx \sqrt{\mathbb{E}(S^2)} \tag{19}$$

is a reasonable approximation. We can then plug equations (4) and (19) into the formulae for the thresholds $q_{Y,\alpha}$ on the transformed scale and obtain

$$\mathbb{E}(\hat{q}_{Y,\alpha}) \approx \mathbb{E}(\bar{Y}) + z_\alpha\sqrt{\mathbb{E}(S^2)},$$

where z_α is the α quantile of the standard normal distribution (with $\alpha \in \{0.4, 0.9, 0.975\}$).

If f was set to the natural logarithm, the question remains how to obtain statements concerning thresholds $q_{X,\alpha}$ on the original scale. Approximation via a second-order Taylor expansion yields

$$\mathbb{E}(\hat{q}_{X,\alpha}) = \mathbb{E}\left\{\exp(\hat{q}_{Y,\alpha})\right\} \approx \exp\left\{\mathbb{E}(\hat{q}_{Y,\alpha})\right\} \times \left\{1 + \frac{\operatorname{Var}(\hat{q}_{Y,\alpha})}{2}\right\}. \tag{20}$$

Empirically, after transformation to the log scale, the variance of the reference observations is low in our applied setting. The resulting variances of $q_{Y,\alpha}$ are then quite small and do not play an important role in equation (20). We can thus use the even simpler approximation

$$\mathbb{E}(q_{\hat{Y},\alpha}) \approx \exp\left\{\mathbb{E}(\hat{q}_{Y,\alpha})\right\} \approx \exp\{\mathbb{E}(\bar{Y}) + z_\alpha \sqrt{\mathbb{E}(S^2)}\}. \tag{21}$$

As can be seen from Figures 2 and 3 from the main manuscript, this approximation works very well in practice.

A.2 Derivations for Section 5.2

Remember that we denote by p_j the peak week of the smoothed incidences in season j and defined

$$x^{\text{smo}}_{j,p_j} = \sum_{d=0}^{l} x^{\text{raw}}_{j,p_j-d} = \begin{pmatrix} 1/l \\ \vdots \\ 1/l \end{pmatrix}^{\top} \mathbf{X}^{\text{raw}}_j .$$

As we assumed that $n = 1$ and f is the identity function (i.e., $y^{(1)}_j = x^{(1)}_j$), equation (1) simplifies to

$$\begin{aligned} \bar{Y} &= \frac{\sum_{j=1}^{m} x^{\text{smo}}_{j,p_j}}{nm} \\ S^2 &= \sqrt{\frac{\sum_{j=1}^{m}(x^{\text{smo}}_{j,p_j} - \bar{y})}{nm-1}}. \end{aligned} \tag{22}$$

It is clear that

$$\mathbb{E}(\bar{Y}) = \begin{pmatrix} 1/l \\ \vdots \\ 1/l \end{pmatrix}^{\top} \mathbb{E}(\mathbf{X}^{\text{raw}\top}_j) = \begin{pmatrix} 1/l \\ \vdots \\ 1/l \end{pmatrix}^{\top} \boldsymbol{\mu}^{\text{raw}} = \frac{1}{l}\sum_{i=1}^{l} \mu^{\text{raw}}_i .$$

The expression for S^2 is known to be an unbiased estimator for the variance of X^{smo}_{j,p_j}, i.e.

$$\mathbb{E}(S^2) = \operatorname{Var}(X^{\text{smo}}_{j,p_j}) = \begin{pmatrix} 1/l \\ \vdots \\ 1/l \end{pmatrix}^{\top} \boldsymbol{\Sigma}^{\text{raw}} \begin{pmatrix} 1/l \\ \vdots \\ 1/l \end{pmatrix} = \frac{1}{l^2}\sum_{i=1}^{l}\sum_{i'=1}^{n} \sigma^{\text{raw}}_{i',i} .$$

This completes the proof.

A.3 Derivations for Section 5.3

As the exceedance of thresholds is invariate to shifting and scaling of the $Y^{(i)}_j$ we can for simplicity assume that they follow a standard normal distribution with $\mu_1 = 0, \sigma_1^2 = 1$. Now consider

$$\hat{q}_{Y,\alpha} = \bar{Y} + z_\alpha S.$$

Basu's theorem tells us that the sample mean and standard deviation are independent. We can thus address their respective distributions separately. We obviously have

$$\bar{Y} \sim \mathrm{N}\left(0, \frac{1}{m}\right).$$

It is moreover known that for the sample variance S^2 of m standard normal random variables

$$[(m-1) \times S^2] \sim \chi^2(m-1)$$

holds (HELM, 2008). The square root of a χ^2 distributed random variable can be approximated well by a normal distribution (Johnson et al., 1994, p426). Specifically, we get

$$\sqrt{2 \times (m-1) \times S^2} \overset{\text{approx}}{\sim} \mathrm{N}(\sqrt{2m-2}, 1)$$

and thus

$$S \overset{\text{approx}}{\sim} \mathrm{N}\left(1, \frac{1}{2 \times (m-1)}\right).$$

Combining these two results and using the independece of sample mean and standard deviation, we get

$$\hat{q}_{Y,\alpha} \overset{\text{approx}}{\sim} \mathrm{N}\left[z_\alpha, \frac{1}{m} + z_\alpha^2 \times \frac{1}{2 \times (m-1)}\right].$$

Equation (11) is then obtained by simple re-scaling and shifting.

We can now compute the sensitivity. It corresponds to the probability that $Y_{m+1}^{(1)} > \hat{q}_{Y,\alpha}$ given exceedance $Y_{m+1}^{(1)} > z_\alpha$ (keep in mind that $y_{m+1}^{(1)}$ it is assumed to follow a standard normal distribution). This conditional probability is straightforward to evaluate using the integral in equation (12). The derivation of the specificity follows the same argument.

B Supplementary figure based on French data

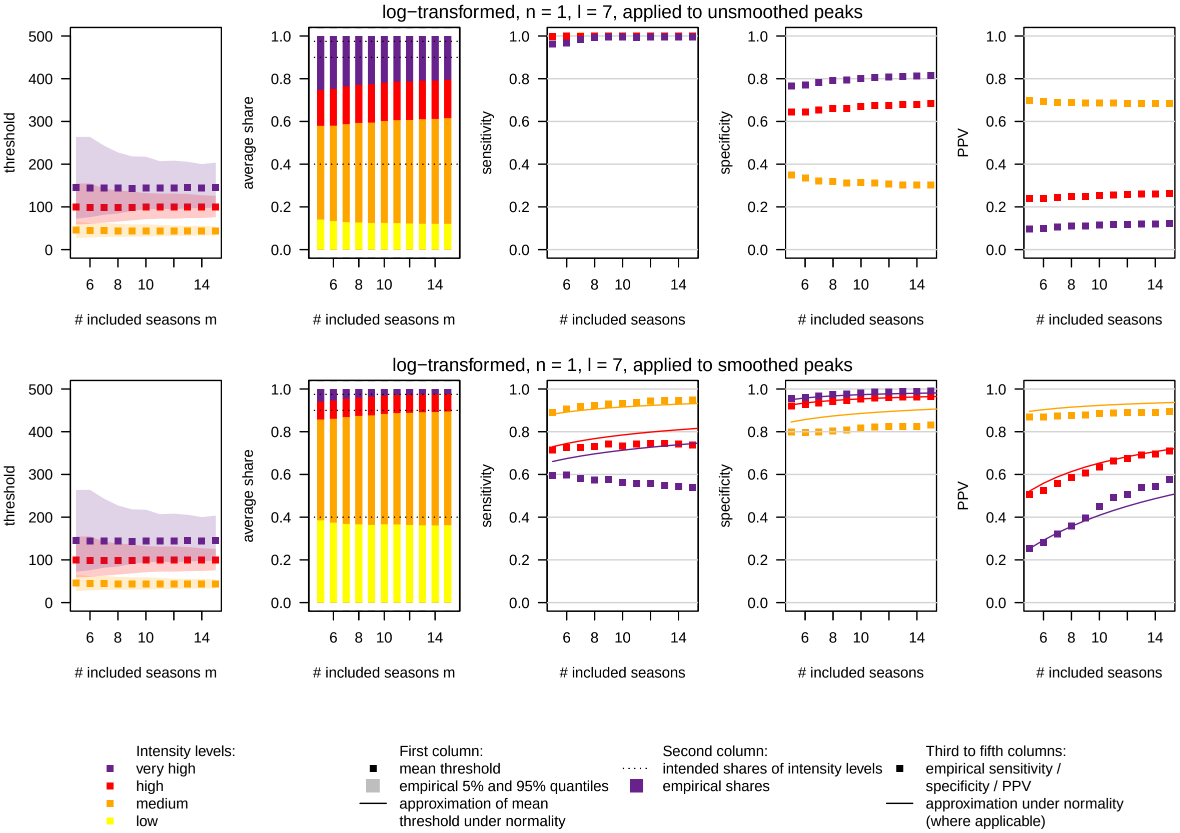


Figure 9: Impact of smoothing of historical data on thresholds. We applied a moving average with $l = 7$ to the historical time series prior to computing thresholds and subsequently applied them to either unsmoothed or smoothed new peak values. Results are shown for thresholds computed with a log transformation. See the legend of Figure 4 for details on the plot elements.

C Supplementary figures based on US data

In the following, Figures 3–8 are reproduced using the data from the US described in Section 6.2.

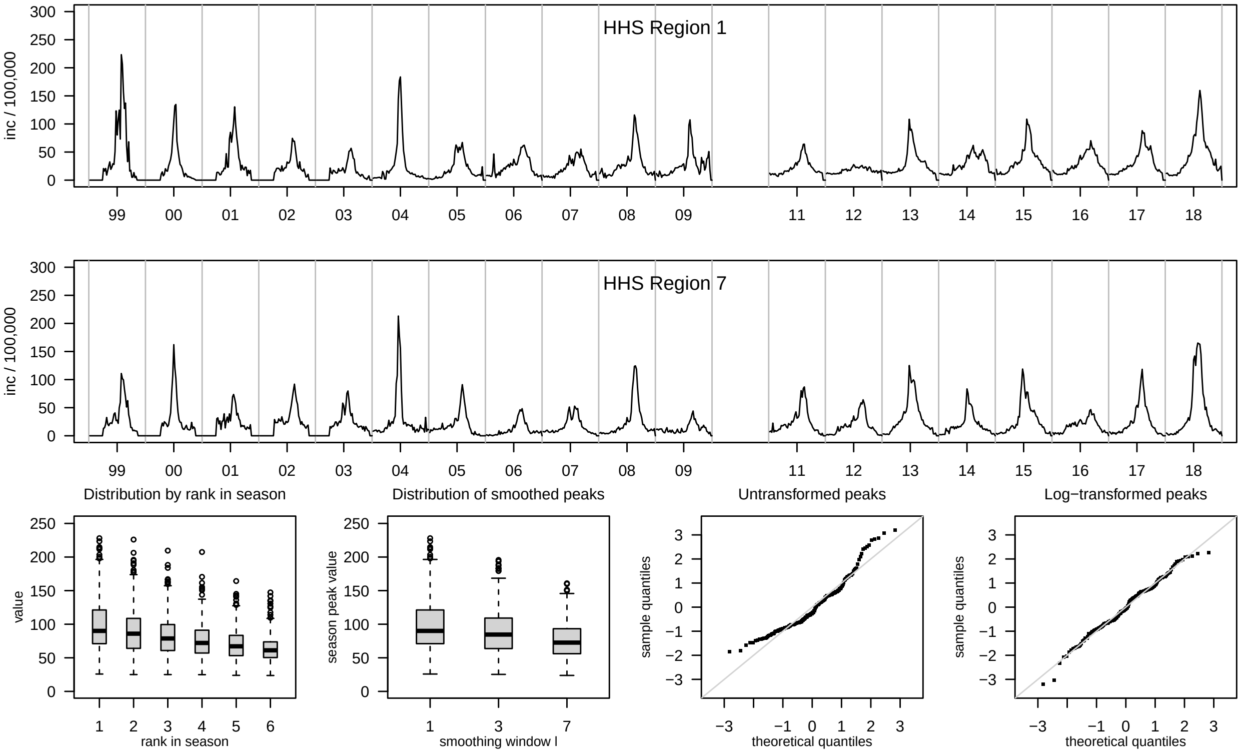


Figure 10: Reproduction of Figure 3 using US data. Time series of weekly weighted ILI percentages in US HHS Regions 1 and 7, 1999–2018. Off-season weeks are omitted in the plot, with grey lines delimiting the different seasons. The bottom row shows descriptive plots of the distribution of season peaks. First: Boxplots of incidence values by rank within season. Second: Boxplot of smoothed peak values as a function of the smoothing window width l. Third: Normal QQ plot of untransformed peak values. Fourth: Normal QQ plot of log-transformed peak values.

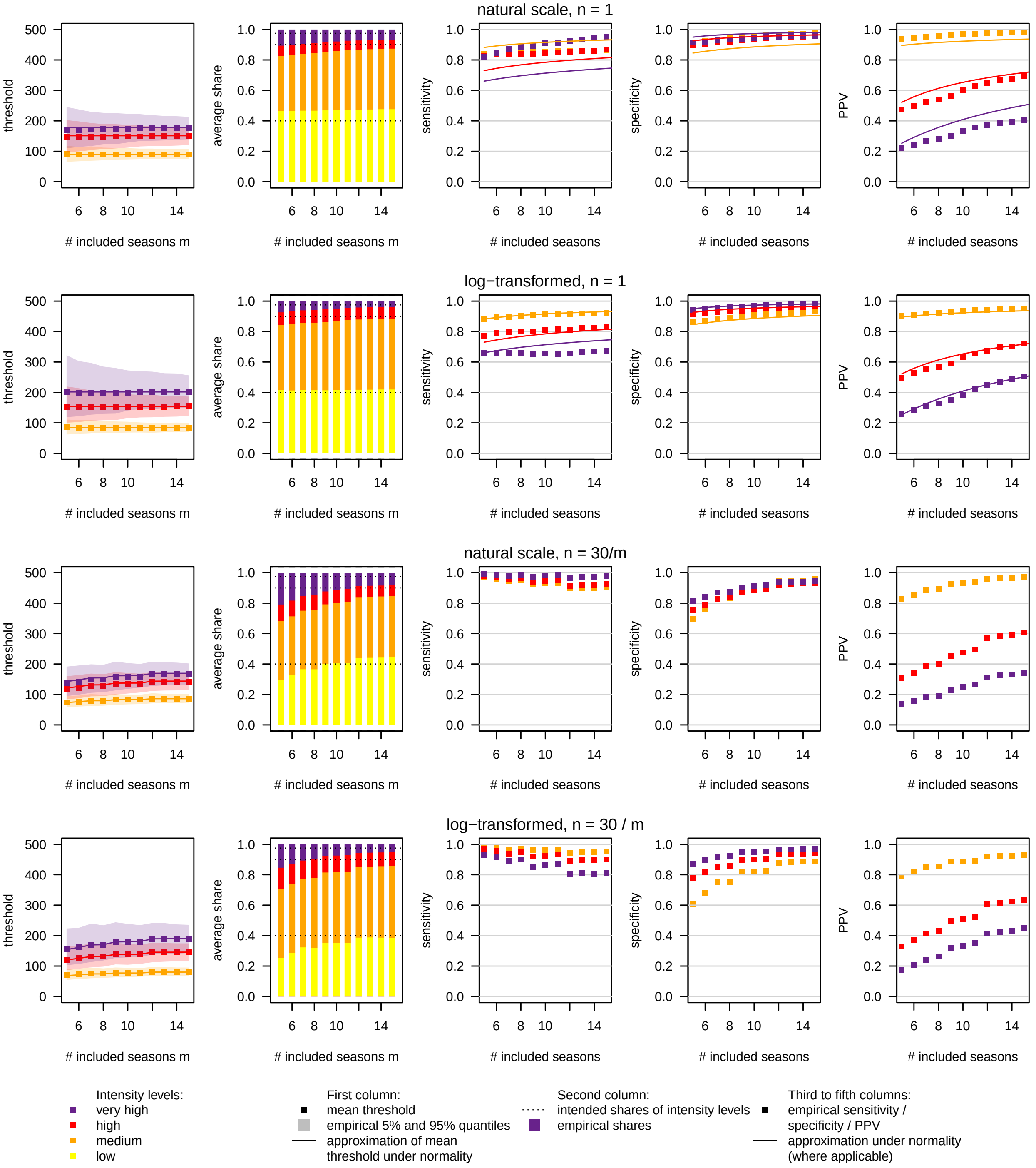


Figure 11: Reproduction of Figure 4 using US data. Impact of the choice of n and transformation function f. First column: simulation-based mean intensity thresholds (squares) along with bands delimited by the empirical 5% and 95% quantiles. Analytical approximations of mean threshold values (computed from empirical means and covariances) are displayed as lines. Second column: resulting average shares of season peaks classified as low, medium, high and very high intensity. Third to fifth columns: sensitivity, specificity and PPVs of the different thresholds. Simulation results are shown as squares. Where available analytical approximations are shown as lines.

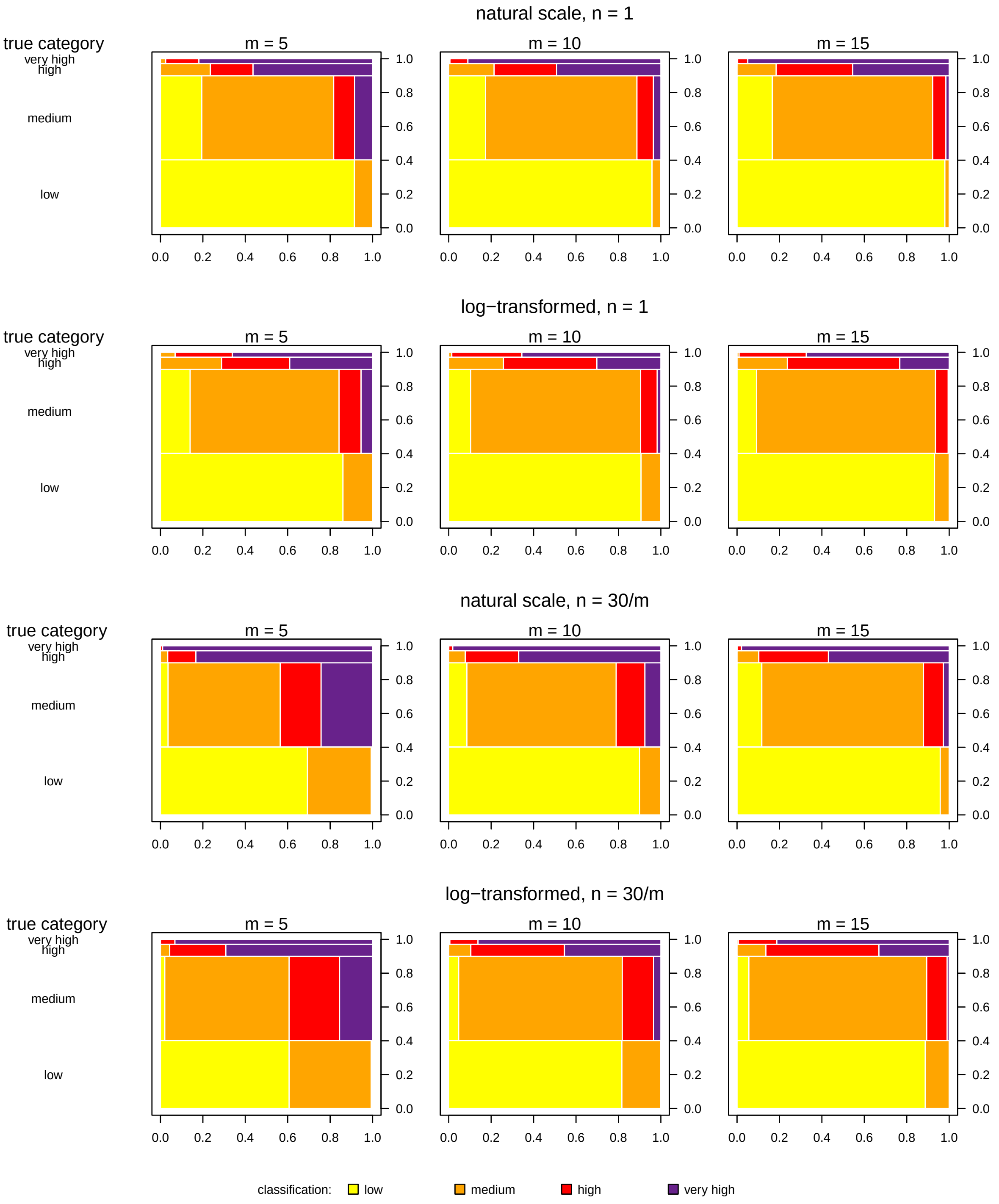


Figure 12: Reproduction of Figure 5 using US data. Confusion matrices for intensity classifications obtained with different choices of n and transformation function f. Mosaic plots show which fractions of season peaks which are truly very high, high, medium or low are classified into the four categories. The true class is determined with respect to the empirical quantiles of the distribution of peaks: very high (highest 2.5% of all peaks), high (next 7.5%), medium (next 50%), low (lowest 40% of all peaks).

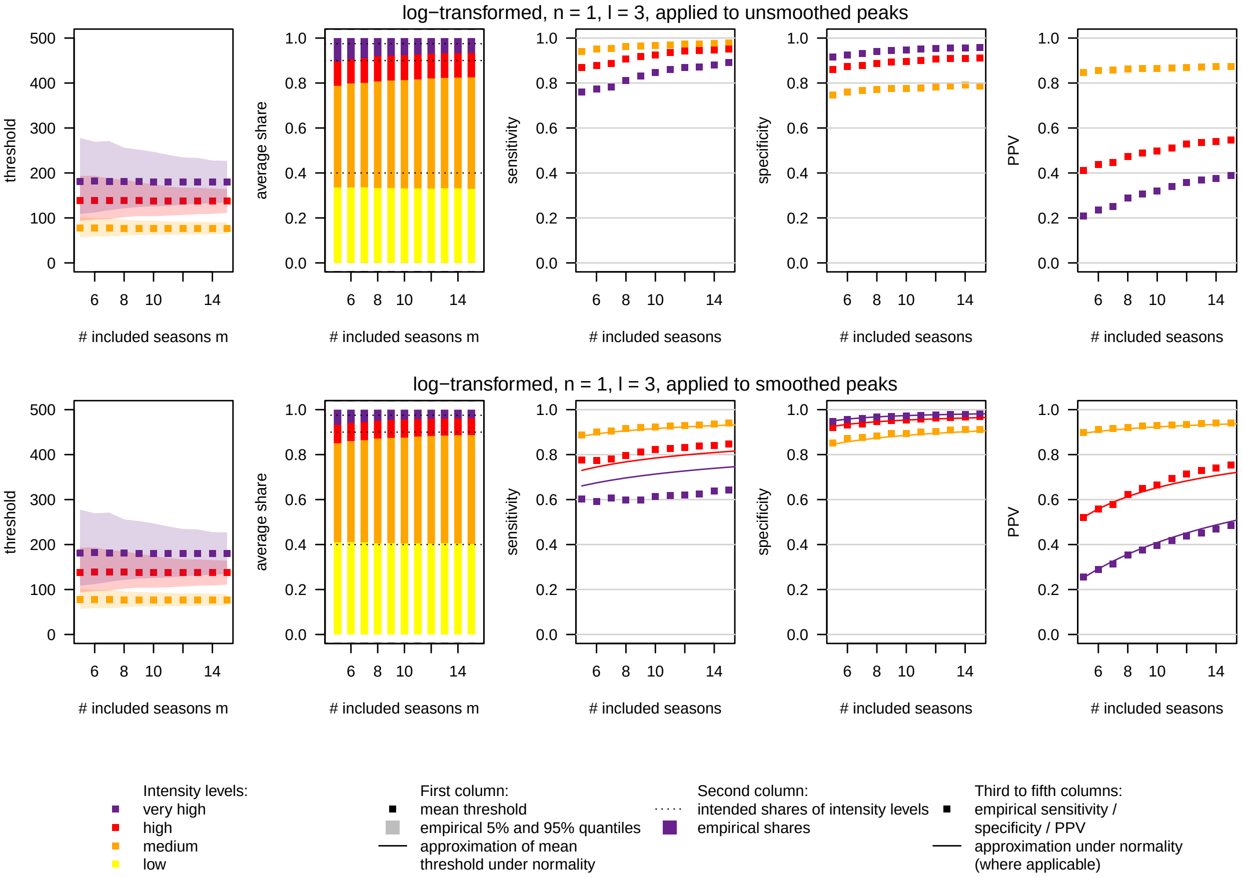


Figure 13: Reproduction of Figure 6 using US data. Impact of smoothing of historical data on thresholds. We applied a moving average with $l = 3$ to the historical time series prior to computing thresholds and subsequently applied them to either unsmoothed or smoothed new peak values. Results are shown for thresholds computed with a log transformation. See the caption of Figure 4 for details on the plot elements.

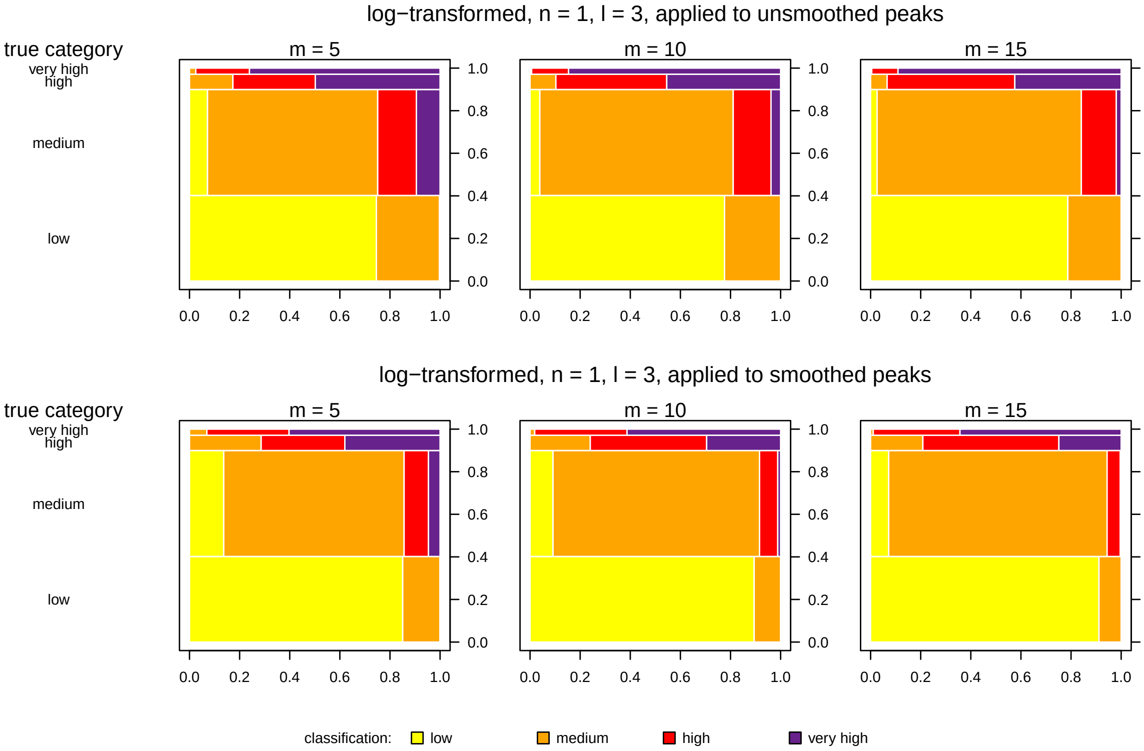


Figure 14: Reproduction of Figure 6.3.2 using US data. Confusion matrices for intensity classifications obtained with a smoothing window of width $l = 3$ and a log transformation. Top: thresholds applied to unsmoothed new peaks; bottom: thresholds applied to smoothed new peaks. Mosaic plots show which fractions of season peaks which are truly very high, high, medium or low are classified into the four categories, see camption of Figure 5 for details.

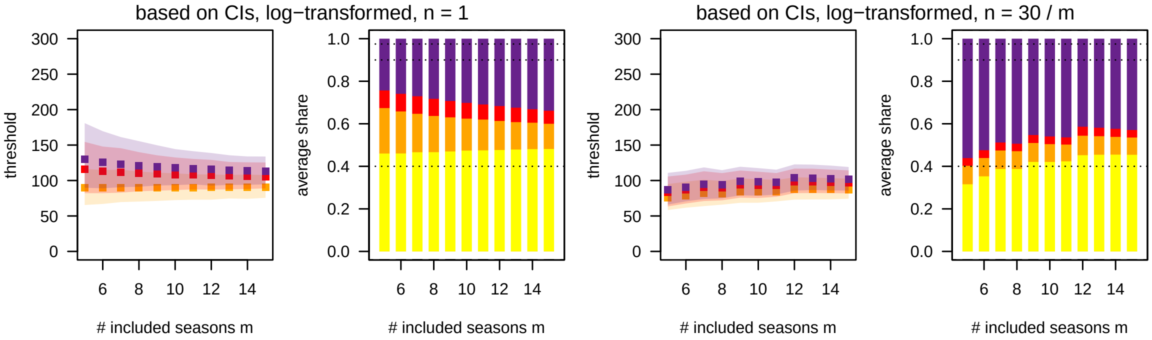


Figure 15: Reproduction of Figure 8 using US data. Average thresholds and exceedance shares when thresholds are based on confidence intervals rather than prediction intervals. See the caption and legend of Figure 4 for details on the plot elements.