Overlapping Time Scales Obscure Early Warning Signals of the Second COVID-19 Wave

Fabian Dablander[1], Hans Heesterbeek[2], Denny Borsboom[1], and John M. Drake[3,4]

[1] *Department of Psychological Methods, University of Amsterdam*
[2] *Department of Population Health Sciences, Utrecht University*
[3] *Odum School of Ecology, University of Georgia*
[4] *Center for the Ecology of Infectious Diseases, University of Georgia*

Abstract

Early warning indicators based on critical slowing down have been suggested as a model-independent and low-cost tool to anticipate the (re)emergence of infectious diseases. We studied whether such indicators could reliably have anticipated the second COVID-19 wave in European countries. Contrary to theoretical predictions, we found that characteristic early warning indicators generally *decreased* rather than *increased* prior to the second wave. A model explains this unexpected finding as a result of transient dynamics and the multiple time scales of relaxation during a non-stationary epidemic. Particularly, if an epidemic that seems initially contained after a first wave does not fully settle to its new quasi-equilibrium prior to changing circumstances or conditions that force a second wave, then indicators will show a decreasing rather than an increasing trend as a result of the persistent transient trajectory of the first wave. Our simulations show that this lack of time scale separation was to be expected during the second European epidemic wave of COVID-19. Overall, our results emphasize that the theory of critical slowing down applies only when the external forcing of the system across a critical point is slow relative to the internal system dynamics.

1 Introduction

Forecasting the (re)emergence of infectious diseases is of great importance to public health (George et al., 2019; Heesterbeek et al., 2015; Morens & Fauci, 2013; Morens et al., 2004; Reich et al., 2019; Viboud et al., 2018). In recent years, early warning indicators based on the phenomenon of critical slowing down have been suggested as a way to anticipate transitions in a wide range of dynamical systems (for overviews, see e.g., Dablander et al., 2020; Drake et al., 2019; Drake et al., 2020; Lenton, 2011; Scheffer et al., 2009; Scheffer et al., 2015). Critical slowing down describes the phenomenon that many systems, as they approach their critical point, return more slowly to their equilibrium after small external perturbations, resulting in an increase in statistics such as the local autocorrelation coefficient and variance (Drake & Griffen, 2010; Wissel, 1984). In standard models of infectious disease transmission, major outbreaks are possible when the effective reproductive number, R_t, is greater than one. The threshold $R_t = 1$ corresponds to a (dynamic) transcritical bifurcation, which is a type of bifurcation that is preceded by critical slowing down (Kéfi et al., 2014; Kuehn, 2011). Early warning indicators based on critical slowing down have been studied extensively and led to a promising research line that aims to utilize them as a tool to forecast the (re)emergence as well as the elimination of infectious diseases (e.g., Brett et al., 2020; Brett et al., 2017; Brett et al., 2018; Brett & Rohani, 2020;

Dessavre et al., 2019; Dibble et al., 2016; Drake et al., 2019; Harris et al., 2020; Miller et al., 2017; O'Dea et al., 2018; O'Regan & Burton, 2018; O'Regan & Drake, 2013; O'Regan et al., 2020; Southall et al., 2020).

In light of this prior research, it seems natural to ask whether early warning indicators based on critical slowing down could have allowed us to anticipate the second COVID-19 wave (e.g., O'Brien & Clements, 2021; Proverbio et al., 2021) and if not, how this can be understood. Here, we question the applicability of early warning indicators in the COVID-19 context, because the COVID-19 situation violates a key assumption of the theory of critical slowing down: a separation of time scales such that the dynamics of the epidemic settle down to a quasi-equilibrium from which there is a slow drift toward the critical point. To our knowledge, there is presently no theory that would indicate whether early warning signals, under such commensurate time scales, can be expected to be reliable. In this paper, we report on a combination of empirical analysis and simulation studies to investigate this issue. Focusing on Europe, we find that a suite of early warning indicators did not reliably rise prior to the second wave in any country as the classical theory of critical slowing down would predict. Using a simulation study that mimics the COVID-19 situation — a first outbreak closely followed by a second one — we show that this contradictory result can be fully explained by the fact that, in the case of COVID-19, in almost all countries R_t already began to creep up again before the number of case reports stabilized at a low value after the first wave. These results indicate that caution is warranted in applying early warning indicators to highly non-stationary settings, such as multi-wave epidemics.

2 Early warning signals for COVID-19

In this section, we quantify the extent to which early warning indicators increased prior to the second wave in a number of European countries.[1] We outline our methodology aided by Figure 1 in Section 2.1, and report our results in Section 2.2.

2.1 Methods

Estimation of R_t. To identify the time at which the COVID-19 epidemic became supercritical for the second time in each country, we followed Gostic et al. (2020) to estimate the instantaneous R_t using the method of Abbott et al. (2020), which improves upon Cori et al. (2013). The method simultaneously estimates the incidence of infections and R_t using Bayesian latent variable modeling. The method proceeds in two steps. First, the incidence at each time step is estimated by convolving the previous number of infections with a probability distribution for the generation interval. This incidence is then convolved over an uncertain incubation period and reporting delay distribution to yield the reported cases (for details, see Abbott et al., 2020). We applied this method to a broad range of European countries using data from March to October 2020.

Selecting the time period between waves. Next, we selected a time period in which to search for evidence of critical slowing down. Early warning indicators are sensitive to changes in the effective reproductive number, and should rise prior to the critical point $R_t = 1$ (Drake et al., 2019; O'Regan & Drake, 2013). Using our country-specific estimate for R_t, we defined the start and end date of the time-series on which we computed the early warning indicators as follows. We chose as start date the date at which R_t is at its lowest point before reaching $R_t = 1$ prior to the second wave. Similarly, we chose as end date the date at which R_t is at its maximum

[1] We analyzed countries in the EU, excluding Spain because of a strong weekend reporting effect that presented difficulties for model convergence, as well as the United Kingdom.

(before going down again) after it crosses $R_t > 1$. Panel (a) in Figure 1 illustrates this selection procedure on a simulated example, with the black line showing R_t and the vertical blue lines indicating its respective minimum and maximum after the first wave receded. We chose this criterion for two reasons. First, after R_t drops below 1, it continues to decrease in all European countries, and we would thus expect early warning indicators to fall, rather than rise. Panel (a) in Figure 1 shows a characteristic bifurcation delay (see also Section 2.3) that describes that cases lag behind the equilibrium value consistent with $R_t < 1$. Choosing for the starting date the time of the minimum value of R_t before R_t rises again allows the system to come closer to its new equilibrium value. Similarly, choosing to end the interval with the maximum of R_t after it crosses the threshold should predispose the analysis toward detecting early warning indicators because it yields the greatest possible length of the time-series and because of the bifurcation delay (see Brett et al., 2017; Dibble et al., 2016, and Section 2.3). Overall, our selection criterion is biased in favor of detecting critical slowing down.

Figure 2 shows the reported (gray) and estimated true number of cases (black) across European countries, with vertical blue lines indicating the segment of the time-series for which we calculated early warning indicators. Figures 6-10 in Appendix A provide a more detailed picture, showing European countries together with their estimated effective reproduction numbers.

Detrending and estimation of early warning indicators. As illustrated in Panel (b) and (c) in Figure 1, we detrended the time segment of interest and then estimated early warning indicators using backwards rolling windows with a uniform kernel (i.e., equally weighted past observations) and window sizes δ_1 and δ_2, respectively. A backward rolling window only uses data from the past to estimate the current value of a particular statistic. For example, to estimate the mean at time point t, we calculate:

$$\bar{y}_t = \frac{1}{\delta_1} \sum_{j=t-\delta_1}^{t} y_j \ ,$$

where y_j is the number of reported cases at a particular time point j (see black line in Panel (b) in Figure 1, for an example). Other early warning indicators we studied were variance, coefficient of variation, index of dispersion, autocovariance, autocorrelation, decay time, skewness, kurtosis, and first differenced variance (for mathematical definitions, see Brett et al., 2018, Table 3). All of these indicators require an estimate of the mean, and so we first estimated the mean and then estimated the particular early warning indicator using a rolling window size of δ_2. For example, the variance at time point t, which is shown in Panel (c) in Figure 1, is calculated as:

$$s_t = \frac{1}{\delta_2} \sum_{j=t-\delta_2}^{t} (y_j - \bar{y}_j)^2 \ .$$

We conducted sensitivity analyses with rolling windows of size $\delta_1 \in [2, 4, \ldots, 18, 20]$ for detrending and rolling windows of size $\delta_2 \in [5, 10, \ldots, 45, 50]$ for indicator estimation using the R package *spaero* (O'Dea, 2016). A window size of 10, for example, means that the previous ten data points are being used to compute the statistic at the current time point. To create a sampling distribution under the null hypothesis of no increase in the early warning indicators that respects the temporal ordering of the data, we fitted an ARMA(p, q) model to the country-specific data. We selected the best fitting model using AIC and subsequently generated 500 surrogate time-series from it, computed the early warning indicators as outlined above, and estimated their rank correlation with time (Kendall's τ). This resulted in the sampling distribution under the null assumption of stationarity, which allowed us to test the actually observed Kendall's τ against a significance level α (Dakos et al., 2012).

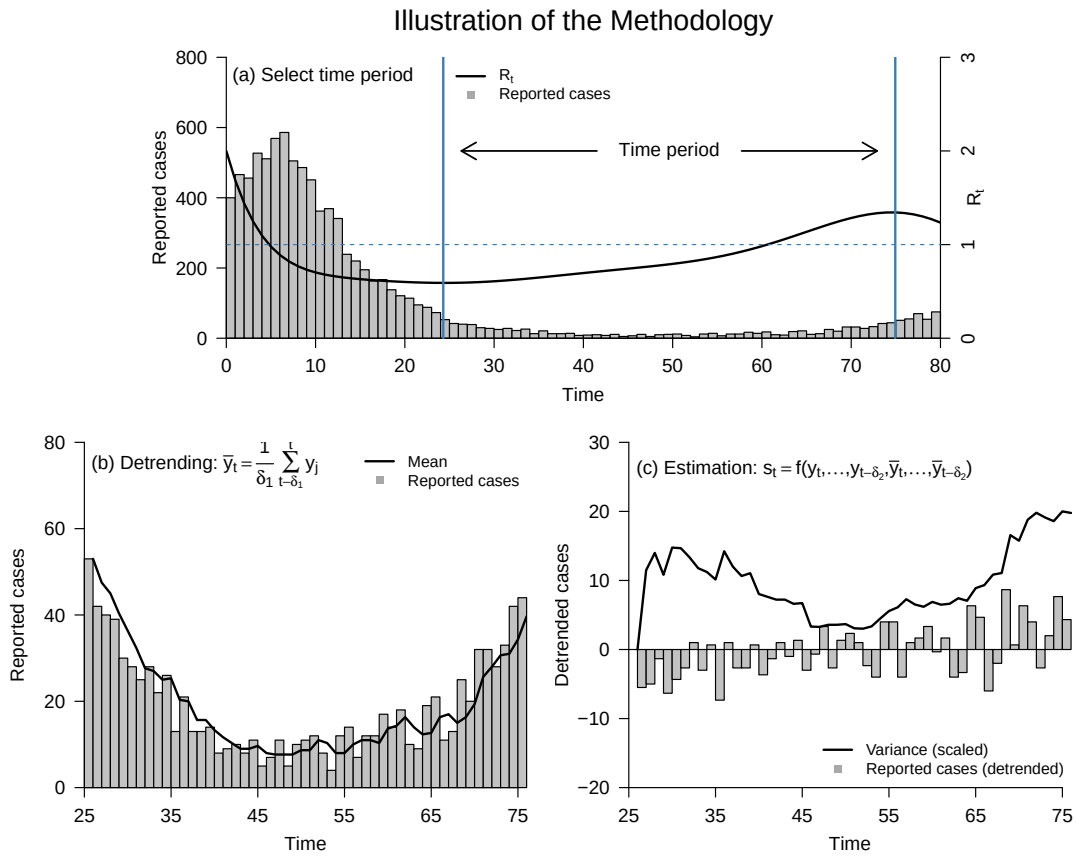

Figure 1: **Illustration of our methodology on simulated data.** Panel (a) shows reported cases (gray) and R_t (black). Vertical blue lines indicate the minimum and maximum R_t after the first wave receded. Panel (b) shows reported cases (gray) during the selected time period and an estimate of the mean (black) using a rolling window of size $\delta_1 = 3$. Panel (c) shows detrended cases (gray) and an estimate of the (scaled) variance (black) using a rolling window of size $\delta_2 = 10$.

2.2 Results

Figure 3 reports results for European countries for $\delta_1 = 4$ and $\delta_2 = 25$. It shows the value of Kendall's τ across all early warning indicators, coloring in red the countries for which τ was either significantly smaller or significantly larger than values generated from the best-fitting country-specific ARMA(p, q) at $\alpha = 0.05$. Notably, many countries displayed a significant *decrease* in the mean, with some showing a decreases in the variance, autocovariance, autocorrelation, and decay time. Several countries exhibited a significant *increase* in the coefficient of variation, which is given by the standard deviation divided by the mean, and in the dispersion index, which is given by the variance divided by the mean. Hence, early warning indicators that were found to display notable signal across a number of countries are the mean, variance, or combinations thereof. Figures 11-20 in Appendix B show sensitivity analyses for the ten early warning indicators across different rolling window sizes for detrending and estimation, indicating that the pattern

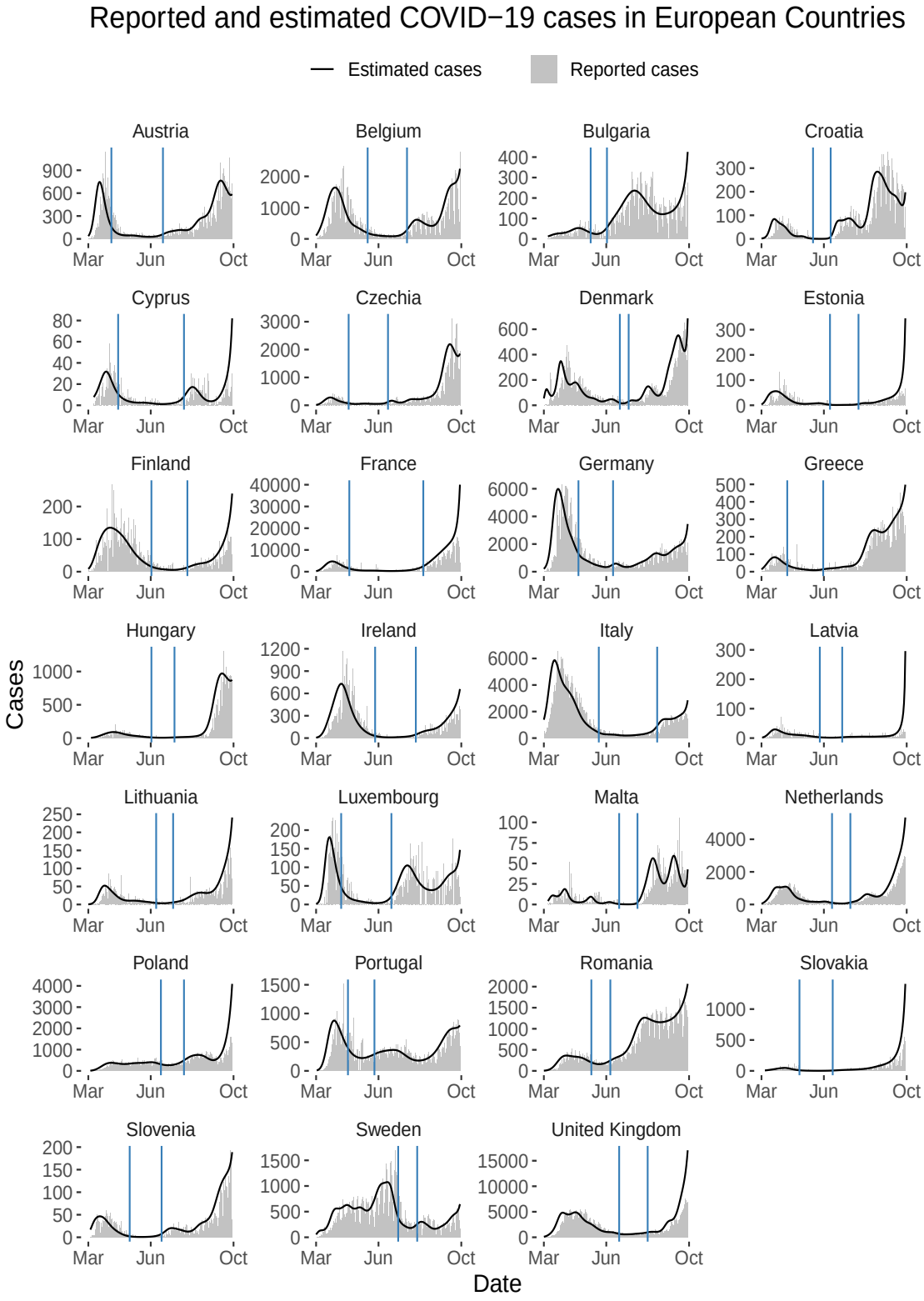

Figure 2: **Reported cases across European countries.** Top: Reported cases (gray) and posterior mean of inferred infected cases (black) for European countries. Vertical blue lines indicate the portion of the time-series for which early warning indicators are computed.

shown in Figure 3 is robust to different choices of these hyperparameters.

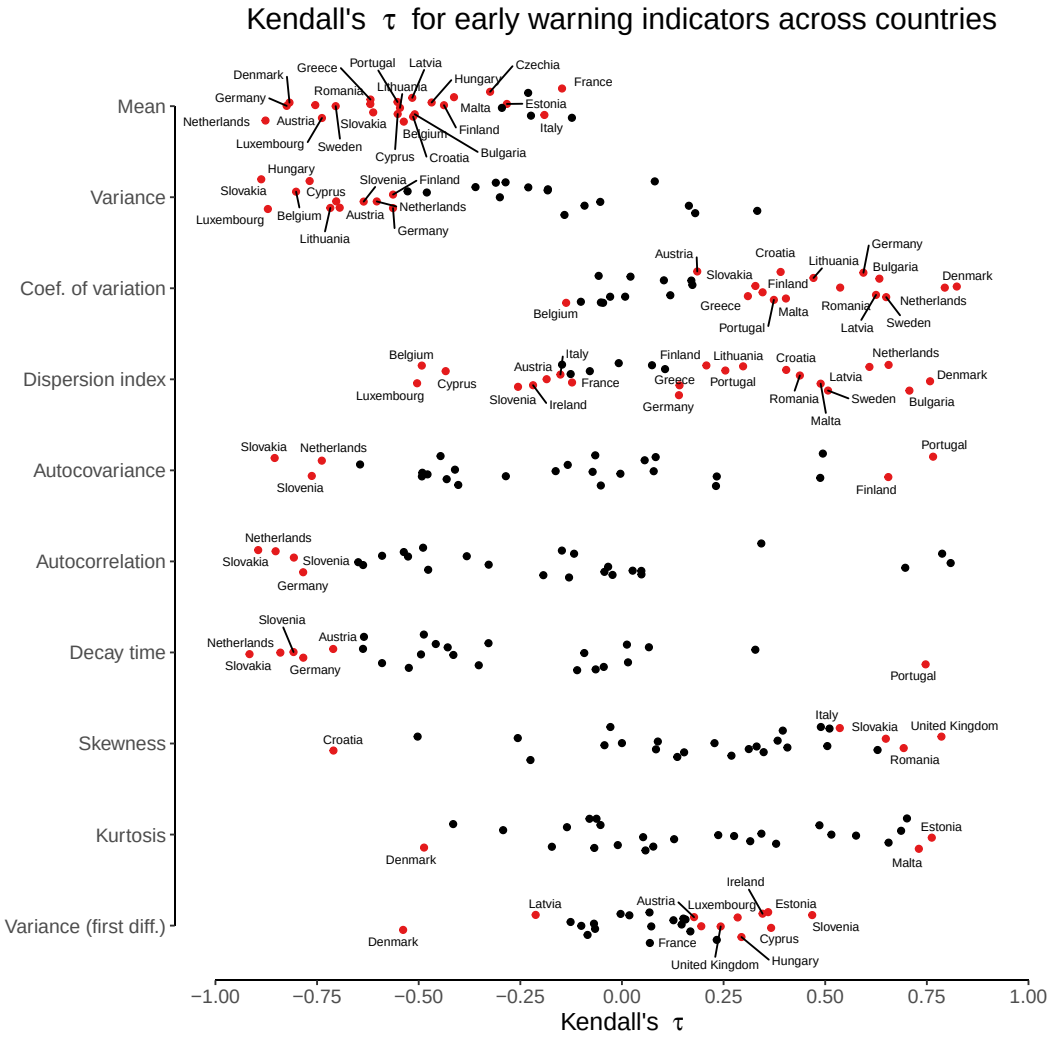

Figure 3: **Summary of results across countries and indicators.** The figure displays Kendall's τ across European countries for ten early warning indicators using $\delta_1 = 4$ for detrending and $\delta_2 = 25$ for indicator estimation. Red points indicate countries for which τ was either significantly smaller or larger than expected under a stationary time-series at $\alpha = 0.05$.

Table 1 shows the number of significantly rising or falling early warning indicators, the length of the selected time-series, the start of the second wave, and the respective posterior mean for R_t. From theory we expect all early warning indicators to rise except the coefficient of variation (Brett et al., 2018), yet we find that half of the indicators show a tendency to fall instead.

We conducted simulations to investigate possible reasons that could underlie the poor performance of early warning indicators to anticipate the second COVID-19 wave. In what follows, we first illustrate how early warning indicators perform under ideal conditions, and then relax the separation of time scales to quantify the erosion in performance.

Country	No. significant ↑	No. significant ↓	Duration	$\mathbb{E}[R_{\min} \mid \mathcal{D}]$	$\mathbb{E}[R_{\max} \mid \mathcal{D}]$
Portugal	4	1	39	0.82	1.07
Finland	3	2	53	0.80	1.22
Malta	3	1	27	0.52	2.38
Romania	3	1	28	0.87	1.14
Netherlands	2	5	27	0.77	1.32
Slovakia	2	5	49	0.66	1.33
Austria	2	4	76	0.63	1.25
Germany	2	4	51	0.77	1.22
Denmark	2	3	13	0.66	1.39
Croatia	2	2	26	0.38	2.85
Latvia	2	2	33	0.77	1.23
Lithuania	2	2	25	0.83	1.19
Bulgaria	2	1	24	0.84	1.31
Estonia	2	1	42	0.61	1.45
Greece	2	1	53	0.81	1.19
Sweden	2	1	28	0.68	1.17
United Kingdom	2	0	42	0.86	1.10
Slovenia	1	5	47	0.63	1.48
Cyprus	1	3	97	0.72	1.42
Luxembourg	1	3	74	0.67	1.48
France	1	2	109	0.77	1.27
Hungary	1	2	34	0.79	1.18
Italy	1	2	86	0.80	1.31
Ireland	1	1	60	0.72	1.28
Belgium	0	4	58	0.83	1.38
Czechia	0	1	58	0.79	1.38
Poland	0	0	34	0.91	1.16

Table 1: The number of significantly rising or falling early warning indicators, out of a total possible of ten, for European countries together with the length of the selected time-series and the respective posterior mean of R_t. $\mathcal{D}$ denotes the (country-specific) data, see Figure 2.

2.3 Model

We illustrate early warning indicators in the context of a first outbreak that is closely followed by a second one by simulating from a stochastic SEIR model calibrated to COVID-19 using the *pomp* R package (King et al., 2016). In particular, let $S(t), E(t), I(t), R(t)$ denote the number of individuals in the susceptible, exposed, infectious, and recovered compartment at time point t, respectively, and let $\Delta N_{S\to E}$, $\Delta N_{E\to I}$, and $\Delta N_{I\to R}$ denote the number of individuals that have transitioned from one compartment to another during the time interval $[t, t+\Delta t]$. The model is updated according to

$$\Delta N_{S\to E} \sim \text{Binomial}\left(S(t), 1 - e^{-\lambda S(t)\Delta_t}\right) \tag{1}$$

$$\Delta N_{E\to I} \sim \text{Binomial}\left(E(t), 1 - e^{-\sigma E(t)\Delta_t}\right) \tag{2}$$

$$\Delta N_{I\to R} \sim \text{Binomial}\left(I(t), 1 - e^{-\gamma I(t)\Delta_t}\right) \quad , \tag{3}$$

where we assume an average incubation and infectious period of $1/\sigma = 5.2$ days (Li et al., 2020) and $1/\gamma = 10$ days (CDC, 2021). The force of infection is given by

$$\lambda = \beta(t)\frac{I(t)}{N} + \eta(t) \ , \tag{4}$$

where $\eta(t)$ is the sparking rate, which we assume to be 0 until day 50, from which point onward cases are imported with a rate of $\eta = 1/50{,}000$. Our goal here is not to produce a simulation model that accurately tracks the COVID-19 outbreak, but instead to investigate critical slowing down in a standardized system that we understand well. To do so, we wish to force R_t to create a multi-wave epidemic. We achieve this by changing $\beta(t)$ accordingly, compensating for the depletion of the susceptible population by multiplying with $S_0/S(t)$ at time point t. Lastly, we assume that each infected person is reported without delay.

To illustrate the phenomenon of critical slowing down under ideal conditions, we start with $10,000$ infected persons out of $N = 1,000,000$ and $R_0 = 3$. This results in a first outbreak, which is rapidly brought down through control measures that we model as bringing R_t down to 0.50 within 25 days. We then force R_t to remain at this low value for 200 days, and then allow it to rise linearly to $R_t = 1$, forcing a second wave. This simulation mimics the situation at the start of the pandemic where the first outbreak caught countries by surprise and lockdown was the key mitigation measure that substantially reduced the effective reproductive number. In the illustration, mitigation measures are maintained for a long period of time. However, in reality mitigation measures were slowly relaxed towards the summer, and with no vaccination in place together with imported infections and increased mixing, the system could not reach a disease-free equilibrium and the reproductive number increased again. This led to a second outbreak in the fall of 2020 in virtually all European countries. Our simple model adequately describes this general pattern as shown in Figure 4a. In particular, the left column in Figure 4a shows the two waves of transmission and their associated early warning indicators, while the right panels in Figure 4a show a similar situation except that no second outbreak occurs. In contrast to the situation with a second wave, variance and autocorrelation do not rise in this case. This illustration demonstrates that under these conditions a second epidemic wave can be anticipated using nonparametric early warning indicators.

It is known that epidemiological systems can experience a *bifurcation delay*, which describes the transient trajectory of an epidemic as its attracting equilibrium changes. One consequence of bifurcation delays is that the time for a large outbreak to settle to its equilibrium even after crossing $R_t > 1$ can be considerable. Dibble et al. (2016) studied bifurcation delays for disease emergence, and Figure 4 indeed shows that it takes a while for the system to show a significant rise in cases even after $R_t > 1$ (see Hungary in Figure 2, for a possible example with regards to COVID-19). As can be seen in Figure 4, a bifurcation delay also occurs for disease elimination. In particular, for $R_t < 1$ the disease is not endemic and the stable equilibrium consists of a number of new cases that depends on the rate of at which cases are imported. There is, however, a substantial delay between the point at which $R_t < 1$ for the first time and a low number of newly reported cases. This means that early warning indicators computed immediately from the period after R_t first declines to less than 1, are tracking a transient far from equilibrium and thus do not provide information about the return rate to equilibrium from small perturbations, i.e. the phenomenon of critical slowing down.

To understand the extent to which this bifurcation delay may influence the performance of early warning indicators, we decreased the time interval for which $R_t = 0.50$ from 200 days (Figure 4a) to 50 days (Figure 4b). We find that both the variance and autocorrelation first *increase* and then *decrease* in the case of both a second outbreak (left panels) and in the case of no second outbreak (right panels). The variance then rises slightly prior to the second wave, a

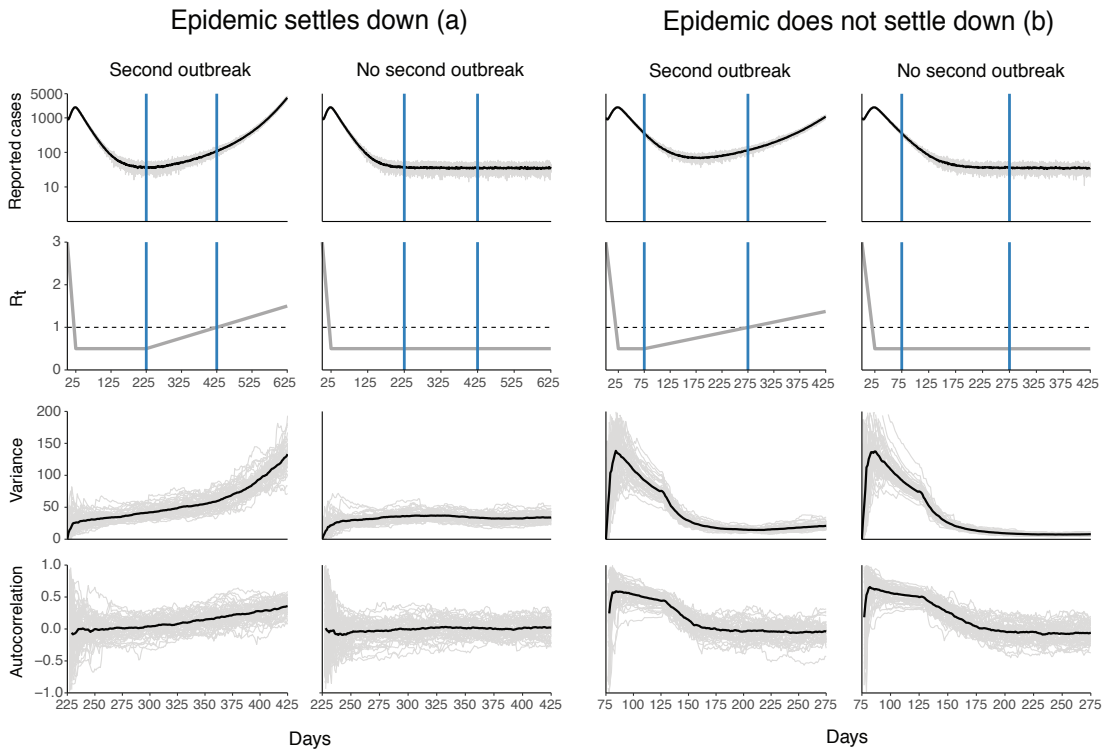

Figure 4: **Signatures of critical slowing down in a simulated second-wave epidemic.** Panels (a): Reported cases of a first outbreak followed by a second (top left) or no outbreak (top right) together with the forcing of R_t (below). Vertical blue lines indicate the period on which we compute the early warning indicators autocorrelation and variance, shown in the two bottom panels. The increase in variance and autocorrelation in the left panels is the manifestation of critical slowing down. Shown are 50 simulation runs (gray) together with their mean (black). Panels (b): Same, but for the case that the epidemic has not settled down after a first outbreak before a second one is forced.

pattern that does not occur for the autocorrelation, nor for the indicators in case of no second wave. This pattern hints at the fact that the bifurcation delay at elimination will interfere with the detection of critical slowing down if the system is not allowed to settle to its new equilibrium because the magnitude of the transient is commensurate with (or larger than) the magnitude of the fluctuations.

2.4 Simulation setup

We conducted additional simulations to systematically assess the extent to which these patterns impact the performance of early warning indicators. The forcing of R_t in the previous illustrations depends on five parameters: the value of R_0; the value of the lowest point R_t reaches; the time it takes R_t to reach it; the time for which R_t stays at the lowest point; and the time it takes R_t to reach criticality again. We again assume that $R_0 = 3$ and that it takes the system 25 days to reach its lowest point of $R_t = 0.50$, but we vary the number of days for which R_t is held constant to be $t_1 \in [25, 50, 100, 200]$ and the time it takes the system to reach $R_t = 1$

to be $t_2 \in [25, 30, \ldots, 95, 200]$. For comparison, we also simulate from a system that stays at $R_t = 0.50$ and does not exhibit a second outbreak. We match the length of the time-series on which we compute early warning indicators (t_2) in case of no outbreak to when an outbreak does occur. As before, backwards rolling windows with a uniform kernel were used for detrending and nonparametric estimation of the mean, variance, coefficient of variation, index of dispersion, skewness, kurtosis, autocovariance, autocorrelation, decay time, and first differences in variance. We used rolling windows a tenth the size of the duration for which R_t stays constant; that is, for $t_1 \in [25, 50, 100, 200]$ we used rolling windows of sizes $3, 5, 10$, and 20, respectively. For indicator estimation, we used rolling window sizes of 50, using the R package *spaero* (O'Dea, 2016). We simulated 500 trajectories for each setting and calculated the area under the curve (AUC), a measure of classification performance, for all indicators. For each indicator, we calculated its rank correlation with time (Kendall's τ), which indicates whether the early warning indicators rise or fall prior to reaching the critical point. The AUC can then be estimated as the probability that τ_{test} is larger than τ_{null} (Brett et al., 2018; Flach, 2016). A value of $|\text{AUC} - 1/2| = 0$ indicates chance performance, with AUC $< 1/2$ and AUC $> 1/2$ indicating a fall or rise in indicators prior to criticality, respectively. Theory predicts a pre-critical increase of all early warning indicators except the coefficient of variation (Brett et al., 2017; Brett et al., 2018). In addition to AUC, which requires comparing the indicator trend in the case of a second outbreak to the case of no second outbreak, we also use the method proposed by Dakos et al. (2012) and outlined in Section 2.1 to ascertain whether an indicator rises significantly. This more closely mimics the real-word situation where we do not have access to the counterfactual situation in which no outbreak occurred. We report the true positive rate (TPR), that is, the proportion of times we find $p < \alpha$ for each indicator and condition, using $\alpha = 0.05$.

2.5 Simulation results

Figure 5a shows that the performance of early warning indicators improves with the time it takes the epidemic to reach a second critical wave. For the case for which the system stays for 200 days at $R_t = 0.50$ (top panel of Figure 5a), we find that all indicators except the kurtosis and the index of dispersion performed well, with the mean and the variance performing best. The coefficient of variation, given by the ratio of the standard deviation to the mean, decreases prior to criticality, indicating that the mean rises more quickly than the standard deviation. Most early warning indicators perform worse when $R_t = 0.50$ for 100 days, yet the mean and variance still perform well overall. Interestingly, the slight decrease in performance in the variance implies a stronger decrease of the coefficient of variation and the index of dispersion especially when the system is forced more quickly (i.e., $t_2 < 125$).

For a period during which $R_t = 0.50$ of 50 days, the performance of the variance decreases, leading to an increasingly strong decrease in the coefficient of variation and the index of dispersion. When forcing is rapid (i.e., $t_2 < 100$), the autocovariance, autocorrelation, and decay time also begin to show a downward trend (AUC $< 1/2$) prior to reaching the critical point. These trends are exacerbated when the system stays at $R_t = 0.50$ for only 25 days. One may think that the simulation shows the *reverse* pattern than the empirical analysis, summarized in Figure 3, because the mean and variance show a positive AUC (hence they *increase* compared to the null simulation) while the mean and variance show a *decrease* in the empirical analysis. There is no contradiction, however, because the mean and variance do in fact decrease in case of a second wave, it is just that they *decrease less* compared to when there is no second wave, as can be seen in Figure 4b.

In the data, the median time for countries to go from their minimum R_t value after the first crossing to their maximum R_t value after the crossing was 42 days. Figures 6-10 further show

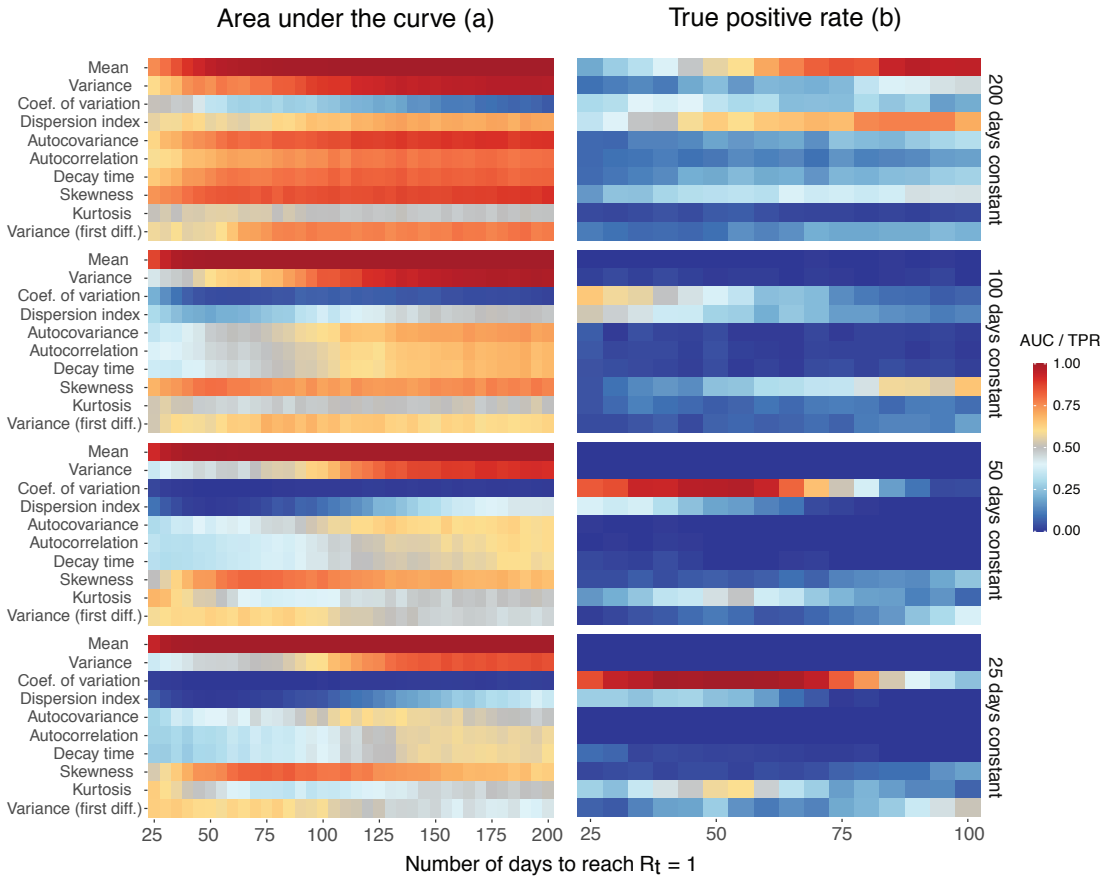

Figure 5: **Indicator performance across simulation settings.** Area under the curve (a) and true positive rate (b) for ten early warning indicators as the number of days for which $R_t = 0.50$ and the number of days it takes the system to reach $R_t = 1$ again vary. True positive rate is calculated by using the best-fitting ARMA(p, q) model to create a stationary null distribution and a decision criterion of finding a significant *increase* at $p < 0.05$.

that R_t basically never stays at a low constant value for a sustained period of time, but is forced immediately towards the critical point. Under the most realistic scenario in our simulation study ($t_1 = 25$ and $t_2 < 50$), many indicators perform poorly, yet we still find excellent performance of a rising mean and excellent performance of a falling coefficient of variation and index of dispersion. This does not imply, however, that they will lead to reliable warnings in practice. While we can quantify discriminatory power using AUC in simulations, in practice early warning indicators have to be *calibrated*. Figure 5b shows that testing for an indicator increase at $\alpha = 0.05$ based on a stationary null distribution created by using the best-fitting ARMA(p, q) model to the time-series under consideration is poorly calibrated, leading to an extremely poor true positive rate which mirrors the empirical results in Section 2.2. This is because the distribution of Kendall's τ under the stationary model is centered around zero, while the actually observed Kendall's τ is negative. As a result, hypothesis tests for an increase in indicator values are expected to suffer from extremely low statistical power in realistic situations. This problem may be exacerbated by a potentially poor fit of the model used to create the null distribution.

3 Discussion

Early warning signals based on the phenomenon of critical slowing have been suggested as a way to anticipate transitions in a wide range of dynamical systems, including the (re)emergence of infectious diseases. We analyzed whether a suite of indicators could have given early warning of the second COVID-19 wave in European countries. We found that the majority of indicators did not rise reliably, instead showing a pronounced decrease, a finding inconsistent with previous applications of the theory of critical slowing down. To understand this pattern, we conducted a simulation study in which we varied the time that is available for the system to settle at its new equilibrium after a first outbreak, as well as the speed with which a second wave is forced. We analyzed the performance of early warning indicators using the area under the curve to quantify classification performance and the true positive rate, using the same methodology with which we analyzed the empirical data. We found that classification performance suffered when the system had too little time to settle to its new (quasi-)equilibrium and the second wave is forced quickly (due to changing conditions in the population, such as reduced adherence to control measures), as we saw in the empirical data. Yet we also found that some indicators, such as the mean, continued to perform well (in terms of AUC) in contrast to what we observed in the empirical analysis. Using the same methodology as in the empirical analysis, however, we found a true positive rate of close to zero when testing for an increase in indicators, except for the coefficient of variation and index of dispersion, which is in line with our empirical results.

Our analyses suggest the following conclusions. First, violating a key assumption of early warning indicators based on critical slowing down — namely that the driver (R_t) changes slowly compared to the time it takes the system to return to its equilibrium after small external perturbations — dramatically reduces their performance. While this may be expected from theory, our analyses underscore this point and show that early warning indicators cannot be used to anticipate future outbreaks that are quickly forced after an initial wave. Second, as a consequence of the fact that the system is not allowed enough time to settle at its new stable equilibrium after an initial outbreak, the first part of the data used for early warning indicator estimation constitutes a transient. Hence there is a bifurcation delay not only after R_t crosses one from below, as previously observed and studied (e.g., Dibble et al., 2016), but also after R_t crosses one from above. If this transient is incorporated in the indicator estimation, then indicators will show a pronounced *decrease* rather than an *increase*. This does not imply, however, that we can use a decrease in indicators as a signal for a future outbreak that quickly follows an initial one, because such a decrease also occurs in case of no outbreak. The poor performance of early warning indicators in our empirical analysis is likely due to a combination of this transient phenomenon and the quick forcing of R_t. The only two indicators that showed a relatively consistent increase across countries are the coefficient of variation and the index of dispersion. This is likely due to the fact that the mean decreased more quickly than the standard deviation and the variance during the transient phase, leading to an increase in the indicators. In other words, the coefficient of variation and the index of dispersion likely increased for reasons other than critical slowing down. Third, our simulation study demonstrated that while early warning indicators can yield high discrimination (i.e., a high AUC), in practice they need to be calibrated. We found that the widely used methodology proposed by Dakos et al. (2012) with decision criterion $p < 0.05$ is poorly calibrated. This leads to poor performance consistent with our empirical results. The key issue is that the sampling distribution created under this methodology is not centered around a negative Kendall's τ (implying a decreasing trend) but a Kendall's τ of around zero (implying no trend). Thus the statistical power to reject the null hypothesis of no increase when actually observing a strong decrease in indicators is too low for these tests to be of practical value in realistic situations. Previous research also suggested that indicators can fail in the COVID-19

context (Proverbio et al., 2021).

Some limitations of this study should be kept in mind. Our empirical analysis takes the reported number of cases across European countries at face value. While we accounted for reporting delays, we disregarded any issues related to changes in reporting or testing that may affect the estimation of R_t. While the flexible method proposed by Abbott et al. (2020) renders any bias induced by a change of testing transient, any bias may have indeed changed the true value at which R_t crosses one. A more extensive analysis would look at all countries that experienced a second wave. However, we chose to limit ourselves to European countries because of the comparatively good reporting standards and the fact that there is sufficiently large heterogeneity in epidemic trajectories across European countries for the purposes of this study. On a similar note, because the time period between the end of the first and the beginning of the second wave was shorter than the time period it takes the system to settle at its new stable equilibrium after the first wave recedes in virtually all countries, we expect our findings to generalize well to non-European countries. We used an admittedly conservative criterion for date stamping the end of the first wave and the start of the second one to reduce the extent of the transient period we incorporate for indicator estimation. In particular, we chose the day at which R_t reaches its lowest value as starting point for the computation of early warning indicators. If anything, based on our finding that incorporating the transient decreases performance, our choice may be too charitable. We chose the end date for the indicator computation as the day at which R_t reaches its maximum after crossing one so as to increase the number of time points and reduce the extent of any bifurcation delay. If anything, this may again have been too generous. At the same time, while the epidemic unfolded quite distinctly in different European countries, R_t never stabilized at a low value and rose quickly after the first outbreak. These are far from the conditions under which to expect a reliable signal in early warning indicators, and our results should not be interpreted as a rejection of their potential in other applications, including other epidemics.

We used backwards rolling windows to avoid the use of data from the "future", and our results can thus translate to a situation in which indicators are computed in real-time. A critical issue when using nonparametric estimation concerns the choice of the size of the rolling windows (Dakos et al., 2012; Dessavre et al., 2019; Lenton et al., 2012). There is a trade-off between a window size that is too small, where estimation accuracy suffers, and a window size that is too large, where stationarity is (more severely) violated (Brett et al., 2017). If a model is available, Dessavre et al. (2019) find that detrending based on model simulation works well, but this route is unavailable as an epidemic unfolds for which accurate models do not yet exist. Similarly, while Miller et al. (2017) found that indicator performance was robust to seasonal forcing, the time scale of such seasonal forcing is much longer compared to the movements of R_t that were observed in some European countries, and which hence may have further reduced performance. We have addressed the issue of window size selection by reporting extensive sensitivity analyses. Our finding that indicators poorly anticipate the second COVID-19 wave is robust to different choices.

Critical slowing down is a phenomenon that has primarily been studied in low-dimensional systems. It is prominent in the study of ferromagnetism and the Lenz-Ising model (Brush, 1967), and has been known to proponents of catastrophe theory since at least the 1970s (Zeeman, 1976). Wissel (1984) suggested critical slowing down as a way to forecast the extinction of a population of rotifers (see also Dai et al., 2012; Drake & Griffen, 2010). Scheffer et al. (2009) brought significant attention to the idea of using critical slowing down as an early warning signal which led to a surge of interest across many fields. Yet there is the obvious question of whether we should expect a phenomenon that pertains primarily to low dimensional systems to occur in the high dimensional real-world. Infectious diseases do not spread in homogeneously mixed

populations with people being distinct only in terms of whether they are susceptible, exposed, infected, or recovered, as our simulation model assumes. Instead, infectious diseases spread between unique individuals on a network that is itself continuously changing. Studying the effect of test sensitivity and frequency on COVID-19 transmission, Larremore et al. (2021) find essentially no difference between a homogeneous compartment model and an agent-based model that is calibrated to New York City micro-census data. More relevant to our investigation, Brett et al. (2020) found that early warning indicators based on critical slowing down do indeed rise prior to an outbreak in high-dimensional network and agent-based models.

A related issue with early warning indicators based on critical slowing down concerns the decision criterion. When do we decide that a rise in indicators is "significant" and constitutes an early warning? In our empirical analysis, we chose a rise in trend to be significant at the $\alpha = 0.05$ level, but this may well require adaption to the specific case at hand. There is a difference between making a statistical inference (e.g., estimating Kendall's τ) and making a decision (e.g., restricting mitigation measures; Boettiger & Hastings, 2012). The latter requires calibration, which is understudied in the context of early warning indicators based on critical slowing down but essential to use in applications. Importantly, some indicators, such as the mean and variance, continue to rise even after R_t crosses one, as predicted by theory (O'Dea & Drake, 2019; Southall et al., 2020). Others are expected to peak at the point at which $R_t = 1$, although the exact maximum may not be clear (O'Dea et al., 2018). This means that it is hard to assess whether, say, a rise in the autocorrelation from 0.50 to 0.70 is already problematic, or whether one should wait until it reaches, say, 0.90 (if it ever will). The extent to which indicators such as autocorrelation rise also depends on a number of reporting details such as the frequency of reporting. It is therefore impossible to provide general guidelines for use in applications. Simulation studies that incorporate reporting issues and focus on specific diseases may provide further insight (Brett et al., 2018; Tredennick et al., under review).

Early warning indicators based on critical slowing down promise to be a quite general and low-cost tool to monitor the emergence and elimination of infectious diseases (e.g., Drake & Hay, 2017; Harris et al., 2020; Tredennick et al., under review). It is understudied how well these indicators perform compared to other tools that may be used as early warning signals. In the context of COVID-19, it seems plausible that by making stronger assumptions about the dynamics of the system or using system-external information such as mobility would lead to much better early warning systems. Simply estimating R_t and forecasting whether and when $R_t > 1$ may be a similarly low-cost but potentially more reliable approach. Conceptually, however, it is not so clear that one would like to have an early warning indicator signalling that R_t is about to cross one. This is due to two related reasons. First, because of the bifurcation delay, it may take weeks or months for the actual outbreak to occur. A method that is able to incorporate this bifurcation delay and produce an early warning of an actual exponential increase in cases may therefore be preferable. Ideally, such a method produces a probabilistic assessment of an outbreak, which can then feed into further decision making. Second, the simple fact that R_t crosses one does not imply that a second wave is incumbent. Instead, it may stay there for a while or fall again, as it did in several European countries during the current pandemic. One cannot impose strong mitigation measures to curb virus spread whenever $R_t > 1$. All this points to a more continuous approach in which multiple, system-external factors are taken into account to assess the risk of future outbreaks. Early warning indicators may be a part of this risk assessment toolbox for (re)emerging diseases when an outbreak is slowly forced — but not, as we have shown, when one outbreak follows closely after another.

Author Contributions. H.H. and D.B. suggested to investigate early warning signals in the context of COVID-19. F.D. analyzed the data, conducted the simulation study, and wrote the first draft of the manuscript. J.D. provided valuable advice on the data analysis and simulation study. F.D., H.H., D.B., and J.D. wrote the manuscript. All authors read and approved the submitted version of the paper. They also declare that there were no conflicts of interest.

Funding. F.D. was supported by ZonMw project 10430022010001.

References

Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., Munday, J. D., Meakin, S., Doughty, E. L., Chun, J. Y., Et al. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, *5*(112), 112.

Boettiger, C., & Hastings, A. (2012). Quantifying limits to detection of early warning for critical transitions. *Journal of the Royal Society Interface*, *9*(75), 2527–2539.

Brett, T. S., Ajelli, M., Liu, Q.-H., Krauland, M. G., Grefenstette, J. J., van Panhuis, W. G., Vespignani, A., Drake, J. M., & Rohani, P. (2020). Detecting critical slowing down in high-dimensional epidemiological systems. *PLoS Computational Biology*, *16*(3), e1007679.

Brett, T. S., Drake, J. M., & Rohani, P. (2017). Anticipating the emergence of infectious diseases. *Journal of The Royal Society Interface*, *14*(132), 20170115.

Brett, T. S., O'Dea, E. B., Marty, É., Miller, P. B., Park, A. W., Drake, J. M., & Rohani, P. (2018). Anticipating epidemic transitions with imperfect data. *PLoS Computational Biology*, *14*(6), e1006204.

Brett, T. S., & Rohani, P. (2020). Dynamical footprints enable detection of disease emergence. *PLoS Biology*, *18*(5), e3000697.

Brush, S. G. (1967). History of the Lenz-Ising Model. *Reviews of Modern Physics*, *39*(4), 883–893.

CDC. (2021). *Interim Guidance on Ending Isolation and Precautions for Adults with COVID-19* [Accessed on 06-07-2021]. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, *178*(9), 1505–1512.

Dablander, F., Pichler, A., Cika, A., & Bacilieri, A. (2020). Anticipating critical transitions in psychological systems using early warning signals: Theoretical and practical considerations. https://doi.org/10.31234/osf.io/5wc28

Dai, L., Vorselen, D., Korolev, K. S., & Gore, J. (2012). Generic indicators for loss of resilience before a tipping point leading to population collapse. *Science*, *336*(6085), 1175–1177.

Dakos, V., Carpenter, S. R., Brock, W. A., Ellison, A. M., Guttal, V., Ives, A. R., Kéfi, S., Livina, V., Seekell, D. A., & van Nes, E. H. (2012). Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PloS One*, *7*(7), e41010.

Dessavre, A. G., Southall, E., Tildesley, M. J., & Dyson, L. (2019). The problem of detrending when analysing potential indicators of disease elimination. *Journal of Theoretical Biology*, 183–193.

Dibble, C. J., O'Dea, E. B., Park, A. W., & Drake, J. M. (2016). Waiting time to infectious disease emergence. *Journal of The Royal Society Interface*, *13*(123), 20160540.

Drake, J. M., Brett, T. S., Chen, S., Epureanu, B. I., Ferrari, M. J., Marty, É., Miller, P. B., O'Dea, E. B., O'regan, S. M., Park, A. W., & Rohani, P. (2019). The statistics of epidemic transitions. *PLoS Computational Biology*, *15*(5), e1006917.

Drake, J. M., & Griffen, B. D. (2010). Early warning signals of extinction in deteriorating environments. *Nature*, *467*(7314), 456–459.

Drake, J. M., & Hay, S. I. (2017). Monitoring the path to the elimination of infectious diseases. *Tropical Medicine and Infectious Disease*, *2*(3), 20.

Drake, J. M., O'Regan, S. M., Dakos, V., Kéfi, S., & Rohani, P. (2020). Alternative stable states, tipping points, and early warning signals of ecological transitions, In *Theoretical Ecology*. Oxford University Press.

Flach, P. A. (2016). ROC analysis, In *Encyclopedia of Machine Learning and Data Mining*. Springer.

George, D. B., Taylor, W., Shaman, J., Rivers, C., Paul, B., O'Toole, T., Johansson, M. A., Hirschman, L., Biggerstaff, M., Asher, J., Et al. (2019). Technology to advance infectious disease forecasting for outbreak management. *Nature Communications*, *10*(1), 1–4.

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., Et al. (2020). Practical considerations for measuring the effective reproductive number, Rt. *PLoS Computational Biology*, *16*(12), e1008409.

Harris, M. J., Hay, S. I., & Drake, J. M. (2020). Early warning signals of malaria resurgence in Kericho, Kenya. *Biology Letters*, *16*(3), 20190713.

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K. T., Edmunds, W. J., Frost, S. D., Funk, S., Et al. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227).

Kéfi, S., Guttal, V., Brock, W. A., Carpenter, S. R., Ellison, A. M., Livina, V. N., Seekell, D. A., Scheffer, M., van Nes, E. H., & Dakos, V. (2014). Early warning signals of ecological transitions: Methods for spatial patterns. *PloS One*, *9*(3), e92097.

King, A. A., Nguyen, D., & Ionides, E. L. (2016). Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *Journal of Statistical Software*, *69*(1), 1–43.

Kuehn, C. (2011). A mathematical framework for critical transitions: Bifurcations, fast–slow systems and stochastic dynamics. *Physica D: Nonlinear Phenomena*, *240*(12), 1020–1035.

Larremore, D. B., Wilder, B., Lester, E., Shehata, S., Burke, J. M., Hay, J. A., Tambe, M., Mina, M. J., & Parker, R. (2021). Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science Advances*, *7*(1), eabd5393.

Lenton, T. M. (2011). Early warning of climate tipping points. *Nature Climate Change*, *1*(4), 201–209.

Lenton, T. M., Livina, V., Dakos, V., van Nes, E., & Scheffer, M. (2012). Early warning of climate tipping points from critical slowing down: Comparing methods to improve robustness. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *370*(1962), 1185–1204.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K., Lau, E., Wong, J., Et al. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *The New England Journal of Medicine*, *382*(13), 1199–1207.

Miller, P. B., O'Dea, E. B., Rohani, P., & Drake, J. M. (2017). Forecasting infectious disease emergence subject to seasonal forcing. *Theoretical Biology and Medical Modelling*, *14*(1), 1–14.

Morens, D. M., & Fauci, A. S. (2013). Emerging infectious diseases: Threats to human health and global stability. *PLoS Pathog*, *9*(7), e1003467.

Morens, D. M., Folkers, G. K., & Fauci, A. S. (2004). The challenge of emerging and re-emerging infectious diseases. *Nature*, *430*(6996), 242–249.

O'Brien, D. A., & Clements, C. F. (2021). Early warning signals predict emergence of COVID-19 waves. *medRxiv*. https://doi.org/10.1101/2021.06.24.21259444

O'Dea, E. B. (2016). spaero: Software for Project AERO. https://cran.r-project.org/web/packages/spaero/index.html

O'Dea, E. B., & Drake, J. M. (2019). Disentangling reporting and disease transmission. *Theoretical Ecology*, *12*(1), 89–98.

O'Dea, E. B., Park, A. W., & Drake, J. M. (2018). Estimating the distance to an epidemic threshold. *Journal of the Royal Society Interface*, *15*(143), 20180034.

O'Regan, S. M., & Burton, D. L. (2018). How Stochasticity Influences Leading Indicators of Critical Transitions. *Bulletin of Mathematical Biology*, *80*(6), 1630–1654.

O'Regan, S. M., & Drake, J. M. (2013). Theory of early warning signals of disease emergenceand leading indicators of elimination. *Theoretical Ecology*, *6*(3), 333–357.

O'Regan, S. M., O'Dea, E. B., Rohani, P., & Drake, J. M. (2020). Transient indicators of tipping points in infectious diseases. *Journal of the Royal Society Interface*, *17*(170), 20200094.

Proverbio, D., Kemp, F., Magni, S., & Goncalves, J. (2021). Performance of early warning signals for disease emergence: A case study on COVID-19 data. https://doi.org/10.1101/2021.03.30.21254631

Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D., Kandula, S., Brooks, L. C., Crawford-Crudell, W., Gibson, G. C., Et al. (2019). Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US. *PLoS Computational Biology*, *15*(11), e1007486.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, *461*(7260), 53–59.

Scheffer, M., Carpenter, S. R., Dakos, V., & van Nes, E. H. (2015). Generic indicators of ecological resilience: Inferring the chance of a critical transition. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 145–167.

Southall, E., Tildesley, M. J., & Dyson, L. (2020). Prospects for detecting early warning signals in discrete event sequence data: Application to epidemiological incidence data. *PLoS Computational Biology*, *16*(9), e1007836.

Tredennick, A., O'Dea, E., Ferrari, M., Rohani, P., & Drake, J. M. (under review). Anticipating disease emergence and elimination: A test of early warning signals using empirically based models.

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., Zhang, Q., Chowell, G., Simonsen, L., Vespignani, A., Et al. (2018). The RAPIDD Ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, *22*, 13–21.

Wissel, C. (1984). A universal law of the characteristic return time near thresholds. *Oecologia*, *65*(1), 101–107.

Zeeman, E. C. (1976). Catastrophe theory. *Scientific American*, *234*(4), 65–83.

A Estimation of R_t across European Countries

Figures 6-9 show countries and their estimated effective reproductive number, with vertical lines indicating the time period on which we computed early warning indicators.

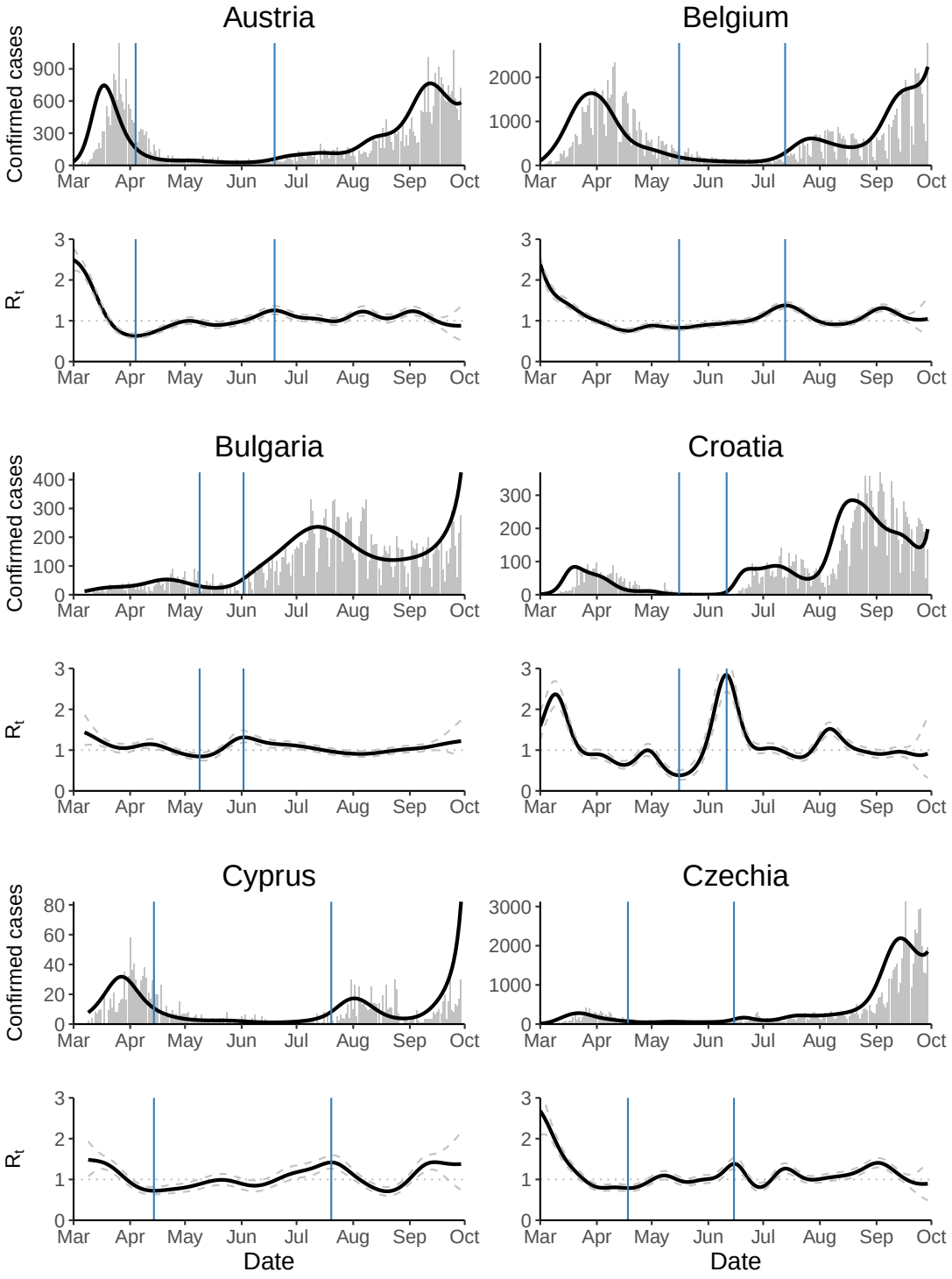

Figure 6: Shows reported cases (gray) and posterior mean of inferred infected cases (black) as well as posterior mean and 95% credible interval of R_t for various countries. Vertical blue lines indicate the time-series on which early warning indicators are computed, see main text.

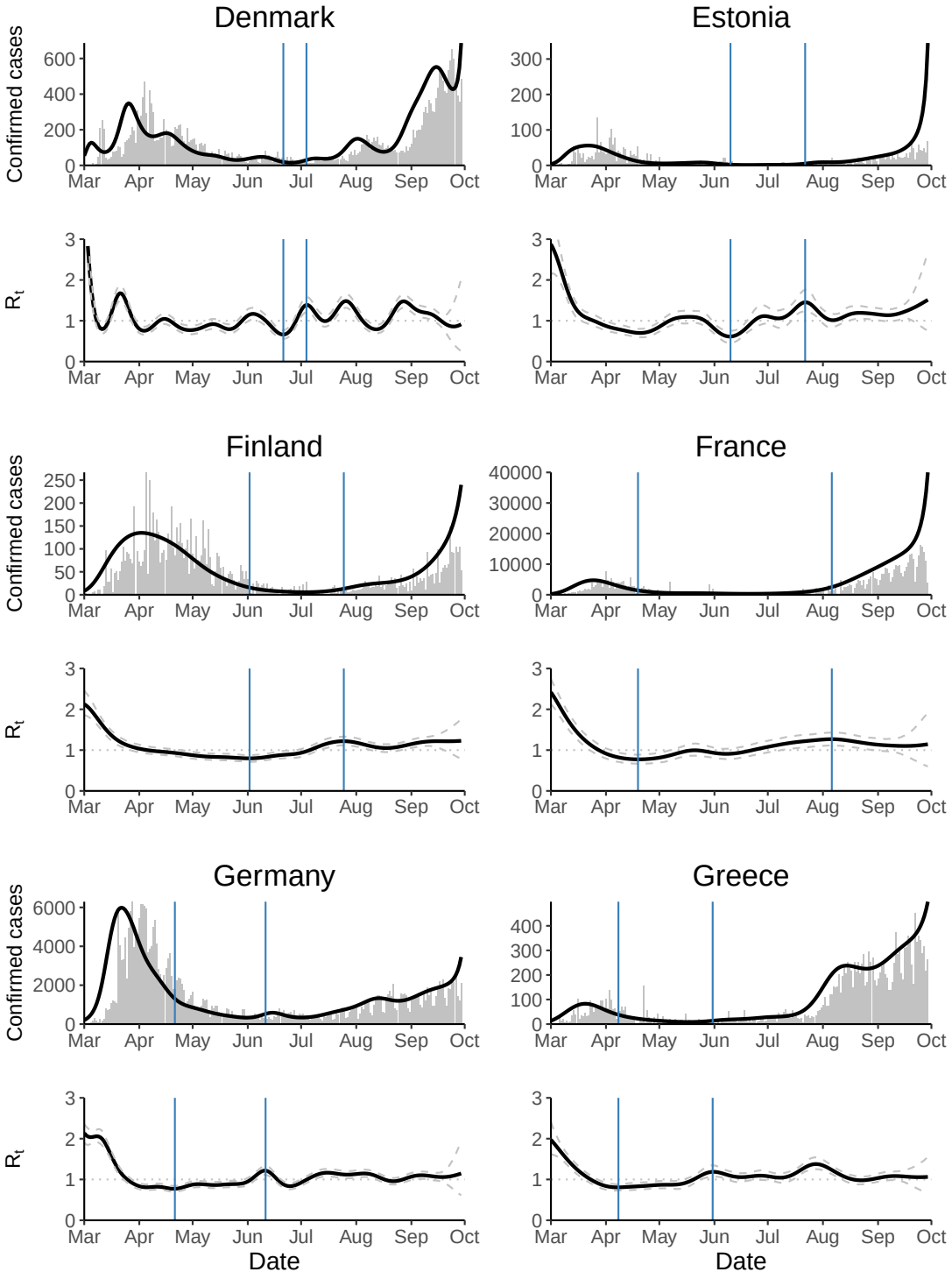

Figure 7: Shows reported cases (gray) and posterior mean of inferred infected cases (black) as well as posterior mean and 95% credible interval of R_t for various countries. Vertical blue lines indicate the time-series on which early warning indicators are computed.

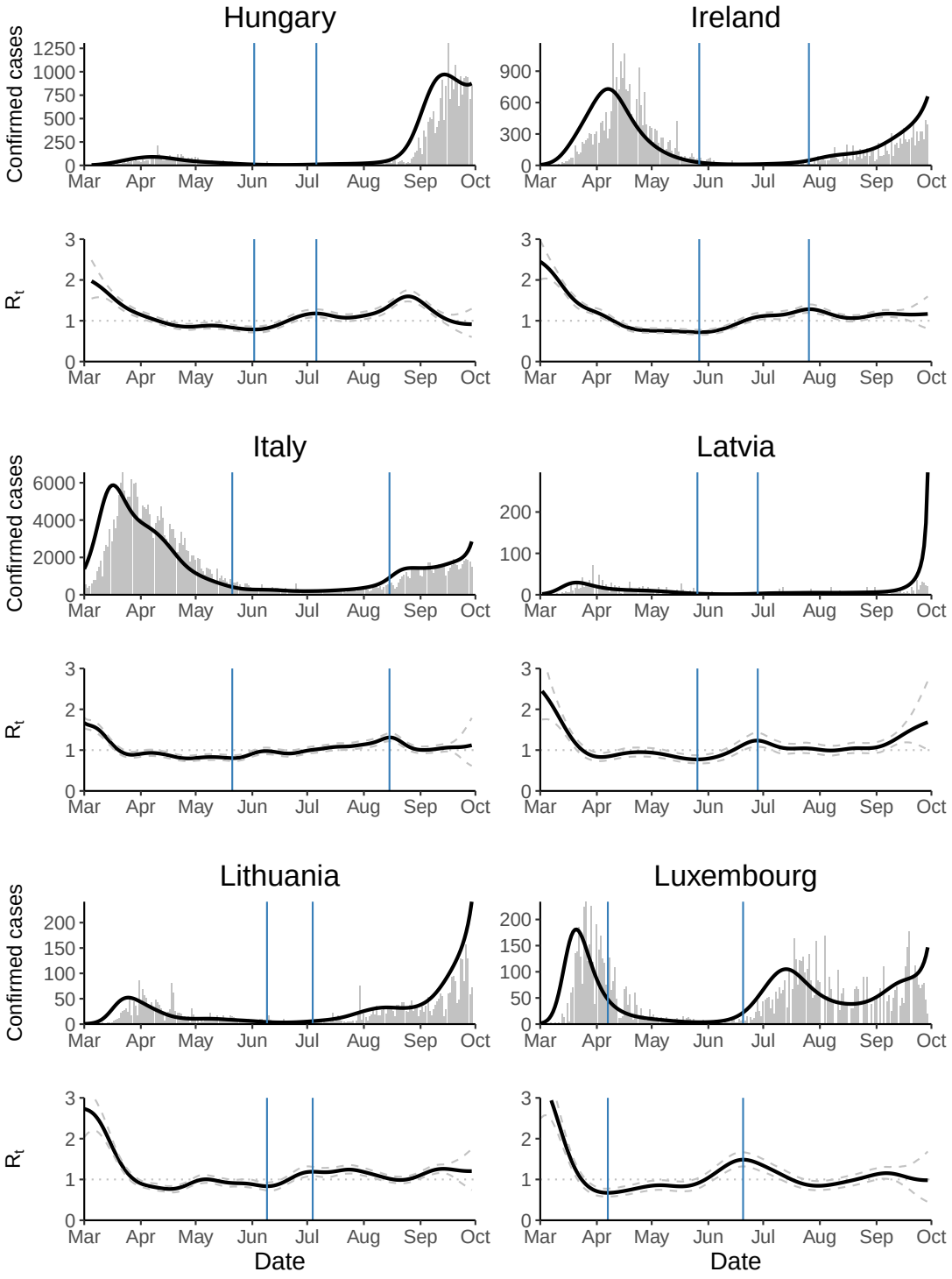

Figure 8: Shows reported cases (gray) and posterior mean of inferred infected cases (black) as well as posterior mean and 95% credible interval of R_t for various countries. Vertical blue lines indicate the time-series on which early warning indicators are computed.

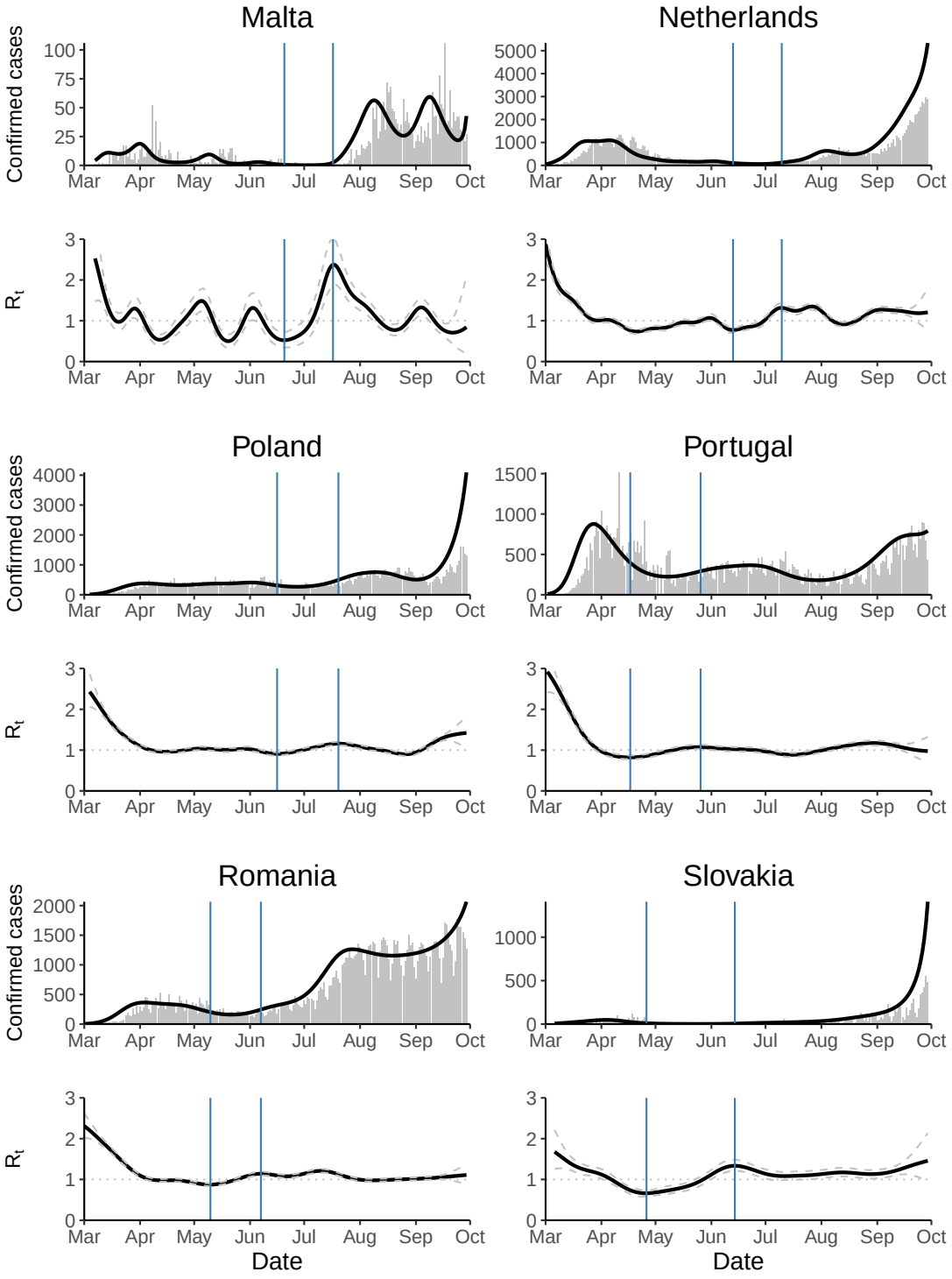

Figure 9: Shows reported cases (gray) and posterior mean of inferred infected cases (black) as well as posterior mean and 95% credible interval of R_t for various countries. Vertical blue lines indicate the time-series on which early warning indicators are computed.

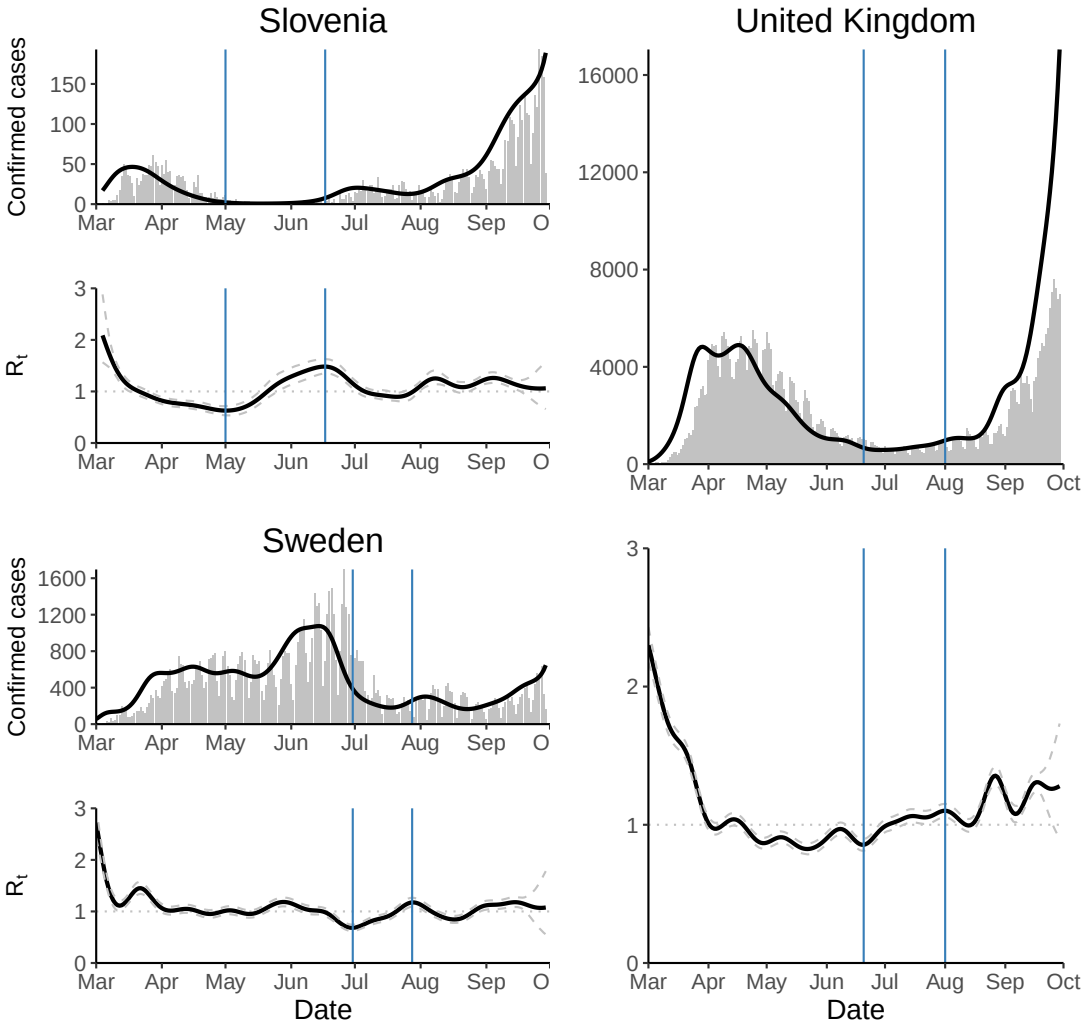

Figure 10: Shows reported cases (gray) and posterior mean of inferred infected cases (black) as well as posterior mean and 95% credible interval of R_t for various countries. Vertical blue lines indicate the time-series on which early warning indicators are computed.

B Sensitivity Analyses

Figures 11-20 show sensitivity analyses for the ten early warning indicators across different rolling window sizes for detrending and estimation.

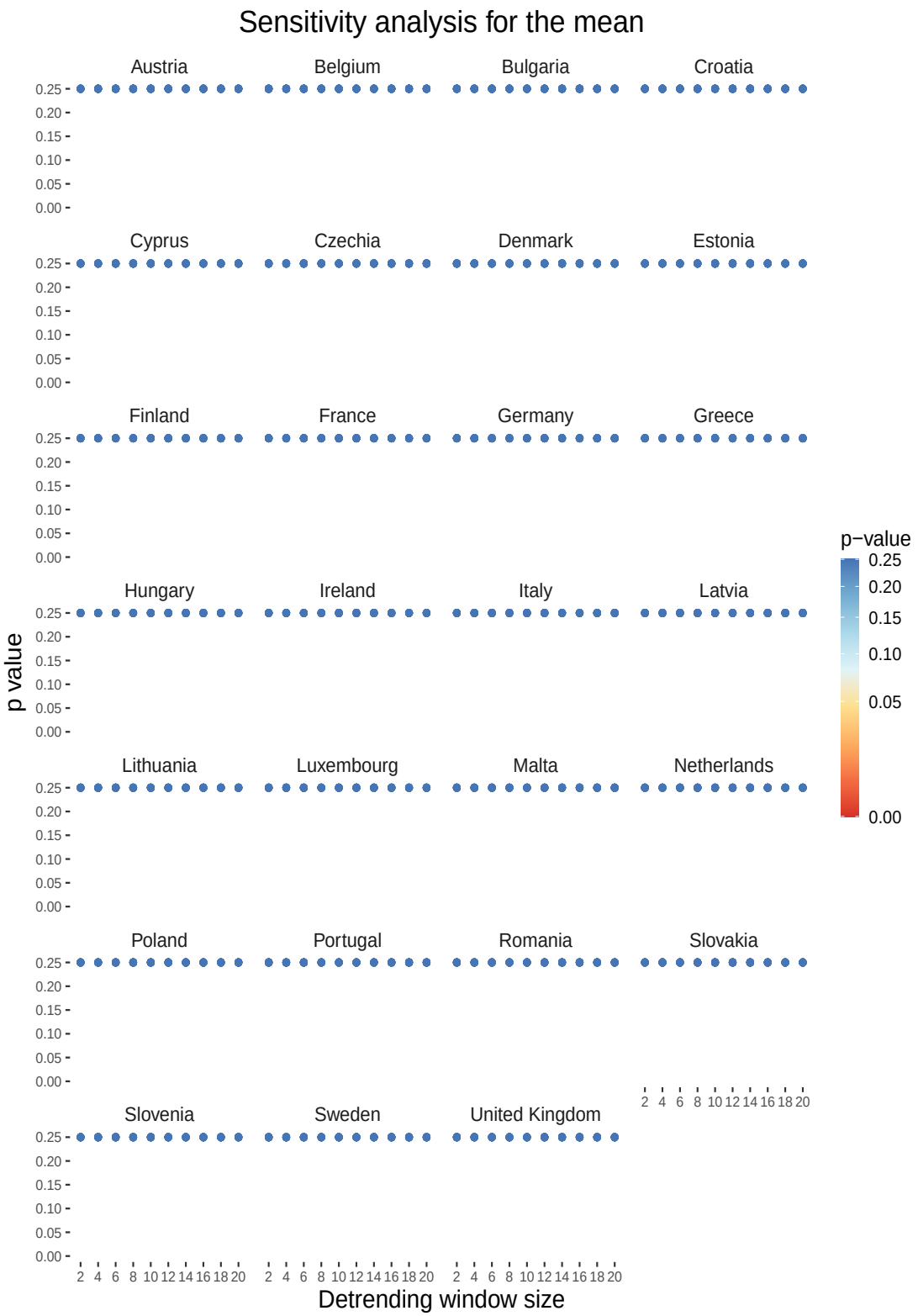

Figure 11: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the mean is significantly larger than expected under the null across detrending rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

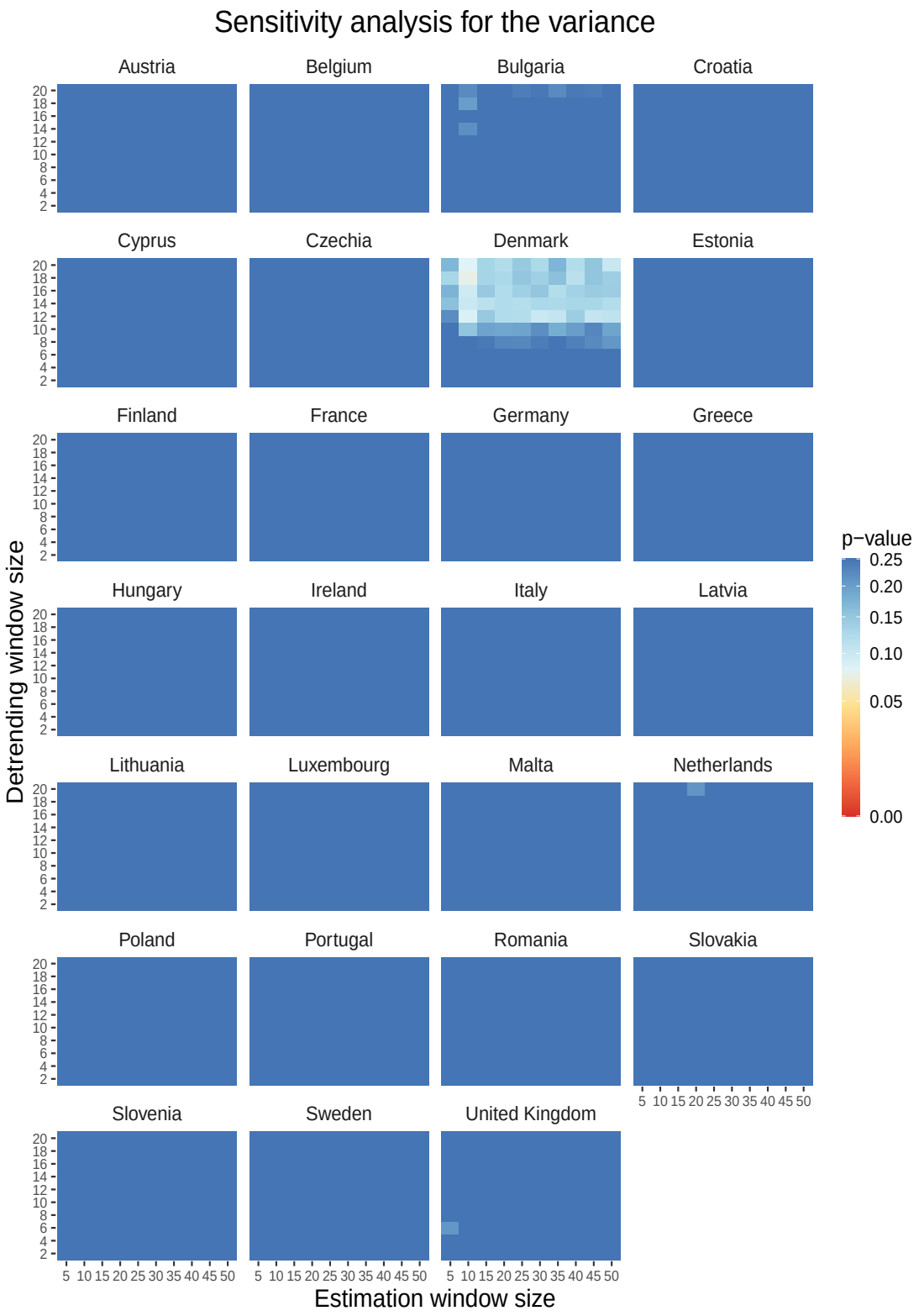

Figure 12: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the variance is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

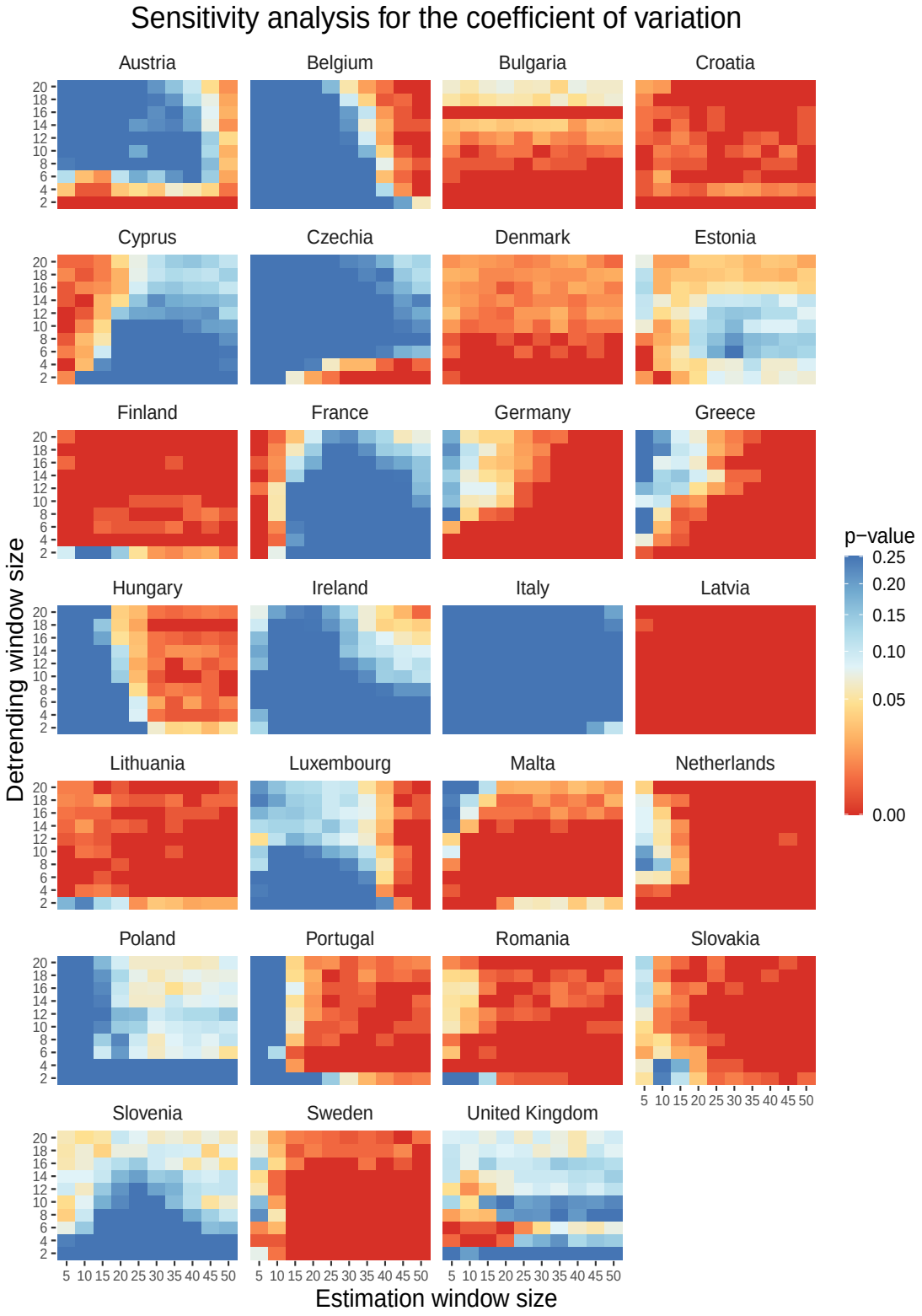

Figure 13: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the coefficient of variation is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

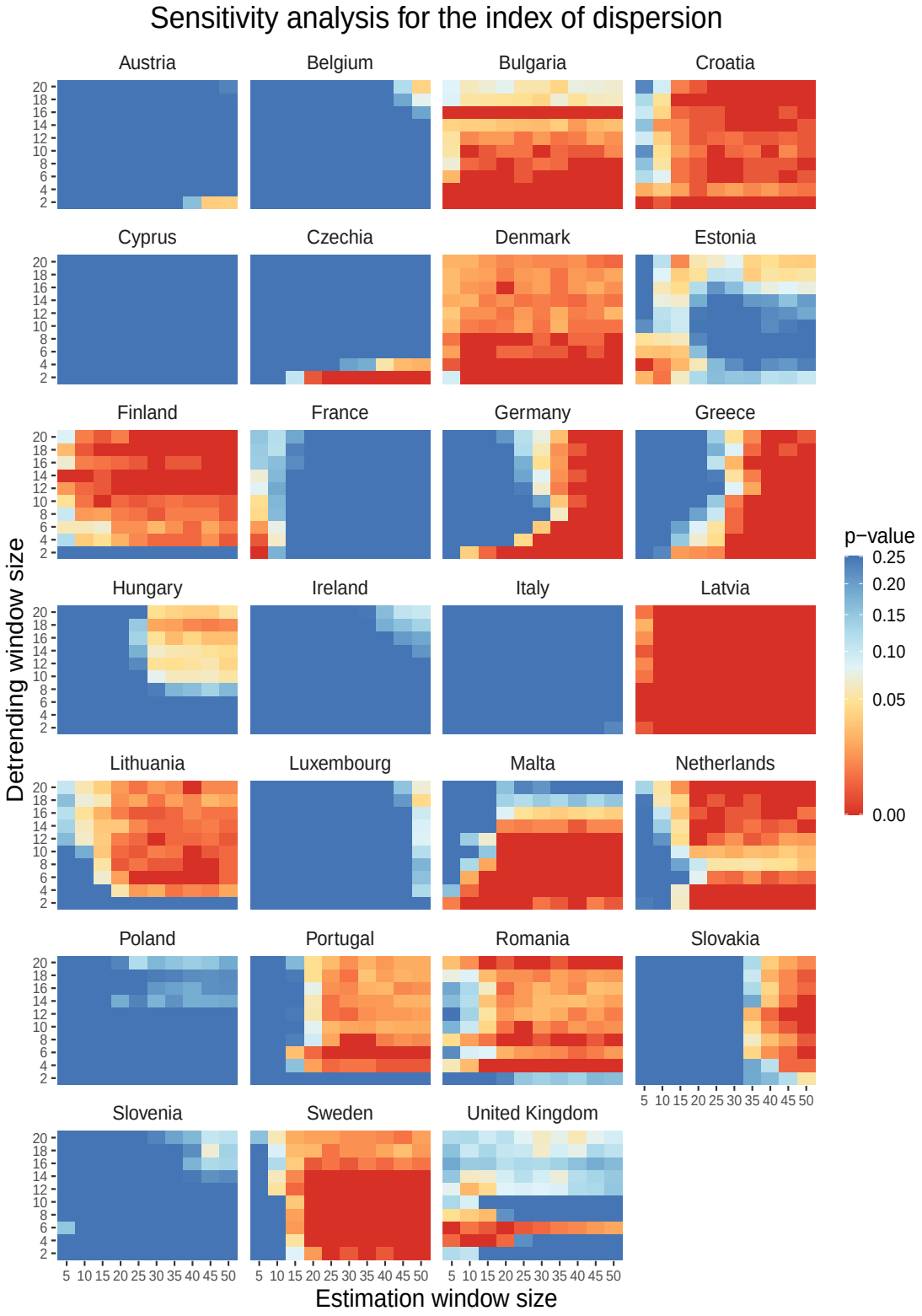

Figure 14: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the index of dispersion is significantly value than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

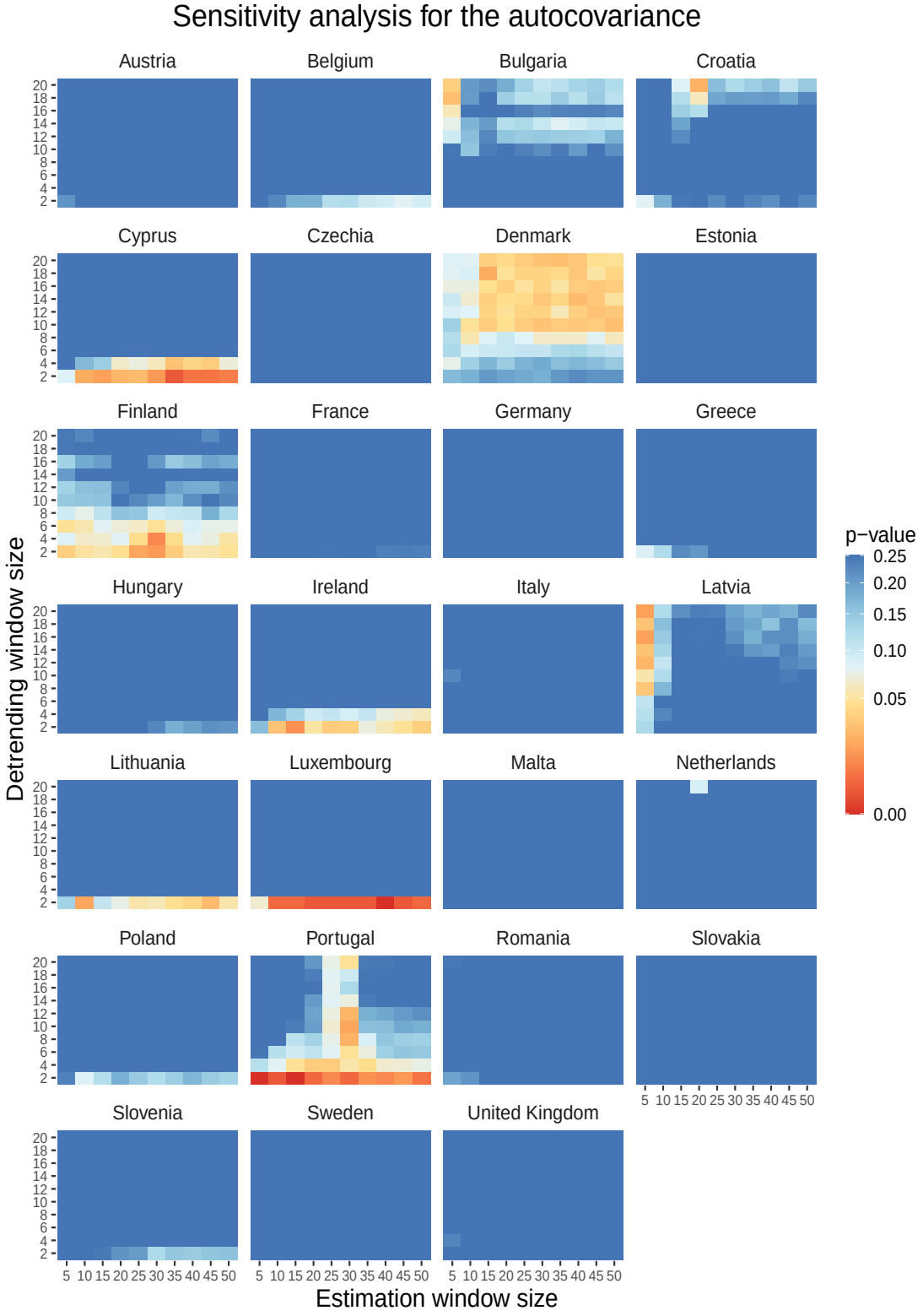

Figure 15: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the autocovariance is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

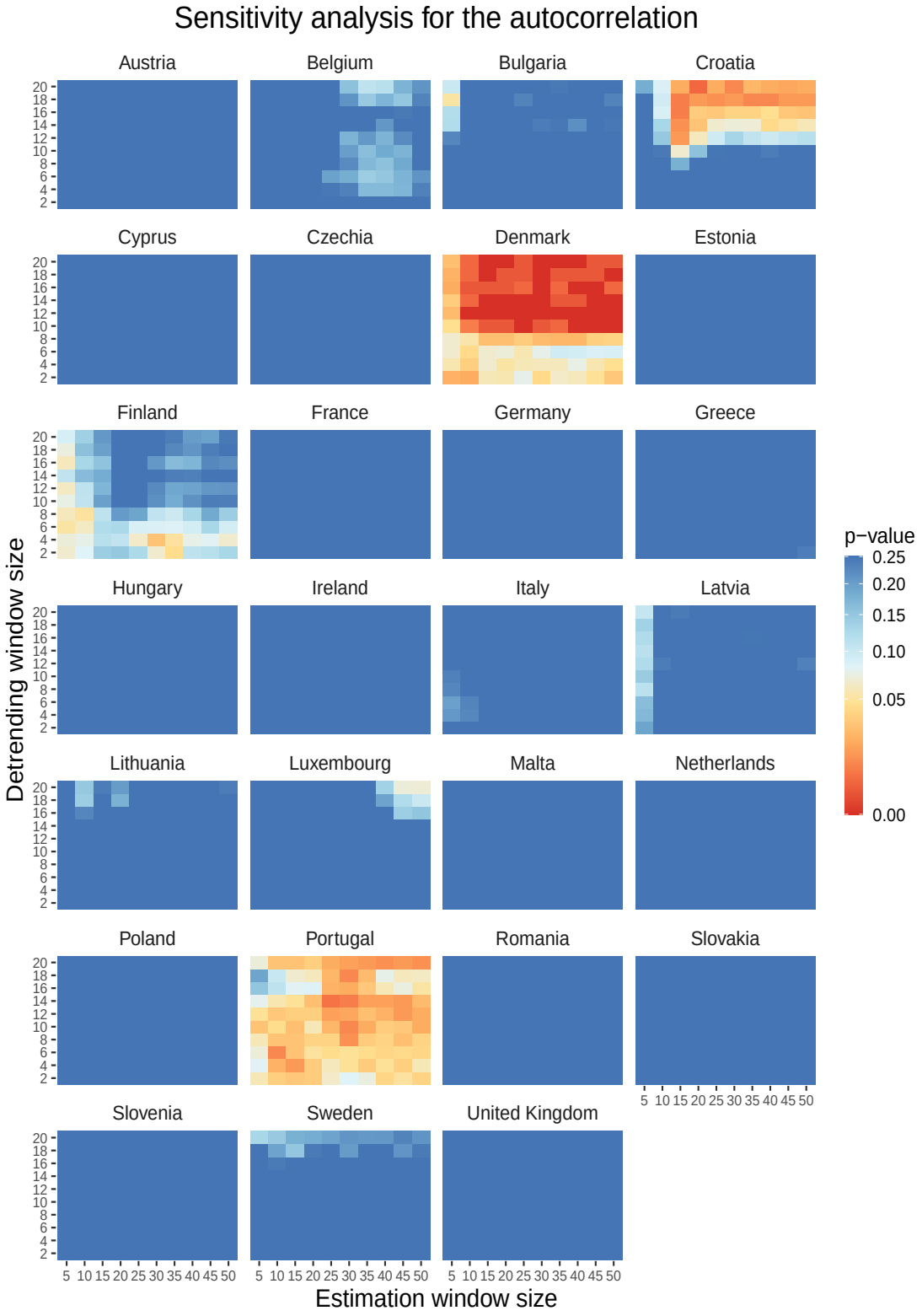

Figure 16: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the autocorrelation is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

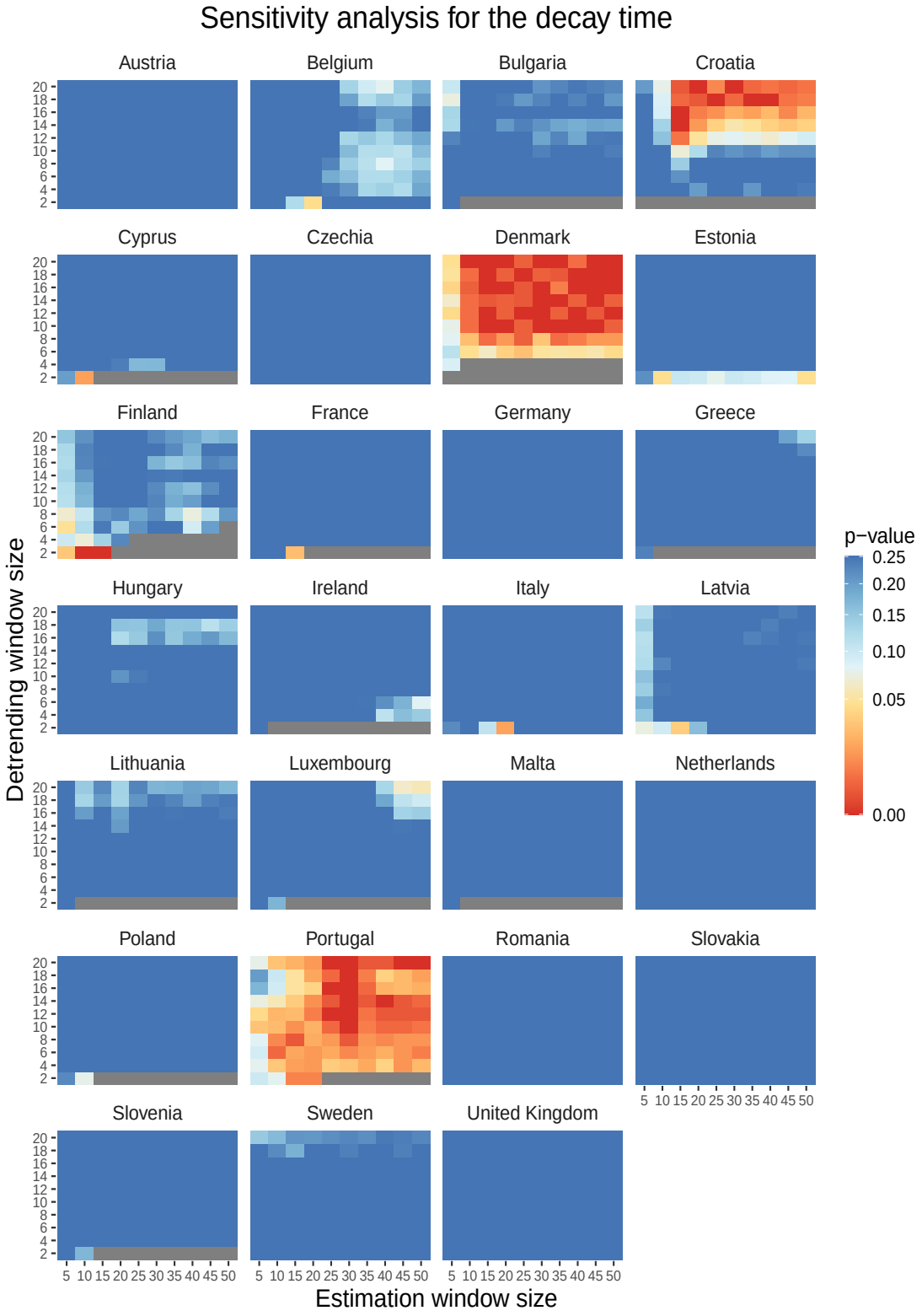

Figure 17: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the decay time is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

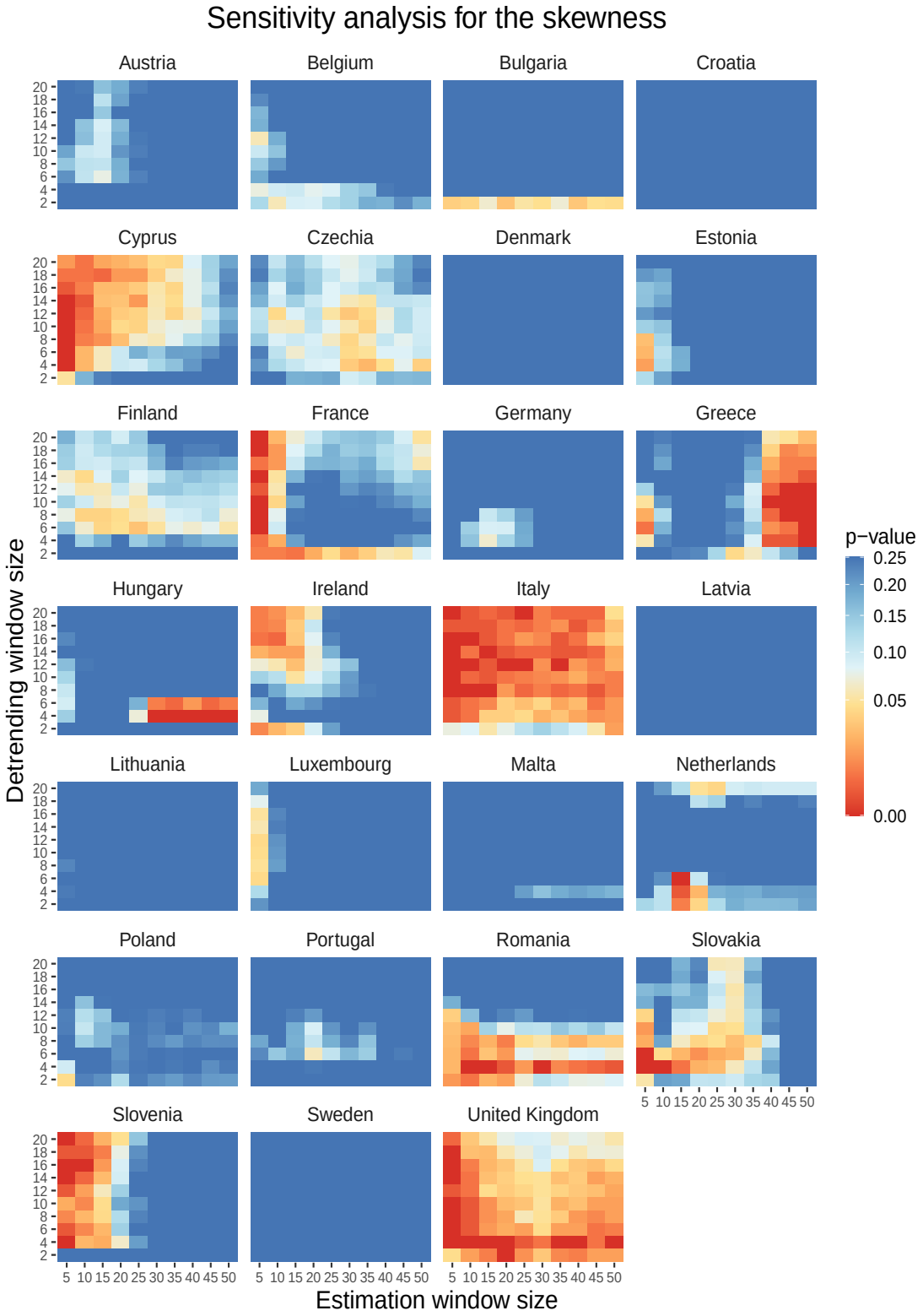

Figure 18: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the skewness is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

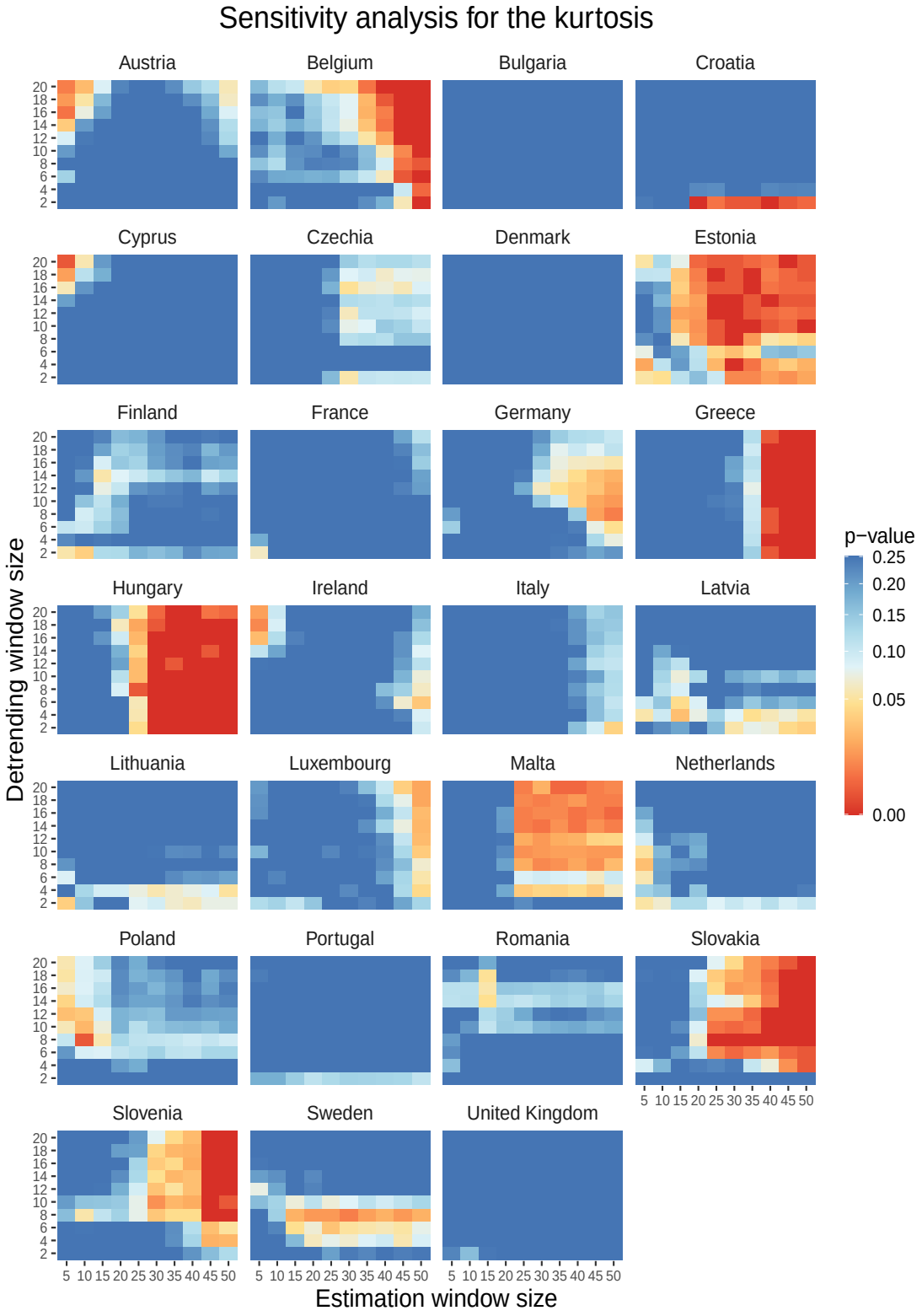

Figure 19: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the kurtosis is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.

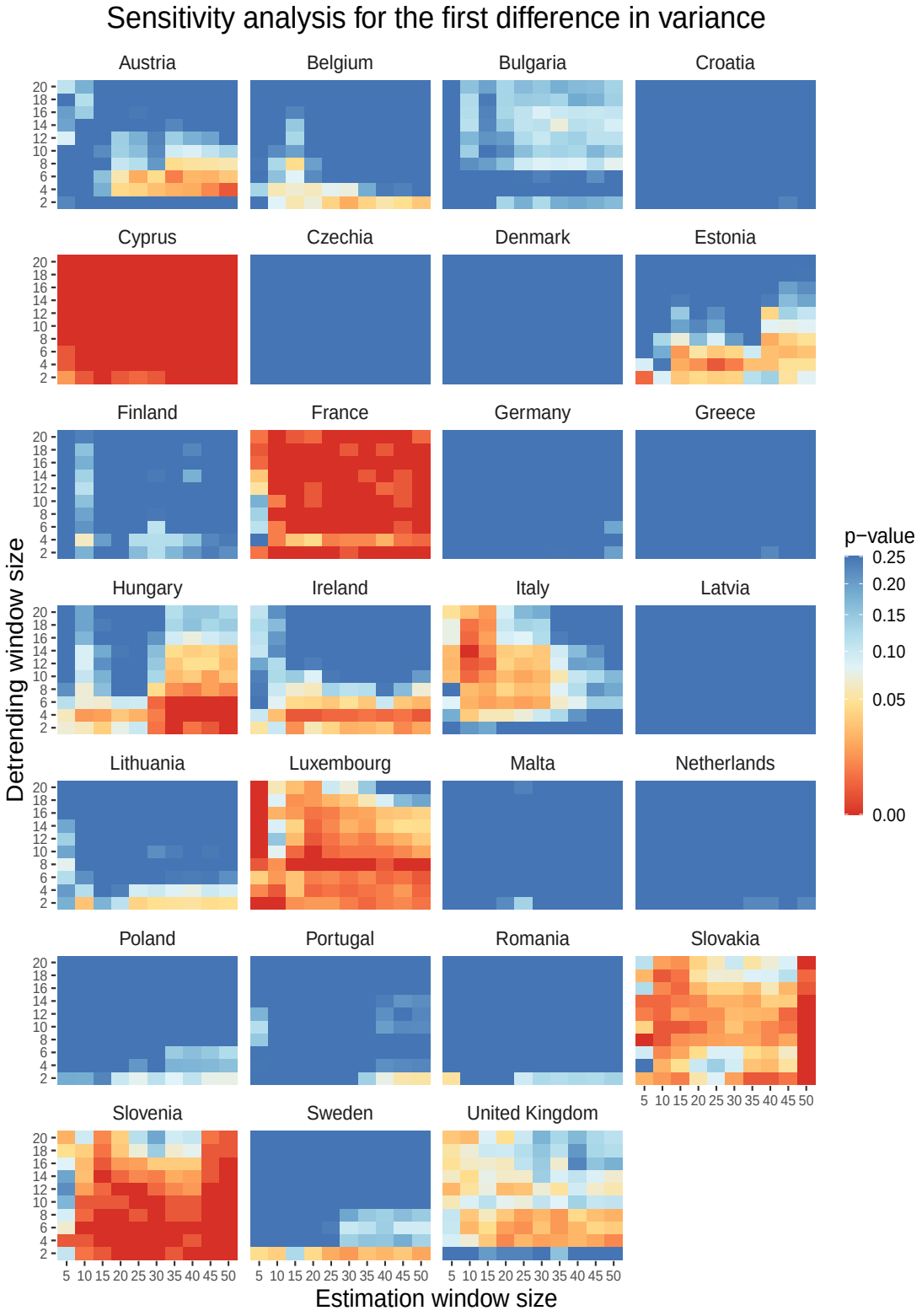

Figure 20: Shows bootstrapped p-values indicating whether the observed Kendall's τ in the first differences in the variance is significantly larger than expected under the null across rolling window sizes. Note that $p = 0.25$ in the legend means $p \geq 0.25$.