Reply to reports on article JBES-P-2022-0567

First of all, we would like to thank the editor, the associate editor and the reviewers for their many comments and suggestions which were very helpful to improve the paper. In the revision, we have addressed all comments of the referees and have rewritten the paper accordingly. Please find our point-by-point responses to the referees below. In our responses, we have indicated where alterations to the manuscript have been made to account for the points raised.

Reply to the editor

1. *The reviewers have diversified opinions, and they raise many issues. Although I think you can address many of them, there is one main issue I don't see how to handle: As the AE points out, you do not have a compelling case for the null hypothesis that time trends are identical across multiple time series. Like the AE, I can't think of different economic time series sharing the same trend. I appreciate your discussions on clustering much more, but the same issue arises within a group.*
   There is a small econometric literature on co-trending and co-breaking where common time trends and breaks are used to capture comovements of multiple time series. Rather than time trends being identical, they are proportional to each other. That might be more plausible for modeling multiple economic time series. Because this may be difficult in your framework, I do not mean to impose it. Instead, I am pointing out that such literature may or may not be helpful.
   Although I doubt the plausibility and usefulness of the null hypothesis, I would like to give the benefit of the doubt. I am willing to consider a revision of your paper, though with no guarantee that it will ultimately be accepted. I ask you to address all of the reviewers' concerns. Most importantly, please make a much more compelling and convincing case for the null hypothesis, which may be very difficult. I should note that this is not a typical R&R and is a rather weak R&R.

   We fully agree with you and the associate editor that the null hypothesis

   $$H_0^{\text{simple}}: \text{ All time trends are the same}$$

   is not very interesting by itself. In most applied cases, it is quite obvious that not all time trends are identical. Hence, there is no need to test this hypothesis formally. In particular, rejecting H_0^{simple} by a statistical test does not generate additional useful information.

   The main aim of our paper is to go beyond such a boring statistical test of H_0 and to come up with an approach that is able to generate interesting information

in practice. However, quite obviously, we have completely failed to convey this in the paper. In the revision, we have rewritten large junks of text to make a more compelling and convincing case for our approach. Let us briefly summarize the two main points here:

(a) As pointed out much more clearly in the revised paper, we consider a null hypothesis that is more general than H_0^{simple}. In fact, we test for co-trending of time series. Put differently, we test for parallelism of time trends. Formally speaking, our null hypothesis is

$$H_0 : m_i = m + c_i \text{ for all } 1 \leq i \leq n,$$

where m_i denotes the time trend of the i-th time series, c_i is a real constant and m is a common time trend. Hence, under H_0, the time trends m_i need not be exactly the same. They are rather vertically shifted version of the curve m.

In the previous version of the paper, it was (unfortunately) only implicit that we consider the more general H_0 rather than H_0^{simple}. The reason is as follows: In our model, the time trends m_i are only identified up to additive constants. In particular, the model equation

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \boldsymbol{\beta}_i^\top \boldsymbol{X}_{it} + \alpha_i + \varepsilon_{it} \qquad (*)$$

is equivalent to

$$Y_{it} = \tilde{m}_i\Big(\frac{t}{T}\Big) + \boldsymbol{\beta}_i^\top \boldsymbol{X}_{it} + \tilde{\alpha}_i + \varepsilon_{it},$$

where $\tilde{m}_i = m_i + \tilde{c}_i$, $\tilde{\alpha}_i = \alpha_i - \tilde{c}_i$ and $\tilde{c}_i$ is an arbitrary real constant. In this general model, it is thus only possible to test for parallelism of time trends (H_0) but not for their exact sameness (H_0^{simple}). For identification purposes, we impose the constraint $\int_0^1 m_i(u)du = 0$ on model equation $(*)$ for all i. This normalizes the trends m_i such that $\tilde{c}_i$ is the same for all i. Put differently, $m_i = m + c$ (with some constant c) for all i under our identification strategy, which implies that H_0 is reduced to H_0^{simple}. We apologize: we should have spelt this out clearly already in the old version of the paper.

(b) The associate editor writes: *Would one really want/need to test for the exact "sameness" of time series trends? Or is it the case that the null hypothesis is uninteresting, but the alternative is, especially if I am able to see which trends are different and where?* Exactly! We want to have an statistical approach which is not only able to tell us (with a certain statistical confidence) whether we are under the null or the alternative. We rather want a method which providesas much information as possible about the type of alternative we are

facing. Specifically, we want a method which allows us to say which trends are different and where they differ from each other. We think that this is exactly the information which is important in practice. Take for example our application on house price trends in different countries (or an application on temperature time trends at different spatial locations, an application on volatility trends of different stocks, ...). Most probably, the trends are not the same in all countries and most probably, they are also not all parallel. Hence, rejecting H_0 does not give us much valuable information. However, proper confidence statements about (i) which countries have different trends and (ii) where, i.e., in which time periods the trends differ may provide very valuable information for practitioners. It is exactly such confidence statements that are produced by our approach.

We hope the above summary and the corresponding discussion in the revised paper make a convincing case why our approach is much more useful and informative than competing methods for the comparison of time trends.

Reply to the associate editor

1. *The less enthusiastic referee wrote: "Perhaps it is my unfamiliarity with the problem (or my tendencies toward Bayesian methodologies...), but I do not find it to be a particularly compelling research question. The goal is to test whether a time trend—after adjusting for covariates—is identical across multiple time series. This does not seem to be a high priority for multiple time series and dynamic regression analysis, and it's not clear whether a hypothesis test generates much useful information in this context." This is a comment I broadly agree with: would one really want/need to test for the exact "sameness" of time series trends? Or is it the case that the null hypothesis is uninteresting, but the alternative is, especially if I am able to see which trends are different and where? I am thinking aloud here, but overall I don't think the testing problem, as stated, is interesting enough for JBES readers.*

   Please see our reply to the editor's comments above for a detailed response to your criticism.

2. *There are two prior papers by the submitting author, which consider a similar problem but in the absence of external covariates. I don't think the current paper makes it clear early enough what is different between the current work and those earlier papers.*

   In the revision, we discuss the main differences directly in the introduction (please see ?? therein).

3. *Due to the various approximations, the size control is only approximate. I don't see it as a "state of the art" way of thinking in these types of FWER control problems; please see e.g.* `https://arxiv.org/abs/2009.05431`*, where size control, in a different but related multiscale testing problem, is exact.*

   As you point out, the size control in our results is only approximate (in the sense of being asymptotic). The reason is that our proofs rely on strong approximation theory which is asympotitc in nature. For simplified versions of our model (e.g. for the most simplistic version $Y_{it} = m_i(t/T) + \varepsilon_{it}$ with errors that are i.i.d. both across i and t), it would be possible to get results on exact size control by using techniques from ??. However, in our general setting, we were not able to use these techniques and thus resorted to strong approximation techniques which allow us to get at least approximate size control. From an applied point of view, this is however not a big problem as far as we can see: usually, results on exact size control (as the ones in the linked paper) cannot be used directly in practice as they depend on certain distributions which are unknown in practice. Hence, one usually needs to resort to asymptotic approximations to derive critical values in practice, making the size control only approximate at the end of the day.

   We have added a short remark on p.?? to point out that the size control in our paper does not hold exactly in finite samples but is asymptotic in nature.

4. *I suspect the procedure must be really difficult to use in practice with confidence, as it depends on so many tuning parameters including the bandwidth. The authors say their software is at* `https://github.com/marina-khi/multiscale_inference`*, but the link is broken.*

   In response to your comment (and to comment ?? of referee ??), we have carried out a number of robustness checks in the revised simulation study where we consider different choices of tuning parameters (see Section ?? in the revised paper as well as our reply to comment ?? of referee ?? for further details).

   The main tuning parameter of our method is the grid $\mathcal{G}_T$ which in particular specifies the bandwidths that are taken into account by the procedure. Most nonparametric tests in the literature depend on one or more bandwidth parameters. Usually, the bandwidth is picked adhoc as there is virtually no theory for optimal bandwidth selection in nonparametric testing (as opposed to optimal bandwidth selection for nonparametric curve estimation). With our multiscale approach, we go one step into the direction of a bandwidth-free test: we consider various bandwidths simultaneously, thus avoiding the need to pick a single bandwidth adhoc. However, our procedure is of course not fully bandwidth-free as we still need to pick a set of bandwidths. Nevertheless, as long as this set is

chosen sufficiently rich (in the sense of including a variety of bandwidth values ranging from very small to very large), its particular choice can be expected to have a negligible effect on the procedure. This is supported by the robustness checks in Section ?? where we consider different grids $\mathcal{G}_T$.

Apart from the grid $\mathcal{G}_T$, our method depends on (i) the number of bootstrap samples L to compute the Gaussian quantile, (ii) the kernel K and (iii) secondary tuning parameters for the computation of the long-run error variance. As long as L is chosen large enough (say $L \geq 1000$), the precise choice of L should have a negligible effect. We use $L = 5000$ throughout the paper and have re-run everything with $L = 1000$ (not reported in the paper), which yields almost identical results. As suggested by classical nonparametric theory, the choice of kernel is much less crucial than the choice of bandwidth. Thus, we have not carried out robustness checks w.r.t. the choice of kernel. For the estimation of the long-run error variance, we can take an estimator off-the-shelf which will of course depend on further tuning parameters. In the paper, we work with the estimator proposed in ?? where extensive robustness checks w.r.t. to the choice of tuning parameters have been carried out. However, it is of course possible to work with other long-run variance estimators. As a rule of thumb, the stronger restrictions we make on the dependence structure of the error process, the easier it gets to estimate the long-run variance and the less tuning parameters are needed. (In the most extreme case where we assume the errors to be i.i.d., the long-run variance coincides with the short-run variance which is very easy to estimate in comparison with the long-run version.)

Finally, many thanks for pointing out to us that the link to the code was broken. We have fixed this. [Fix the link!]

5. *Both referees, including the more enthusiastic one, mention several further issues with the paper, including issues related to the practicalities of the method, the simulation study and the asymptotic nature of the method.*

   Please see our reply to referees 1 and 2 for a detailed point-by-point response to the issues raised.

Reply to referee 1

Thank you very much for the careful reading of our manuscript and the interesting suggestions. In our revision, we have addressed all your comments. Please see our replies to them below.

1. *The assumptions and requirements for the variance σ^2 deserve further consideration. First, it is claimed that the variances are assumed to be constant across*

series, but that a different estimator is used for each series. Which is the correct assumption for practice and theory? Second, given the economic and potential financial applications, how might volatility (or time-varying variance) be incorporated into the testing procedure? Is this plausible within the proposed framework, even if additional assumptions are required? If it is not plausible to account for volatility explicitly, then is the procedure robust in the presence of volatility?

2. *There are several issues with the simulation study.*
   (i) Setting the fixed effect to zero and including a single covariate both make for a much simpler design than considered in the theory. More challenging scenarios, including nonzero fixed effects and multiple predictors (e.g., using the estimated values and/or covariates from the application) would better demonstrate the capabilities of this approach.
   (ii) The data from the null fix $m_i = 0$ and claim this is WLOG. However, this is also quite a simple case: the shared $m_i()$ curve could be quite complex under the null, which only maintains that the trends are shared among the series.
   (iii) There are no competing methods considered; some alternative approach or benchmark must be added. A reasonable alternative might consider an additive model and compute confidence intervals (or bands) for the trends, with a simple heuristic to determine whether the functions are identical. The proposed approach should do better, but demonstrating improvements over a reasonable alternative is important.
   (iv) only a small number of series is considered. How does the approach perform when n is large?

   (i) We have taken your suggestions into account and consider the following, more challenging setup for the simulation study. In addition to the described setup, we also perform several robustness checks that are described further.

   – As before, we choose $n = 15$ and $T = 100, 250, 500$. Simulation scenario with $n = 15$ and $T = 150$ reflects well both applications.
   – We include 3 covariates and model them by a following VAR(3) process:

$$\underbrace{\begin{pmatrix} X_{it,1} \\ X_{it,2} \\ X_{it,3} \end{pmatrix}}_{=:\boldsymbol{X}_{it}} = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{pmatrix} \begin{pmatrix} X_{it-1,1} \\ X_{it-1,2} \\ X_{it-1,3} \end{pmatrix} + \underbrace{\begin{pmatrix} \nu_{it,1} \\ \nu_{it,2} \\ \nu_{it,3} \end{pmatrix}}_{=:\boldsymbol{\nu}_{it}}.$$

   We choose $a_1 = a_2 = a_3 = 0.25$. The innovations $\boldsymbol{\nu}_{it}$ are drawn i.i.d. from

a multivariate normal $N(0, \Phi)$ with

$$\Phi = \begin{pmatrix} 1 & \varphi & \varphi \\ \varphi & 1 & \varphi \\ \varphi & \varphi & 1 \end{pmatrix},$$

where $\varphi = 0.1, 0.25$. We can also set $\varphi = 0$ which will result in simulating 3 independent covariate processes.

- We set $\boldsymbol{\beta}_i = (\beta_{i,1}, \beta_{i,2}, \beta_{i,3}) = (1, 1, 1)$ for all i.
- We assume that the errors ε_{it} follow the AR(1) model $\varepsilon_{it} = a\varepsilon_{i(t-1)} + \eta_{it}$, where $a = 0.25$ and the innovations η_{it} are i.i.d. normal with zero mean $E[\eta_{it}] = 0$ and variance $E[\eta_{it}^2] = 0.25^2$.
- We let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a normally distributed random vector. In particular, $\alpha \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix},$$

where $\rho = 0.1, 0.25$ gives the correlation across time series i.

- To generate data under the null $H_0 : m_1 = \ldots = m_n$, we let $m_i = 0$ for all i as before. To produce data under the alternative, we use the bump functions $m_1(u) = b \cdot \mathbb{1}(u \in [0.3, 0.7]) \cdot \left(1 - \left\{\frac{u-0.5}{0.2}\right\}^2\right)^2$ for $b = 0.25, 0.5, 1$ (depicted in Figure 1) and $m_i = 0$ for $i \neq 1$.
- We take the grid $\mathcal{G}_T$ to be the same as before: $\mathcal{G}_T = U_T \times H_T$, where $U_T = \left\{u \in [0, 1] : u = \frac{5t}{T} \text{ for some } t \in \mathbb{N}\right\}$ and $H_T = \left\{h \in \left[\frac{\log T}{T}, \frac{1}{4}\right] : h = \frac{5t-3}{T} \text{ for some } t \in \mathbb{N}\right\}$.
- As before, in order to estimate the long-run variance σ_i^2, we follow the procedure described in Khismatullina and Vogt (2020) with the following tuning parameters: $q = 25$ and $r = 10$.
- As before, we calculate the Gaussian quantiles based on 5000 samples, and the size and power calculations are done based on 5000 repetitions.

The results of the simulation study with the covariates are presented in Tables 1 and 2 for $\phi = 0.1$ and $\rho = 0.1$ and in Tables 3 and 4 for $\phi = 0.25$ and $\rho = 0.25$. As can be seen, the empirical size gives a reasonable approximation to the target α in all scenarios under investigation, even though the size numbers have a slight upward bias. This bias gets smaller as the sample size T increases, which reflects the fact that we get more and more information for each of the tests that we need to carry out. We can

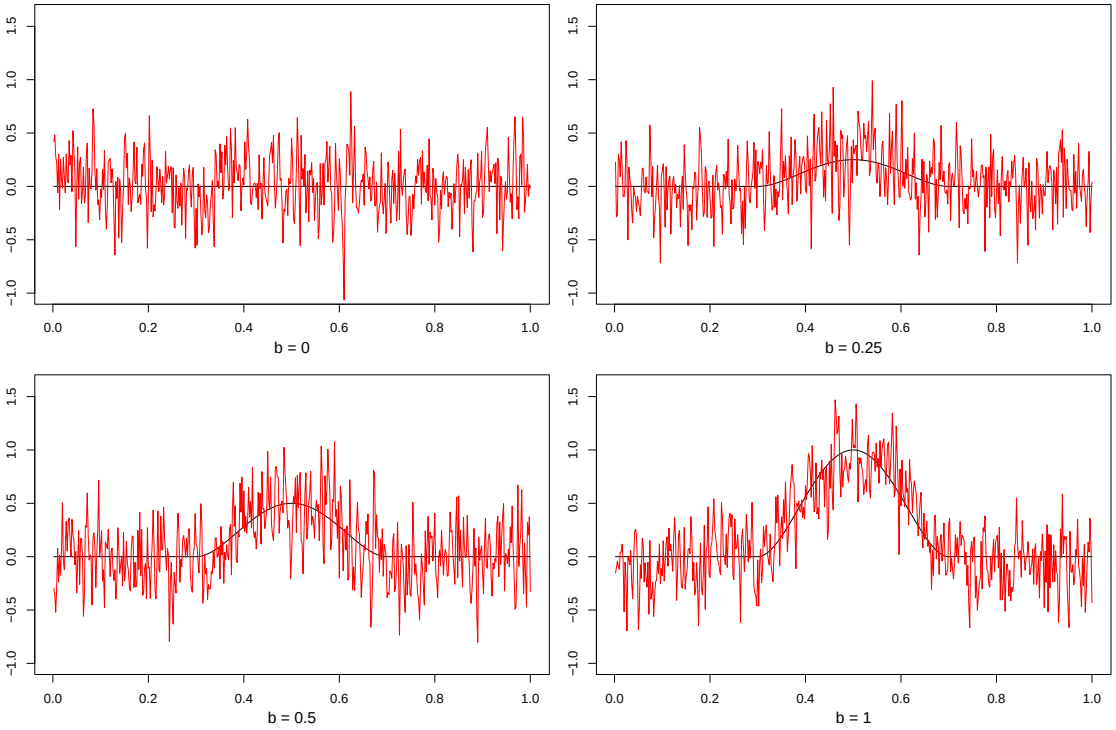

Figure 1: In black, the bump function $m_1(t/T)$ is plotted for different heights of the bump: $b = 0, 0.25, 0.5, 1$ ($b = 0$ corresponds to the data under the null H_0). In red, we depict the one specific instance of a bump function plus the error term, $m_1(t/T) + \varepsilon_{it}$ for $T = 500$.

also see that the upward bias become slightly more pronounced in case of stronger degrees of correlation across the covariates and the time series, i.e., with higher values of ϕ and ρ, but the change in test performance is almost negligible and can be attributed to random sampling. To summarize, even though slightly liberal, the test controls the FWER quite accurately in the simulation settings that we consider.

As for the empirical power calculations, the test has substantial power in all the considered simulation settings as can be seen in 4. For the smallest height of the bump function $b = 0.25$ and the smallest time series length $T = 100$, the power is only moderate, reflecting the fact that the alternative with $b = 0.25$ is not very far away from the null. However, as we increase the height b and the sample size T, the power increases quickly. Already for the height value $b = 0.5$, we reach a power of 0.98 for $T = 250$ and for all nominal sizes α.

As a robustness check, we calculate the actual size values within a slightly simplified simulation context. Specifically, we do not include any covari-

Table 1: Size of the multiscale test for $\phi = 0.1$ and $\rho = 0.1$ for different sample sizes T and nominal sizes α.

	nominal size α		
T	0.01	0.05	0.1
100	0.007	0.041	0.075
250	0.015	0.068	0.128
500	0.013	0.061	0.109

Table 2: Power of the multiscale test for $\phi = 0.1$ and $\rho = 0.1$ for different sample sizes T and nominal sizes α. Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$

	nominal size α		
T	0.01	0.05	0.1
100	0.019	0.073	0.128
250	0.242	0.440	0.565
500	0.731	0.864	0.914

(b) $b = 0.50$

	nominal size α		
T	0.01	0.05	0.1
100	0.058	0.198	0.299
250	0.981	0.996	0.999
500	1.000	1.000	1.000

(c) $b = 1.00$

	nominal size α		
T	0.01	0.05	0.1
100	0.012	0.060	0.114
250	1.000	1.000	1.000
500	1.000	1.000	1.000

ates in the model but only the fixed effects α_i with $\rho = 0.1$. Additionally, we explore larger sample sizes ($T = 100, 250, 500, 750, 1000$), albeit with a sparser grid $\mathcal{G}_T$ for computational efficiency: we define the grid as $\mathcal{G}_T = U_T \times H_T$, where $U_T = \left\{u \in [0,1] : u = \frac{10t}{T} \text{ for some } t \in \mathbb{N}\right\}$ and $H_T = \left\{h \in \left[\frac{\log T}{T}, \frac{1}{4}\right] : h = \frac{10t-3}{T} \text{ for some } t \in \mathbb{N}\right\}$. The findings are reported in Table 5. Notably, the results exhibit minimal deviation from the primary simulation scenarios. One thing that is potentially worth mentioning is that we can see a discernible "bump" in the actual size values for moderate sample sizes ($T = 250, 500$). This phenomenon persists across various specifications tested, including scenarios with $\rho = 0$ which indicates no dependence between the time series. This specific pattern can be explained by the trade-off between the sample size and the dimensionality of the problem, i.e., the number of tests we carry out simultaneously. For moderate value of the sample sizes ($T = 250, 500$) the number of comparisons is already quite large (15750 and 63000 simultaneous comparisons, respectively), while the effective sample size for testing individual null hypothesis remains relatively modest. However, empirical evidence suggests that the size values stabilize for larger sample sizes ($T = 750, 1000$) suggesting robust test performance even in the face of significant high dimensionality of the problem (149625 and 262500 simultaneous comparisons, respectively).

Table 3: Size of the multiscale test for $\phi = 0.25$ and $\rho = 0.25$ for different sample sizes T and nominal sizes α.

	nominal size α		
T	0.01	0.05	0.1
100	0.010	0.038	0.074
250	0.014	0.067	0.127
500	0.014	0.065	0.115

Table 4: Power of the multiscale test for $\phi = 0.25$ and $\rho = 0.25$ for different sample sizes T and nominal sizes α. Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$

	nominal size α		
T	0.01	0.05	0.1
100	0.017	0.069	0.139
250	0.235	0.459	0.583
500	0.704	0.860	0.914

(b) $b = 0.50$

	nominal size α		
T	0.01	0.05	0.1
100	0.062	0.196	0.318
250	0.982	0.997	0.999
500	1.000	1.000	1.000

(ii) Verbal argument.

(iii) To the best of our knowledge, the only other test for comparing trend curves with similar properties has been developed in Park et al. (2009) (SiZer). We have added comparison of our method with SiZer to the Appendix. However, we would like to note that their analysis is mainly methodological and the theory was developed only for the case of $n = 2$ time series. In Park et al. (2009), the authors propose a possible extension to their approach in case of more than 2 time series, but the extension does not allow for pairwise comparison of the time series. Moreover, it is not clear how to calculate the actual size and power in that case.

Furthermore, the model considered in Park et al. (2009) is much simpler than ours: it does not include neither the covariates, nor the fixed effects. Hence, in order to allow for fair comparison between two methods, we consider a simplified version of the simulation setup from (i). In particular, we do the following.

- We choose $n = 2$ and $T = 100, 250, 500, 750, 1000$.
- We consider a simplified model $Y_{it} = m_i\left(\frac{t}{T}\right) + \varepsilon_{it}$ that does not include the covariates or the fixed effects as they are not part of the model in Park et al. (2009).

Table 5: Size of the multiscale test for the case without any covariates and $\rho = 0.1$ for different sample sizes T and nominal sizes α.

	nominal size α		
T	0.01	0.05	0.1
100	0.005	0.032	0.064
250	0.021	0.074	0.134
500	0.010	0.061	0.114
750	0.010	0.062	0.122
1000	0.014	0.058	0.117

- As before, we assume that the errors ε_{it} follow the AR(1) model $\varepsilon_{it} = a\varepsilon_{i(t-1)} + \eta_{it}$, where $a = 0.25$ and the innovations η_{it} are i.i.d. normal with zero mean $E[\eta_{it}] = 0$ and variance $E[\eta_{it}^2] = 0.25^2$.
- To generate data under the null $H_0 : m_1 = m_2$, we let $m_i = 0$ for $i = 1, 2$ as before. To produce data under the alternative, we use the bump functions $m_1(u) = b \cdot \mathbb{1}(u \in [0.3, 0.7]) \cdot \left(1 - \left\{\frac{u-0.5}{0.2}\right\}^2\right)^2$ for $b = 0.25, 0.5$ (depicted in Figure 1) and $m_2 = 0$.
- We take the grid $\mathcal{G}_T$ as before: $\mathcal{G}_T = U_T \times H_T$, where $U_T = \left\{u \in [0, 1] : u = \frac{5t}{T} \text{ for some } t \in \mathbb{N}\right\}$ and $H_T = \left\{h \in \left[\frac{\log T}{T}, \frac{1}{4}\right] : h = \frac{5t-3}{T} \text{ for some } t \in \mathbb{N}\right\}$.
- In order to make the comparison between the methods as fair as possible, we do not estimate the long-run variance σ_i^2 from the data but consider σ_i^2 as known. Specifically, we use the theoretical value of the long-run variance calculated based on the true parameter values:

$$\sigma_i^2 = \frac{E[\eta_{it}^2]}{(1-a)^2} = \frac{1}{9} \text{ for all } i.$$

  Furthermore, since SiZer depends not on the long-run variance, but on the autocovariance functions of the time series $\gamma_i(\cdot)$ for $i = 1, 2$, in the calculation of the critical values of SiZer we use the following formula for the autocovariance function:

$$\gamma_i(k) = \frac{E[\eta_{it}^2]a^{|k|}}{1-a^2} = \frac{0.25^{|k|}}{15}.$$

- As before, we calculate the Gaussian quantiles based on 5000 samples, and the size and power calculations are done based on 5000 repetitions.
- All of the details of the exact implementation of SiZer can be found in the Appendix in Khismatullina and Vogt (2020).

Table 6: Size comparison of the proposed multiscale test ($\mathcal{T}_{\text{MS}}$) and SiZer ($\mathcal{T}_{\text{SiZer}}$, Park et al. (2009)) for different sample sizes T and various significance levels α.

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$
$T = 100$	0.010	0.099	0.047	0.332	0.106	0.524
$T = 250$	0.008	0.162	0.041	0.510	0.102	0.699
$T = 500$	0.009	0.214	0.044	0.598	0.089	0.787
$T = 750$	0.008	0.252	0.045	0.657	0.095	0.844
$T = 1000$	0.012	0.271	0.053	0.691	0.102	0.873

Table 7: Power comparison of the proposed multiscale test ($\mathcal{T}_{\text{MS}}$) and SiZer ($\mathcal{T}_{\text{SiZer}}$, Park et al. (2009)) for different sample sizes T and various significance levels α for the bump function with $b = 0.25$.

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$
$T = 100$	0.075	0.251	0.193	0.551	0.317	0.714
$T = 250$	0.253	0.629	0.486	0.876	0.633	0.951
$T = 500$	0.640	0.905	0.817	0.985	0.887	0.996
$T = 750$	0.867	0.982	0.957	0.998	0.981	1.000
$T = 1000$	0.970	0.998	0.993	1.000	0.997	1.000

Table 8: Power comparison of the proposed multiscale test ($\mathcal{T}_{\text{MS}}$) and SiZer ($\mathcal{T}_{\text{SiZer}}$, Park et al. (2009)) for different sample sizes T and various significance levels α for the bump function with $b = 0.5$.

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{SiZer}}$
$T = 100$	0.482	0.733	0.703	0.922	0.815	0.967
$T = 250$	0.966	0.998	0.994	1.000	0.998	1.000
$T = 500$	1.000	1.000	1.000	1.000	1.000	1.000
$T = 750$	1.000	1.000	1.000	1.000	1.000	1.000
$T = 1000$	1.000	1.000	1.000	1.000	1.000	1.000

In what follows, we denote the proposed multiscale procedure and SiZer (Park et al. (2009)) by $\mathcal{T}_{\text{MS}}$ and $\mathcal{T}_{\text{SiZer}}$, respectively.

The results of the size and power simulation studies that compare $\mathcal{T}_{\text{MS}}$ and $\mathcal{T}_{\text{SiZer}}$ are presented in Tables 6 and Tables 7, 8 respectively. There

is an important difference between $\mathcal{T}_{\text{MS}}$ and $\mathcal{T}_{\text{SiZer}}$ to be noticed: $\mathcal{T}_{\text{MS}}$ is a *global* test procedures while $\mathcal{T}_{\text{SiZer}}$ is a scale-wise in its essence. This means that $\mathcal{T}_{\text{MS}}$ tests $H_0(u,h)$ simultaneously for all locations $u \in U_T$ and scales $h \in H_T$ and controls the size simultaneously over both locations u and scales h, whereas $\mathcal{T}_{\text{SiZer}}$ tests the hypothesis $H_0(u,h)$ simultaneously for all $u \in U_T$ but separately for each scale $h \in H_T$ and, hence, controls the size for each scale $h \in H_T$ separately. This results in a much more liberal nature of $\mathcal{T}_{\text{SiZer}}$ as can be seen in Table 6. Specifically, the size numbers of the multiscale test $\mathcal{T}_{\text{MS}}$ are reasonably close to the target nominal size levels α. Contrarily, the actual size numbers of $\mathcal{T}_{\text{SiZer}}$ are much larger than the target α. Since the number of scales h in the grid $\mathcal{G}_T$ increases with T, they even move away from α as the sample size T increases. To summarize, as expected, the global test $\mathcal{T}_{\text{MS}}$ holds the size reasonably well, whereas the row-wise method $\mathcal{T}_{\text{SiZer}}$ is much too liberal.

As for the power simulation studies that are presented in Tables 7 and 8, we can note that in all simulation scenarios, $\mathcal{T}_{\text{SiZer}}$ is more powerful than $\mathcal{T}_{\text{MS}}$. This gain of power is presumably due to the fact that $\mathcal{T}_{\text{SiZer}}$ is in general too liberal in terms of size as observed in Table 6.

(iv) In order to illustrate the performance of the test with larger number of time series, we rerun the simulations from (i) with a smaller grid. This allows us to deal not only with $n = 15$ but also with larger values of n: $n = 25, 50, 100$. Specifically, we take the grid $\mathcal{G}_T$ to be a dyadic scheme (as in Wavelet analysis) with scales in the set

$$\mathcal{H}_T = \big\{h = 2^k h_{\min} \text{ for } k = 0, \dots, K\big\},$$

where

$$h_{\min} = \frac{\lceil \log T \rceil}{T}$$

and K is such that $2^K h_{\min} \leq \frac{1}{4}$, i.e.,

$$K \leq \Big\lfloor \log\Big(\frac{T}{4\lceil \log T \rceil}\Big)\Big\rfloor \frac{1}{\log(2)},$$

and

$$\mathcal{G}_T = \Big\{(u,h) \subseteq [0,1] : (u,h) = ((2s+1)h, h) \text{ for } s = 0, \dots, \Big\lfloor \frac{h^{-1}-1}{2} \Big\rfloor \text{ and } h \in \mathcal{H}_T\Big\}.$$

We rerun the simulations precisely as specified in (i), deviating solely in the grid specification and focusing on the scenario where $\phi = 0.1$ and $\rho = 0.1$

Table 9: Size of the multiscale test for $n = 15$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

	nominal size α		
T	0.01	0.05	0.1
100	0.005	0.035	0.066
250	0.012	0.051	0.094
500	0.015	0.063	0.116

Table 10: Power of the multiscale test for $n = 15$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$

	nominal size α		
T	0.01	0.05	0.1
100	0.015	0.068	0.121
250	0.096	0.232	0.322
500	0.524	0.722	0.805

(b) $b = 0.50$

	nominal size α		
T	0.01	0.05	0.1
100	0.061	0.192	0.293
250	0.742	0.886	0.934
500	1.000	1.000	1.000

(c) $b = 1.00$

	nominal size α		
T	0.01	0.05	0.1
100	0.013	0.068	0.126
250	1.000	1.000	1.000
500	1.000	1.000	1.000

across varying sample sizes: $T = 100, 250, 500$, and diverse number of time series $n = 15, 25, 50, 100$. The results are presented in Tables 9 - 16. As can be seen in Tables 9, 11, 13 and 15, the empirical size provides an appropriate approximation to the target alpha across all scenarios under consideration. In regards to the empirical power calculations, the test demonstrates significant power across all simulated scenarios, as illustrated in Tables 10, 12, 14 and 16. Notably, despite slight variations from the benchmark model, the observed changes are not substantial which affirms the robustness of the proposed test across different specifications of the grid.

3. *Similarly, there are many related clustering methods, including (Bayesian and non-Bayesian) methods for clustering functional data. The proposed approach is reasonable, yet should be placed in a broader context and evaluated against appropriate competitors.*

   We next investigate the finite sample performance of the clustering algorithm.To do so, we consider a very simple scenario: we generate data from the model $Y_{it} = m_i(\frac{t}{T}) + \varepsilon_{it}$, that is, we assume that there are no fixed effects and no covariates. The error terms ε_{it} are specified as before. Moreover, as before, we set the number of time series to $n = 15$ and we consider different time series lengths T. We partition the $n = 15$ time series into $N = 3$ groups, each

Table 11: Size of the multiscale test for $n = 25$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

	nominal size α		
T	0.01	0.05	0.1
100	0.007	0.037	0.074
250	0.010	0.056	0.116
500	0.017	0.052	0.100

Table 12: Power of the multiscale test for $n = 25$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$

	nominal size α		
T	0.01	0.05	0.1
100	0.014	0.050	0.099
250	0.081	0.213	0.303
500	0.476	0.665	0.763

(b) $b = 0.50$

	nominal size α		
T	0.01	0.05	0.1
100	0.049	0.143	0.234
250	0.705	0.870	0.921
500	1.000	1.000	1.000

(c) $b = 1.00$

	nominal size α		
T	0.01	0.05	0.1
100	0.012	0.051	0.103
250	1.000	1.000	1.000
500	1.000	1.000	1.000

containing 5 time series. Specifically, we set $G_1 = \{1, \dots, 5\}$, $G_2 = \{6, \dots, 10\}$ and $G_3 = \{11, \dots, 15\}$, and we assume that $m_i = f_l$ for all $i \in G_l$ and all $l = 1, 2, 3$. The group-specific trend functions f_1, f_2 and f_3 are defined as $f_1(u) = 0$, $f_2(u) = \mathbb{1}\left\{\frac{|u-0.25|}{0.25} \leq 1\right\}\left(1 - \frac{(u-0.25)^2}{0.25^2}\right)^2 - \mathbb{1}\left\{\frac{|u-0.75|}{0.25} \leq 1\right\}\left(1 - \frac{(u-0.75)^2}{0.25^2}\right)^2$ and $f_3(u) = 4 \cdot \mathbb{1}\left\{\frac{|u-0.75|}{0.025} \leq 1\right\}\left(1 - \frac{(u-0.75)^2}{0.025^2}\right)^2 - 4 \cdot \mathbb{1}\left\{\frac{|u-0.25|}{0.025} \leq 1\right\}\left(1 - \frac{(u-0.25)^2}{0.025^2}\right)^2$. In order to estimate the groups G_1, G_2, G_3 and their number $N = 3$, we use the same implementation as before followed by the clustering procedure.

As a benchmark, we use the following clustering procedure:

- Estimate the trends m_i by a local linear estimator $\hat{m}_i$ with a fixed bandwidth chosen as the smallest bandwidth from H_T.
- Compute a $\mathcal{L}^2$ distance measure d_{ij} between $\hat{m}_i$ and $\hat{m}_j$:

$$d_{ij} = \int_0^1 (\hat{m}_i(w) - \hat{m}_j(w))^2 dw.$$

- Construct the following dissimilarity measure from these distances:

$$\hat{\Delta}(S, S') = \max_{i \in S, j \in S'} d_{ij}.$$

- Run a HAC algorithm with the computed dissimilarities.

Table 13: Size of the multiscale test for $n = 50$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

	nominal size α		
T	0.01	0.05	0.1
100	0.006	0.031	0.063
250	0.010	0.051	0.102
500	0.023	0.076	0.126

Table 14: Power of the multiscale test for $n = 50$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$

	nominal size α		
T	0.01	0.05	0.1
100	0.011	0.055	0.100
250	0.069	0.172	0.260
500	0.449	0.646	0.736

(b) $b = 0.50$

	nominal size α		
T	0.01	0.05	0.1
100	0.026	0.102	0.179
250	0.680	0.836	0.895
500	0.999	1.000	1.000

(c) $b = 1.00$

	nominal size α		
T	0.01	0.05	0.1
100	0.006	0.040	0.081
250	1.000	1.000	1.000
500	1.000	1.000	1.000

Table 15: Size of the multiscale test for $n = 100$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

	nominal size α		
T	0.01	0.05	0.1
100	0.007	0.035	0.067
250	0.014	0.057	0.111
500	0.017	0.071	0.126

Table 16: Power of the multiscale test for $n = 100$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$

	nominal size α		
T	0.01	0.05	0.1
100	0.008	0.037	0.076
250	0.053	0.152	0.244
500	0.426	0.596	0.689

(b) $b = 0.50$

	nominal size α		
T	0.01	0.05	0.1
100	0.018	0.076	0.141
250	0.647	0.807	0.874
500	0.999	1.000	1.000

(c) $b = 1.00$

	nominal size α		
T	0.01	0.05	0.1
100	0.005	0.034	0.074
250	1.000	1.000	1.000
500	1.000	1.000	1.000

Our procedure can be regarded as a further development of this very simple and natural benchmark procedure. In particular: our procedure replaces the simple distance measure d_{ij} by a more advanced multiscale distance measure and provides a way to estimate the number of clusters, which is not part of the simple benchmark procedure.

Here we have a problem: the benchmark procedure is actually pretty good, much better than our method. It is very slow though. Discuss during the next call?

The simulation results are reported as follows. For our multiscale method, the empirical probabilities with which which the estimate $\widehat{N}$ is equal to the true number of groups $N = 3$ for the significance level $\alpha = 0.05$ for $T = 100, 250$ and 500 are $0.028, 0.117$ and 0.9678, respectively. Furthermore, the empirical probabilities with which the estimated group structure $\{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{N}}\}$ equals the true one $\{G_1, G_2, G_3\}$ for the significance level $\alpha = 0.05$ for $T = 100, 250$ and 500 are $0, 0.1112$ and 0.9674, respectively. The benchmark procedure is able to detect the difference between the groups in all possible case for all of the sample sizes. There must be something wrong here, no?

As the benchmark procedure does not provide an estimate of the number of clusters K, I have also compared the results for the known number of clusters ($K = 3$) for $T = 100$ and the results are not better... We almost never detect the true grouping anyway.

4. *A related Bayesian strategy is to use simultaneous band scores (simBaS) to assess whether a function differs from zero. This could be applied pairwise to the differences between functions to establish a Bayesian competitor to the proposed approach, and simply requires posterior draws from an analogous Bayesian model.*

   We may argue here that we use the benchmark discussed in the previous comment instead of the proposed Bayesian strategy because it is naturally linked to our approach.

5. *The application includes numerous tuning parameters (including kernels, intervals, etc.). Are the results robust to these choices? Further details are needed.*

   The tuning parameters are:

   (a) the grid $\mathcal{G}_T$

   (b) tuning parameters to estimate the error variances σ_i^2

   (c) the number of bootstrap samples L to compute the Gaussian quantile

   (d) the kernel K.

   (Have I forgotten anything here?)

We should run robustness checks:

(a) We should consider different grids $\mathcal{G}_T$. In the simulation study, we use $\mathcal{G}_T = U_T \times H_T$ with

$$U_T = \left\{u \in [0,1] : u = \tfrac{5t}{T} \text{ for some } t \in \mathbb{N}\right\}$$
$$H_T = \left\{h \in \left[\tfrac{\log T}{T}, \tfrac{1}{4}\right] : h = \tfrac{5t-3}{T} \text{ for some } t \in \mathbb{N}\right\}.$$

We could additionally consider a finer grid with, e.g.,

$$U_T = \left\{u \in [0,1] : u = \tfrac{t}{T} \text{ for some } t \in \mathbb{N}\right\}$$
$$H_T = \left\{h \in \left[\tfrac{\log T}{T}, \tfrac{1}{4}\right] : h = \tfrac{t}{T} \text{ for some } t \in \mathbb{N}\right\}$$

and a sparser one with, e.g.,

$$U_T = \left\{u \in [0,1] : u = \tfrac{10t}{T} \text{ for some } t \in \mathbb{N}\right\}$$
$$H_T = \left\{h \in \left[\tfrac{\log T}{T}, \tfrac{1}{4}\right] : h = \tfrac{10t-3}{T} \text{ for some } t \in \mathbb{N}\right\}.$$

I guess it is enough to consider only $n = 15$ and $T = 100$ for the robustness check. But I guess we should consider both the null and the alternative(s).

(b) I don't know whether it is necessary to consider different tuning parameters for the estimation of the error variance. Maybe we could run the procedure with the true $\sigma^2 = \sigma_i^2$ as a benchmark. Ideally, we can then report that the results produced by the benchmark are very similar to those produced by the feasible algorithm with estimated σ_i^2. I guess this should be enough.

(c) One may compute the Gaussian quantile with different values of L. I guess the results should be very similar.

(d) I think there is no need to try out different kernels K as from nonparametrics it is well known that the kernel is not so important.

6. *The multiscale tests are designed to control the FWER. Why is that the right criterion for the types of applications in mind (compared to e.g., FDR)? Given that other reasonable choices exist, additional motivation for this objective is warranted.*

7. *I'm wondering if there might be some clarification about the independence of ε_{it} across series i. In particular, suppose the intercepts α_i were instead considered random, like in mixed modeling (or Bayesian inference). Then marginally, the "new" errors $(\alpha_i + \varepsilon_{it})$ would be dependent across series i. Similar reasoning might apply to the covariates. From this perspective, the class of models might be considered more general.*

8. *It is claimed on p.6 that the mean function integrating to zero is "required" for identification of the intercept. I think this is a sufficient, not necessary, condition, since others might suffice.*

Reply to referee 2

Thank you very much for the careful reading of our manuscript and the interesting suggestions. We have addressed all your comments in the revision. Please see our replies to them below.

1. *Although you correctly cite Khismatullina and Vogt (2020, 2021) on which quite a bit of this new work seems to be based can you please summarize more in particular about the test proposed in Khismatullina and Vogt (2021, Journal of Econometrics), and explain where and how your proofs differ from that (e.g. by the complexity in needing to treat the covariates).*

2. *As your results are of asymptotic nature, it would be good to discuss limitations – even give an example where the procedure would cease to work.*

3. *Moreover, can you at least sketch out if by something like a Bootstrap procedure (cf. Zhang et al, 2012) more of the "asymptotic flavour" of your test/cluster procedure could be remedied?*

4. *I am having a slight (finite sample) identification concern with (not only your) model(s) mixing deterministic (nonparametric) trends with covariate (and also error) structure which is allowed to be positively serially dependent, e.g. autoregressive (as in your examples): I think that for "any" fixed sample size it might always occur that the trajectory of a stochastic trend, an autoregressive process with roots relatively close to the unit circle, say, cannot be distinguished from the deterministic trend. Wouldn't that be potentially a problem for your (and any related) test procedure? As a follow-up on this, wouldn't you need (or to say it differently, wouldn't it be perhaps beneficial to add) some extra conditions on the nature of your covariates (and potentially also your errors ε?) to avoid this problem?*

5. *What about a naive competitor that is just based on the second derivative (= change of the slope parameters) rather than the distance based on the curves and the first derivative (as in your local linear estimator)? I believe that this could also work rather well on your economic example data in Figures 3–6? Maybe you can "benchmark" your procedure against such a simple competitor (as such a comparison is somewhat missing explicitly – although you orally compare sometimes with Zhang et al (2012)).*

6. *Can your proposed test procedure be considered somehow to be equivalent to constructing a uniform confidence region where you would need to control if two (or more curves) are within the same tube (not just pointwise)? If so, it would be perhaps interesting to explain the link, and why for your test procedure it is sufficient/adequate to control the "familywise" error (does this correspond to what one does for a "uniform" region?).*

7. *How in all of this does the number of curves (larger than two) play a role, in practice, for correctly calibrating your test (as least asymptotically as possible)? On the other hand, do your results reflect the fact that obviously they depend on the number of time series (or rather the number of series where trends are different, a number that you would have access to in an oracle situation)?*

8. *Here is a small series of remarks towards needing to choose (u, h) – an example for a practical choice is given in Section 7, only (a bit late): Your localised multiscale method requires to discretize the continuous (u, h). I am wondering if the way to do this plays a role for the properties of the resulting practical procedure. Can you please also compare with wavelet-based multiscale methods which are based somehow on a "built-in" way of choosing the location-scale parameters (u, h)?*

9. *page 12, around equation (3.6): it took me a moment to understand that you are talking about the standard local linear estimator (of Fan and Gijbels) here, you might want to make this clearer.*

10. *I understand the heuristics behind using the Gaussian version (3.12) of the test statistics in the "idealised" situation but what about the "non-idealised" situation of unknown variances σ^2 and unknown parameters β? Is the Gaussian-based MC simulation method still valid when you need to estimate those parameters?*

11. *Again about the choice of (u, h): what happens with expressions (such as in equation (4.1)) which depend on $\max_{(u,h)}$ in practice where you have to discretize this (u, h)? I do not think that a maximum over a continuous location-scale parameter can be treated the same way as one over a discrete one? Does the choice of the grid $\mathcal{G}_T$ influence the results here, don't you need some (additional) conditions on the grid (its spacing etc)? This refers, e.g. to the simulation section 6, page 24 where – in passing – you might want to change the strange wording there where you say "for some t in N" and rather detail the specification of the grid in t here as you do later in Section 7.*

12. *Section 7, page 41, lines 45-50: can you develop this conjecture a bit?*

13. *You might want to add a Conclusion Section which could both serve to recall the difficulties encountered in treating the more general situation of more than two curves and the presence of covariates, and also discuss some of the aforementioned points on Bootstrap alternatives or on potential competitors.*

14. *Develop more to which extent the second data application (in the Supplement) brings insights beyond the one of the first (and why you chose to present the first and not the second in the main body of the text).*

15. *Supplement section page 15, line 49 – a notational detail: should the first* o_p *be* O_p *if the* $\rho_T = o(1/\log(T))$ *or vice versa?*

16. *It would be good to explain somewhere in the main body (Section 3 or 4?) the additional difficulties in proving the results in the presence of the covariates.*

References

Khismatullina, M. and Vogt, M. (2020). Multiscale inference and long-run variance estimation in non-parametric regression with time series errors. *Journal of the Royal Statistical Society: Series B*, **82** 5–37.

Park, C., Vaughan, A., Hannig, J. and Kang, K.-H. (2009). SiZer analysis for the comparison of time series. *Journal of Statistical Planning and Inference*, **139** 3974–3988.