Environmental and Ecological Statistics
https://doi.org/10.1007/s10651-024-00605-x

Sequential hypothesis testing for selecting the number of changepoints in segmented regression models

Andrea Priulla[1] · **Nicoletta D'Angelo**[1]

Received: 19 September 2023 / Accepted: 7 February 2024

Abstract
Segmented regression is widely used in many disciplines, especially when dealing with environmental data. This paper deals with the problem of selecting the correct number of changepoints in segmented regression models. A review of the usual selection criteria, namely information criteria and hypothesis testing, is provided. We enhance the latter method by proposing a novel sequential hypothesis testing procedure to address this problem. Our sequential procedure's performance is compared to methods based on information-based criteria through simulation studies. The results show that our proposal performs similarly to its competitors for the Gaussian, Binomial, and Poisson cases. Finally, we present two applications to environmental datasets of crime data in Valencia and global temperature land data.

Keywords Changepoint · Hypothesis testing · Information criterion · Segmented regression · Score test

1 Introduction

Segmented regression is a standard tool in many fields, including epidemiology (Ulm 1991), occupational medicine, toxicology, ecology, biology (Betts et al. 2007), and more recently, higher education (Li et al. 2019; Priulla et al. 2021).

Segmented or broken-line models are regression models where the relationships between the response and one or more explanatory variables are piecewise linear, namely represented by two or more straight lines connected at unknown values. These values are commonly referred to as *changepoints* or *breakpoints*. The main

Handling Editor: Luiz Duczmal.

✉ Nicoletta D'Angelo
nicoletta.dangelo@unipa.it

Andrea Priulla
andrea.priulla@unipa.it

[1] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

Published online: 21 March 2024

advantage of these models lies in the results' interpretability while also achieving a good trade-off with flexibility, typically achieved by non-parametric approaches.

This paper deals with the task of selecting the number of changepoints in segmented regression models, a topic widely discussed by many authors, as Lerman (1980) and Kim et al. (2000). This is a common problem in segmented regression models. Indeed, if the number of changepoints is too small, the model may not capture all the changes in the relationship between the variables, resulting in bias and reduced model fit. On the other hand, if the number of changepoints is too large, the model may overfit the data and not generalize well to new data.

Other common problems in segmented regression models include the identification of the location of the changepoints in a segmented variable, and the estimation of its effects on the response variable, to cite a few. For instance, Horváth et al. (2004) proposes two classes of monitoring schemes to (sequentially) detect a structural change in a linear model. Aue et al. (2006) develop an asymptotic theory for two monitoring schemes aimed at detecting a change in the regression parameters, showing to have a correct asymptotic size and detecting a change with probability approaching unity. Then, Chen et al. (2011) deal with two problems concerning locating changepoints in a linear regression model, namely, the one involving jump discontinuities (not covered in this paper) in a regression model and the other involving regression lines connected at unknown points. The latter is the main framework covered in our work. Muggeo and Adelfio (2011) present a computationally efficient method to obtain estimates of the number and location of the changepoints in genomic sequences, or, more generally, in mean-shift regression models. Adelfio (2012) introduces a new approach based on the fit of a generalized linear regression model for detecting changepoints in the variance of heteroscedastic Gaussian variables with piecewise constant variance function, and D'Angelo et al. (2022) extend such approach in order to detect changepoints in the variance of multivariate Gaussian variables, allowing to provide simultaneous detection of changepoints in functional time series.

Moving to the main topic of our work, in literature several approaches have been proposed to select the optimal number of changepoints in segmented regression models. One common approach is to use information criteria, such as the Akaike Information Criterion (AIC) (Akaike 1974) or the Bayesian Information Criterion (BIC) (Schwarz 1978), which balance the model's goodness of fit with the model's complexity. These criteria penalize models with more parameters, which can prevent overfitting.

Another approach is to use cross-validation (Zou et al. 2020; Pein 2023), which involves splitting the data into training and validation sets, fitting models with different numbers of changepoints to the training set, and then selecting the number of changepoints that minimizes the prediction error on the validation set. The model that performs best on the validation set is selected. Cross-validation can be computationally intensive, but it can also provide a more accurate estimate of model performance than AIC or BIC.

A further approach is hypothesis testing to determine the significance of adding a changepoint to the model. Typically, this consists of performing different hypothesis tests starting from testing $\mathcal{H}_0 : K_0 = 0$ vs $\mathcal{H}_1 : K_0 = K_{max}$ where K_0 is the true

number of changepoints and K_{max} is the maximum number of potential changepoints fixed a priori, as done by Kim et al. (2000). However, this well-established procedure, which requires sequentially testing for the existence of a changepoint, makes testing for any additional changepoints unfeasible.

In light of this, this paper proposes a novel sequential hypothesis testing procedure that overcomes this problem, having the perk of not being limited to testing for a maximum number of additional changepoints fixed a priori. Starting from the work of Kim et al. (2000), we enhance such a method by proposing a novel sequential hypothesis testing, to identify the correct number of changepoints.

First, we provide an overview of the segmented regression models and a review of the main tools useful for selecting the number of changepoints. Regarding the information-based criteria, we consider the AIC, the BIC and the generalized Bayesian Information Criterion (gBIC). As regards hypothesis testing, we consider Davies' test (Davies 1977) and the Score test (Muggeo 2016). The performance of the different tools and our proposed procedure is then assessed through simulation studies. Finally, to explore the applicability of the considered framework, two original applications are proposed: the first one deals with the crime events that occurred in Valencia in 2019, available from the `stopp` package (D'Angelo and Adelfio 2023) of the software R (R Core Team 2023); the second one deals with global temperature anomalies data from the NOAA Merged Land Ocean Global Surface Temperature Analysis Data set (Smith et al. 2008). All the analyses are performed using the `segmented` package (Muggeo 2008) of the R Core Team (2023) statistical software, and original codes from the authors.

The structure of the paper is as follows. Section 2 introduces the segmented regression model and Section 3 reviews suitable criteria for model selection in this context. Section 4 illustrates our proposal. Section 5 presents simulations to study the performance of the given criteria, and Section 6 proposes two applications dealing with crime events in Valencia and with global anomalies temperature data. The paper ends with conclusions in Section 7. The Appendix contains supplementary material in support of the run experiments.

2 Background on the segmented regression models

The segmented linear regression is expressed as

$$g(\mathbb{E}[Y|x_i, z_i]) = \beta_0 + \theta z_i + \beta_1 x_i + \sum_{k=1}^{K_0} \delta_k (x_i - \psi_k)_+ \tag{1}$$

where g is the link function, x is a broken-line covariate and z is a covariate whose relationship with the response variable is not broken-line. Multiple covariates can be accounted for, but we limit our study to unique covariates. We denote by K_0 the true number of changepoints and by ψ_k the K_0 locations of the changes in the relationship that we call, from now on, *changepoints*. These are selected among all the possible values in the range of x. The notation $(x_i - \psi_k)_+$ is to be read as $(x_i - \psi_k) I(x_i > \psi_k)$.

The coefficient θ represents the non-broken-line effect of z, β_1 represents the effect when $x_i < \psi_1$, while $\boldsymbol{\delta} = \{\delta_k\}_{k=1}^{K_0}$ is the vector of the differences in the effects.

The basic statistical problem dealt in this paper is the identification of the number of changepoints K_0. The estimation of their locations, that is the vector of ψ_k, and the broken-line effects, represented by β_1 and the vector $\boldsymbol{\delta}$, may also be of interest.

For estimation purposes, a reparametrization of the segmented model in Equation (1) is considered, dropping the non-segmented covariate z without any loss of generality. This reparametrization has the advantage of an efficient estimating approach via the algorithm discussed in Muggeo (2003, 2008), fitting iteratively the generalized linear model:

$$g(\mathbb{E}[Y|x_i]) = \beta_1 x_i + \sum_k \delta_k \tilde{U}_{ik} + \sum_k \gamma_k \tilde{V}_{ik}^{-}, \tag{2}$$

where $\tilde{U}_{ik} = (x_i - \tilde{\psi}_k)_+$, $\tilde{V}_{ik}^{-} = -I(x_i > \tilde{\psi}_k)$. The parameters β and δ_k are the same as in Equation (1), while the γ are the working coefficients useful only for the estimation procedure. At each iteration, the working model in Equation (2) is fitted and new estimates of the changepoints are obtained via: $\hat{\psi}_k = \tilde{\psi}_k + \frac{\hat{\gamma}_k}{\hat{\delta}_k}$ iterating the process up to convergence. Inferences on $\hat{\psi}$ is usually the main interest, and can be drawn by means of bootstrap, likelihood-based or Wald-type methods. In particular, Muggeo (2003) discusses and implements the usage of the Wald statistics. The standard error of $\hat{\psi}$ is obtained through a linear approximation for the ratio of two random variables using the Delta Method: $SE(\hat{\psi}) = \{[\text{var}(\hat{\gamma}) + \text{var}(\hat{\beta})(\hat{\gamma}/\hat{\beta})^2 + 2(\hat{\gamma}/\hat{\beta})\text{cov}(\hat{\gamma}, \hat{\beta})]/\hat{\beta}^2\}^{\frac{1}{2}}$, with $\text{var}(\cdot)$ and $\text{cov}(\cdot;\cdot)$ the variance and covariance, respectively.

For further details on the estimation procedure, we refer to Muggeo (2003).

3 Selecting the number of changepoints

In a more general context, with multiple changepoints as in Equation (1), we need to select only the significant changepoints by removing the spurious ones. Indeed, whether the generic $\hat{\psi}_k$ is not significant, the corresponding covariate V_k in Equation (2) should be a noise variable, as it would be $\hat{\delta}_k \approx 0$. Therefore, selecting the number of significant changepoints in model (1) means selecting the significant variables among $V_1, \ldots, V_{K^*}$, from model (2), where K^* is the number of estimated changepoints. The fitted optimal model will have $\hat{K} \leq K^*$ changepoints selected by any criterion. It is important to notice that these models are not nested, so likelihood ratio tests for model selection cannot be used.

Furthermore, the usual statistics cannot be used to verify the existence of a changepoint, since it is present only under the alternative hypothesis. This leads to a non-linear problem because the regularity conditions of the log-likelihood are not satisfied.

Basically, we need to select the $\hat{\psi}_1, \ldots, \hat{\psi}_{\hat{K}}$ among the $\hat{\psi}_1, \ldots, \hat{\psi}_{K^*}$ via a selection criterion. The changepoints $\hat{\psi}_1, \ldots, \hat{\psi}_{\hat{K}}$ will be a subset of the estimates $\hat{\psi}_1, \ldots, \hat{\psi}_{K^*}$,

since one or more changepoints are not included due to the deletion of one or more variables V_k by means of the given selection criterion. Therefore, it should be noticed that, while $\hat{\psi}_1, \ldots, \hat{\psi}_{K^*}$ are the estimates maximising the likelihood with K^* changepoints, there is no guarantee that the subset $\hat{\psi}_1, \ldots, \hat{\psi}_{\hat{K}}$ constitutes also the best estimate for the number of changepoints.

Much of the literature deals with the problem of determining the 'best' subset of independent variables: Hocking (1976) summarizes various selection criteria, reviewed below. These can be classified under two major approaches: information criteria and hypothesis testing.

The first information criterion is the well-known Akaike Information Criterion (Akaike 1974), expressed as $AIC = -2 \log L + 2p$, where L represents the likelihood function and p stands for the actual model dimension quantified by the number of estimated parameters, including $\hat{\beta}$, the $\hat{\delta}$ and $\hat{\psi}$ vectors in the segmented regression models, that is $p = 1 + 2\hat{K}$. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

The second criterion is the Bayesian Information Criterion (Schwarz 1978) $BIC = -2 \log L + p \log(n)$ that includes both a penalty for the number of estimated parameters p and for the logarithm of the number of observations n. In the most common Gaussian case, let's denote by y_i the response variable and by $\hat{\mu}_i$ the estimated expectation through a generalized linear model. We can therefore express the BIC as $BIC = n \log \hat{\sigma}^2 + p \log(n)$, where $\hat{\sigma}^2$ is the error variance, defined as $\frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$, which is an unbiased estimator for the true variance. The one with the lowest BIC is preferred when picking from several models. The BIC is an increasing function of the error variance $\hat{\sigma}^2$ and an increasing function of p. That is, unexplained variation in the dependent variable and the number of explanatory variables increases the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit, or both. The BIC generally penalizes parameters more strongly than the AIC, though it depends on n and p. For a typical linear regression model, it is well understood that the traditional best subset selection method with the BIC can identify the true model consistently (Shao 1997; Shi and Tsai 2002). With a fixed predictor dimension, Wang et al. (2009) showed that the tuning parameters for high dimensional model selection procedures selected by a BIC type criterion can identify the true model consistently, and similar results are further extended to the situation with a diverging number of parameters for both unpenalized and penalized estimators. Therefore, the definition of the generalized BIC based on Gaussian distributed *iid* errors is $gBIC = \log(\hat{\sigma}^2) + p \frac{\log(n)}{n} C_n$, where C_n is a known constant (e.g. 1, $\sqrt{n}$, $\log n$, $\log \log n$). The definition reduces to $gBIC = -2 \log L + p \log(n) C_n$ in the case of non-Gaussian errors (which we will also refer to when dealing with Binomial and Poisson responses). In general, the larger C_n, the more parsimonious the selected model. Note that the gBIC reduces to the usual BIC when $C_n = 1$. The same considerations for the BIC hold, that is, when choosing from several models, the one with the lowest gBIC is the one to be preferred.

An alternative approach to selecting the number of changepoints relies on sequential hypothesis testing. Typically, this consists of performing different hypothesis tests starting from

$$\begin{cases} \mathcal{H}_0 : K_0 = 0 \\ \mathcal{H}_1 : K_0 = K_{max} \end{cases}$$

where K_{max} is fixed a priori. Depending on the rejection or not of the null hypothesis, the procedure can test for the next hypothesis system by either increasing the number of changepoints specified under $\mathcal{H}_0$ or decreasing the one under $\mathcal{H}_1$, respectively (Kim et al. 2000).

4 Proposed sequential hypothesis testing

In this paper, we propose a novel sequential procedure to identify the correct number of changepoints resorting to the pseudo-score (Muggeo 2016) or Davies' test (Davies 1977).

Testing for the existence of a changepoint means that we are dealing with the following system of hypotheses:

$$\begin{cases} \mathcal{H}_0 : \delta_k = 0 \\ \mathcal{H}_1 : \delta_k \neq 0 \end{cases}.$$

Evaluating the existence of a changepoint is actually a non-regular problem, because ψ_k is present only under the alternative $\mathcal{H}_1$. This problem makes usual statistical tests, such as the Wald or the likelihood ratio test, useless, because of the lack of a reference null distribution, even asymptotically. Therefore, we review below two tests used to evaluate the presence of a changepoint.

The first test proposed is the Davies' Test (Davies 1977), an asymptotic test useful for dealing with hypothesis testing with a nuisance parameter present only under the alternative. Assuming fixed and known changepoints, the procedure computes K 'naive' test statistics $S(\psi_k)$ for the difference-in-slope δ_k, seeks the lowest value and corresponding naive p-value (according to the alternative hypothesis), and then corrects the selected (minimum) p-value by means of the K values of the test statistic.

Considering the case of multiple changepoints $\psi_1 < \psi_2 < \cdots < \psi_k$ and relevant K test statistics, Davies defined an upper bound for the p-value given by

$$\text{p-value} \approx \Phi(-M) + V \exp\left(-M^2/2\right)(8\pi)^{-1/2}$$

where $\Phi(\cdot)$ is the cumulative Normal distribution function. M is the supremum of the test statistics $S(\psi)$, that is, $M = \sup\{S(\psi) : \mathcal{L} \leq \psi \leq \mathcal{U}\}$ where $\{\mathcal{L}, \mathcal{U}\}$ is the range of possible values of ψ (typically, the support of the segmented covariate).

Then, V is the total variation of $S(\psi)$, computed as $V = \int_{\mathcal{L}}^{\mathcal{U}} \frac{\partial S(\psi)}{\partial \psi} \mathrm{d}\theta = |S(\psi_1) - S(\mathcal{L})| + |S(\psi_2) - S(\psi_1)| + \ldots + |S(\mathcal{U}) - S(\psi_n)|$, with $\psi_1, \ldots, \psi_n$ the successive changepoints of $S(\psi)$.

Although Davies' test is useful to test for the existence of a changepoint, it is not considered ideal to identify the number of changepoints or their location. Indeed, the alternative hypothesis $\mathcal{H}_1$ actually states the existence of at least one additional changepoint, that is $K_0 > k$ when $\delta_k \neq 0$.

The second one is a pseudo-score test proposed by Muggeo (2016), which is based on an adjustment of the score statistic. This approach requires quantities only from the null fit and thus it has the advantage that it is not necessary to estimate the nuisance parameter under the alternative. The proposed statistic has the form:

$$s_0 = \frac{\bar{\varphi}^T(I_n - A)y}{\sigma\{\bar{\varphi}^T(I_n - A)y\}^{\frac{1}{2}}}$$

where $(I_n - A)y$ is the residual vector under $\mathcal{H}_0$, with I_n the identity matrix, A the hat matrix, y the observed response vector, and $\bar{\varphi} = \{\bar{\varphi}_1, \dots, \bar{\varphi}_n\}^T$ the vector of the means of the nuisance parameter ψ_k averaged over the range $\{\mathcal{L}, \mathcal{U}\}$, i.e. $\bar{\varphi} = K^{-1} \sum_{k=1}^{K} \varphi(x_i, \psi_k), i = 1, \dots, n$. This does not depend on ψ_k, so the score can be computed even under $\mathcal{H}_0 : \delta_k = 0$ when ψ_k is not defined. The function $\varphi(x_i, \psi_k)$ includes the case of discontinuous changepoint $\varphi(x_i, \psi_k) = I(x_i > \psi_k)$ and the linear segmented $\varphi(x_i, \psi_k) = (x_i - \psi_k)_+$, which is the one covered in this paper.

Contrary to the procedure of Kim et al. (2000), our proposal has the advantage of not being limited to testing for a maximum number of additional changepoints fixed a priori. Indeed, the previously explained procedure makes testing for more than two additional changepoints with the pseudo-score unfeasible. Our proposal overcomes this problem by making it possible to test for any number of additional changepoints thanks to the sequential procedure.

Starting from

$$\begin{cases} \mathcal{H}_0 : K_0 = 0 \\ \mathcal{H}_1 : K_0 = 1 \end{cases}$$

and depending on the tests' results, the procedure ends testing at most

$$\begin{cases} \mathcal{H}_0 : K_0 = K_{max} - 1 \\ \mathcal{H}_1 : K_0 = K_{max} \end{cases}$$

and selecting up to K_{max} changepoints. The p-value for each hypothesis can be obtained via the Davies' or the pseudo-score test. Furthermore, we control for over-rejection of the null hypotheses at the overall level α employing the Bonferroni correction, comparing each p-value with α/K_{max}. Of course, setting the Bonferroni correction to α/K_{max} means putting ourselves in the most conservative setting.

For simplicity, we outline the algorithm when the maximum number of changepoints is $K_{max} = 3$, restricting the analyses to a contained limited number of changepoints.

The procedure works iteratively fitting models following Muggeo (2003), as sketched in Section 2, as follows:

1. Fit a segmented model to the data with $K = 1$ and test

$$\begin{cases} \mathcal{H}_0 : \delta_1 = 0 & (K_0 = 0) \\ \mathcal{H}_1 : \delta_1 \neq 0 & (K_0 \geq 1) \end{cases}$$

via the Score or Davies' test. If it is not rejected then $\hat{K} = 0$ and the procedure stops at this step. Otherwise, we proceed with the algorithm;

2. Fit a segmented model with $K = 2$ and test

$$\begin{cases} \mathcal{H}_0 : \delta_2 = 0 & (K_0 = 1) \\ \mathcal{H}_1 : \delta_2 \neq 0 & (K_0 \geq 2) \end{cases}$$

If $\mathcal{H}_0$ is not rejected then $\hat{K} = 1$ and the procedure stops. Otherwise, we proceed to fit the following model;

3. Fit a segmented model with $K = 3$ and test

$$\begin{cases} \mathcal{H}_0 : \delta_3 = 0 & (K_0 = 2) \\ \mathcal{H}_1 : \delta_3 \neq 0 & (K_0 \geq 3) \end{cases}$$

If $\mathcal{H}_0$ is not rejected then $\hat{K} = 2$. Otherwise, $\hat{K} \geq 3$.

It is important to remind that when using the Davies' test, even if, based on the rejection of the last test, the number of changepoints selected is equal to 3 (or in general K_{max}), the actual number could be larger, as we are actually testing for (at least) one additional changepoint at each step.

5 Simulation studies

This section is devoted to simulation studies for comparing the performance of our proposed method to the previously introduced criteria for selecting the true number of changepoints, considering Gaussian, Binomial, and Poisson responses.

We simulated from four different scenarios, generating and then fitting models with different true values of the number of changepoints, namely $K_0 \in \{0, 1, 2, 3\}$. We consider three different sample sizes $n \in \{100, 250, 500\}$, including one covariate x_i, whose effect on the response is assumed broken-line, taking equispaced values ranging from 0 to 1. An example for each scenario, with $n = 100$, is represented in Figure 3. The segmented models used for the simulations are reported in Table 1, firstly considering *iid* Gaussian errors with standard deviation equal to $\sigma = 0.3$.

We set $\alpha = 0.05$ for the hypothesis testing, while the penalization of the gBIC is chosen as $C_n = \log \log n$. For each K_0, we fit four models with $K \in \{0, 1, 2, 3\}$: the estimated number of changepoints is obtained by fitting segmented models using the `segmented` library (Muggeo 2003, 2008) over 500 simulations.

Table 2 reports the simulation results in terms of the percentage of the correctly selected number of changepoints for each criterion.

With regard to information-based criteria, we select the 'best' model by choosing the one with the lowest value of the given information criterion. As for the hypothesis testing, we choose the best model by applying the procedure proposed in Sect. 4.

Table 1 Linear segmented regression models fitted for the simulations

K_0	model
0	$y_i = 2 + 15x_i + \epsilon_i$
1	$y_i = 2 + 15x_i - 8(x_i - 0.2)_+ + \epsilon_i$
2	$y_i = 2 + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + \epsilon_i$
3	$y_i = 2 + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + 10(x_i - 0.75)_+ + \epsilon_i$

Table 2 Percentages of the correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes $n \in \{100, 250, 500\}$) - Gaussian response variable

	$n = 100$				$n = 250$				$n = 500$			
AIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.600**	0.192	0.130	0.078	**0.612**	0.164	0.142	0.082	**0.556**	0.198	0.120	0.126
1	0.000	**0.656**	0.214	0.130	0.00	**0.590**	0.256	0.154	0.000	**0.572**	0.254	0.174
2	0.000	0.002	**0.712**	0.286	0.000	0.000	**0.656**	0.344	0.000	0.000	**0.656**	0.344
3	0.000	0.000	0.010	**0.990**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
BIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.966**	0.030	0.004	0.000	**0.994**	0.006	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.970**	0.022	0.008	0.000	**0.982**	0.016	0.002	0.000	**0.996**	0.004	0.000
2	0.000	0.020	**0.950**	0.030	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.996**	0.004
3	0.000	0.000	0.096	**0.904**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
gBIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.996**	0.004	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
2	0.000	0.072	**0.926**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000
3	0.000	0.000	0.270	**0.730**	0.000	0.000	0.008	**0.992**	0.000	0.000	0.000	**1.000**
Davies	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.982**	0.014	0.004	0.000	**0.994**	0.006	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.994**	0.004	0.002	0.000	**0.986**	0.006	0.008	0.000	**0.994**	0.002	0.004
2	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000
3	0.000	0.000	**0.682**	0.318	0.000	0.000	0.112	**0.888**	0.000	0.000	0.000	**1.000**
Score	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.986**	0.014	0.000	0.000	**0.976**	0.024	0.000	0.000	**0.992**	0.008	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.998**	0.002	0.000
2	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.999**	0.001
3	0.012	0.048	0.286	**0.654**	0.000	0.000	0.016	**0.984**	0.000	0.000	0.000	**1.000**

Conditional frequencies are reported in the rows of the Table, and the interpretation of the results is as follows. For instance, simulating a model with $K_0 = 0$ and $n = 100$, the AIC picks the right number of changepoints 60% of the times. Therefore, a criterion that perfectly selects the right number of changepoints should report values equal to 1 in the main diagonal of the table, and zeros in all the other entries.

It appears evident that the AIC overestimates the number of changepoints more frequently than the other considered criteria. This is a reasonable result since it is well known that the AIC tends to overestimate the number of parameters. Also, this is the reason why it might seem to correctly pick the number of changepoints when $K_0 = 3$, as the percentage of correct identification approaches 1. This is because we have not considered alternative hypotheses with $K_0 > 3$, that would likely be selected. The BIC and gBIC seem to behave better, as well as the Davies' and pseudo-score tests. An exception is represented by the case in which $K_0 = 3$ and $n = 100$, since Davies' test underestimates the number of changepoints on average. Overall, we notice that the gBIC outperforms its competitors in almost all the considered scenarios, especially as n increases. Other simulation studies, omitted for brevity, show that a sample size larger than $n = 500$ leads to the same results.

We also explore 5-fold cross-validation (CV). Table 12 of the Appendix contains the percentages of the correctly selected number of changepoints by the CV criterion, based on the same 500 runs and three different sample sizes $n \in \{100, 250, 500\}$ of the Gaussian response variable. We notice that CV leads to similar results if compared to the AIC ones, performing only slightly better as n increases but with the additional disadvantage of requiring much more computational time. Based on these results, we believe the CV criterion is not worth exploring further with other response variables distributions.

Then, we perform other simulations, again considering the same models of Table 1, with an additional non-broken-line variable z_i, whose effect is set equal to $\theta = 4$. The models considered are reported in Table 13 of the Appendix. We consider both a continuous variable $Z \sim Beta(\alpha_1 = 1, \alpha_2 = 2)$ and a dichotomous variable $Z \sim Bernoulli(\pi = 0.5)$. These additional results are reported in Tables 14 and 15 of the Appendix, respectively. Overall, we do not identify any relevant differences in the results when a non-broken-line variable is added to the linear predictor, especially as n increases.

In Table 3, we report the results of fitting logit models, whose linear predictors are reported in Table 4. We sum an additional error term to the linear predictors to make data more jittered. This is done to obtain the same degree of variability for the simulations from each considered distribution and to achieve a realistic compromise between simulated and real data.

Both information criteria and tests struggle to individuate the third changepoint even when n increases. The only exception is the AIC, whose performance is worse for $K_0 \in \{0, 1, 2\}$, but better with $K_0 = 3$. As evident from Figure 3, this third changepoint corresponds to a moderate change in the slope of the relationship between the response variable y and the segmented covariate x. The better performance of the AIC in this last scenario could indicate its ability to spot even slight changes in the segmented relation, in the Binomial case.

Table 3 Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes $n \in \{100, 250, 500\}$) - Binomial response variable

	$n = 100$				$n = 250$				$n = 500$			
AIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.804**	0.140	0.040	0.016	**0.798**	0.124	0.058	0.020	**0.762**	0.128	0.084	0.026
1	0.042	**0.726**	0.086	0.146	0.002	**0.750**	0.138	0.110	0.000	**0.690**	0.178	0.132
2	0.000	**0.810**	0.150	0.030	0.000	0.030	**0.740**	0.230	0.000	0.000	**0.704**	0.296
3	0.000	**1.000**	0.000	0.000	0.000	0.012	**0.598**	0.390	0.000	0.002	0.486	**0.512**
BIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.986**	0.012	0.002	0.000	**0.996**	0.004	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.370	**0.620**	0.008	0.002	0.064	**0.926**	0.010	0.000	0.000	**1.000**	0.000	0.000
2	0.014	**0.898**	0.086	0.002	0.000	0.314	**0.674**	0.012	0.000	0.052	**0.930**	0.018
3	0.000	**1.000**	0.000	0.000	0.000	0.306	**0.652**	0.042	0.000	0.034	**0.888**	0.078
gBIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
1	**0.720**	0.280	0.000	0.000	0.314	**0.686**	0.000	0.000	0.022	**0.978**	0.000	0.000
2	0.058	**0.918**	0.024	0.000	0.000	**0.698**	0.302	0.000	0.000	0.300	**0.698**	0.002
3	0.000	**1.000**	0.000	0.000	0.028	**0.688**	0.276	0.008	0.000	0.296	**0.698**	0.006
Davies	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.996**	0.002	0.002	0.000	**0.998**	0.002	0.000	0.000	**0.996**	0.004	0.000	0.000
1	0.472	**0.526**	0.000	0.002	0.058	**0.938**	0.004	0.000	0.000	**0.998**	0.000	0.002
2	0.026	**0.948**	0.024	0.002	0.000	0.460	**0.528**	0.012	0.000	0.076	**0.912**	0.012
3	0.000	**1.000**	0.000	0.000	0.000	0.400	**0.562**	0.038	0.000	0.036	**0.908**	0.056
Score	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.998**	0.002	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.978**	0.022	0.000	0.000
1	**0.558**	0.442	0.000	0.000	0.088	**0.904**	0.008	0.000	0.002	**0.992**	0.006	0.000
2	0.042	**0.916**	0.038	0.004	0.002	0.384	**0.592**	0.022	0.000	0.082	**0.898**	0.020
3	0.000	**1.000**	0.000	0.000	0.008	0.244	**0.668**	0.080	0.000	0.026	**0.842**	0.132

Table 4 Generalized linear segmented regression models fitted for the simulations - Binomial case

K_0	linear predictor
0	$\eta_i = -1 + 11x_i$
1	$\eta_i = -1 + 11x_i - 20(x_i - 0.2)_+$
2	$\eta_i = -1 + 11x_i - 20(x_i - 0.2)_+ + 25(x_i - 0.5)_+$
3	$\eta_i = -1 + 11x_i - 20(x_i - 0.2)_+ + 25(x_i - 0.5)_+ - 14(x_i - 0.8)_+$

Finally, we carry out simulations under the Poisson case. Table 5 contains the results of fitting Poisson models, whose linear predictors are reported in Table 6. For ease of comparison, we set the same sample sizes used for the Binomial and Gaussian cases, even though the results show that a larger sample size could be needed to achieve equally good results. In detail, the information criteria, especially the AIC, almost completely fails in selecting the correct number of changepoints. Among the information criteria, the gBIC achieves a good performance when $K_0 \neq 0$. The same holds for the Davies' test for all the sample sizes.

Table 5 Percentages of the correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes $n \in \{100, 250, 500\}$) - Poisson response variable

	$n = 100$				$n = 250$				$n = 500$			
AIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	0,000	0,028	0,216	**0,756**	0,000	0,016	0,212	**0,772**	0,000	0,012	0,192	**0,796**
1	0,000	0,082	0,264	**0,654**	0,000	0,020	0,238	**0,742**	0,000	0,020	0,278	**0,702**
2	0,000	0,000	0,316	**0,684**	0,000	0,000	0,196	**0,804**	0,000	0,000	0,154	**0,846**
3	0,000	0,000	0,022	**0,978**	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000	**1,000**
BIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	0,004	0,050	0,268	**0,678**	0,002	0,034	0,240	**0,724**	0,000	0,030	0,226	**0,744**
1	0,000	0,352	0,272	**0,376**	0,000	0,296	0,310	**0,394**	0,000	0,300	**0,354**	0,346
2	0,000	0,008	**0,564**	0,428	0,000	0,000	**0,536**	0,464	0,000	0,000	0,478	**0,522**
3	0,000	0,000	0,040	**0,960**	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000	**1,000**
gBIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	0,016	0,084	0,292	**0,608**	0,020	0,078	0,254	**0,648**	0,024	0,092	0,256	**0,628**
1	0,000	**0,554**	0,238	0,208	0,000	**0,620**	0,208	0,172	0,000	**0,672**	0,222	0,106
2	0,000	0,024	**0,664**	0,312	0,000	0,000	**0,734**	0,266	0,000	0,000	**0,758**	0,242
3	0,000	0,004	0,084	**0,912**	0,000	0,000	0,002	**0,998**	0,000	0,000	0,000	**1,000**
Davies	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	0,032	0,054	0,032	**0,880**	0,050	0,028	0,028	**0,888**	0,034	0,038	0,026	**0,900**
1	0,000	**0,624**	0,174	0,200	0,000	**0,614**	0,174	0,212	0,000	**0,646**	0,186	0,168
2	0,000	0,032	**0,698**	0,270	0,000	0,000	**0,710**	0,290	0,000	0,000	**0,678**	0,322
3	0,000	0,024	0,156	**0,820**	0,000	0,000	0,008	**0,992**	0,000	0,000	0,000	**1,000**
Score	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0,350**	0,288	0,152	0,210	**0,364**	0,242	0,186	0,208	**0,358**	0,294	0,174	0,174
1	0,000	**0,790**	0,166	0,044	0,000	**0,752**	0,198	0,050	0,000	**0,806**	0,174	0,020
2	0,000	0,030	**0,750**	0,220	0,000	0,000	**0,816**	0,184	0,000	0,000	**0,778**	0,222
3	0,000	**0,612**	0,132	0,256	0,000	**0,542**	0,000	0,458	0,000	0,454	0,000	**0,546**

Table 6 Generalized linear segmented regression models fitted for the simulations - Poisson case

K_0	linear predictor
0	$\eta_i = 4 + 3x_i$
1	$\eta_i = 4 + 3x_i - 6(x_i - 0.25)_+$
2	$\eta_i = 4 + 3x_i - 6(x_i - 0.25)_+ - 4(x - 0.5)_+$
3	$\eta_i = 4 + 3x_i - 6(x_i - 0.25)_+ - 4(x - 0.5)_+ + 7.5(x_i - 0.75)_+$

6 Applications to real data

6.1 Application to crime data

In this subsection, we apply the sequential procedure to select the number of change-points to the `valenciacrimes` dataset available in the `stopp` package in R (D'Angelo and Adelfio 2023). This database includes information about the crime events that occurred in Valencia, Spain, in 2019. Time and space location of the events represent the most important variables of the dataset, but variables based on distances from events to "places of concentration" are also available, including the Euclidean distance from the nearest: atm, bank, bar, cafe, industrial site, marked, nightclub, police, pub, restaurant, or taxi.

In detail, we are interested in exploring the relationship between the number of crimes that occurred and the hour of the event occurrence. To this aim, we compute the hourly number of crimes within a day (`number_of_crimes`), as the yearly accumulated number of crimes according to the variable `week_day`. This makes the number of statistical units equal to 168, which is the number of weekdays times the hours within a day. Therefore, we first fit a Poisson regression model

$$\log(\mathbb{E}[\texttt{number\_of\_crimes}_i]) = \beta_0 + \beta_1 \texttt{crime\_hour}_i \tag{3}$$

that does not assume any segmented relationship between the covariate `crime_hour` and the response variable `number_of_crimes`, i.e. $K = 0$.

The summary of the estimated coefficients of model (3) is reported in Table 7.

The second step is to test if this relationship can actually be assumed to be broken-line. This would indicate that the expected number of crimes significantly changes as hours go by. To this aim, we estimate three segmented regression models of the form:

$$\log(\mathbb{E}[\texttt{number\_of\_crimes}_i]) = \beta_0 + \beta_1 \texttt{crimes\_hour}_i + \sum_{k=1}^{K} \delta_k (\texttt{crimes\_hour}_i - \psi_k)_+ \tag{4}$$

Table 7 Coefficients of non-segmented model (3)

	estimate	s.e.	p-value
α	3.863	0.021	<2e-16 ***
β_1	0.024	0.001	<2e-16 ***

Table 8 Values of the information-based criteria of the fitted models (4) and p-values of each step of the sequential procedure to be compared to $\alpha/3 = 0.016$ for the crime data

	K			
Criterion	0	1	2	3
AIC	1972	1418	1394	**1380**
BIC	1978	1430	1412	**1405**
gBIC	1984	1443	1432	**1431**
Davies' test	0.000	0.000	**0.025**	0.020
Score test	0.000	0.000	**0.171**	0.656

Table 9 Coefficients of the chosen segmented model with $\hat{K} = 2$ for the crime data

	estimate	s.e.	p-value
β_0	4.483	0.032	<2e-16 ***
β_1	−0.093	0.006	<2e-16 ***
δ_1	0.274	0.081	-
δ_2	−0.128	0.081	-
ψ_1	10.108	0.465	-
ψ_2	12.819	0.885	-

with $K = 1, 2, 3$.

In Table 8, we report the information criteria values for each estimated model and the p-values for each step of the sequential hypothesis testing procedure outlined in Section 4.

A segmented relationship is clearly more appropriate compared to a classical linear relationship, given that both the information-based criteria and the sequential procedure never select the model with $K = 0$. In detail, all the considered information-based criteria select the model with $K = 3$. Differently, the procedure based on sequential hypothesis testing selects the number of changepoints for which the corresponding test is no longer significant, that is unequivocally $K = 2$.

Following the results of the simulation studies, which have shown that the proposed sequential procedure outperforms its information-based criteria competitors in Poisson segmented models, we choose the model with $\hat{K} = 2$.

The summary of the coefficients of the selected model is reported in Table 9, and the broken-line relationship between the two variables is in Figure 1.

In particular, $\hat{\beta}_1$ is the effect of `crime_hour` when $x_i < \hat{\psi}_1$, that is, when the crime occurred before 10 am, while $\hat{\delta}_1$ and $\hat{\delta}_2$ are the changes in the slope when $\hat{\psi}_1 < x_i < \hat{\psi}_2$ and $x_i > \hat{\psi}_2$, respectively. The positive value of $\hat{\delta}_1$ indicates an increase in the number of crimes after 10 am, while the negative value of $\hat{\delta}_2$ indicates a decrease after 1 pm. Nevertheless, the sum of $\hat{\beta}_1 + \hat{\delta}_1 + \hat{\delta}_2$ is still positive, indicating a positive relationship.

In summary, the relationship between the number of crimes and the time of crime occurrence is confirmed. Furthermore, we have proven the presence of two changepoints in the hour of crime occurrence, after which the number of crimes changes.

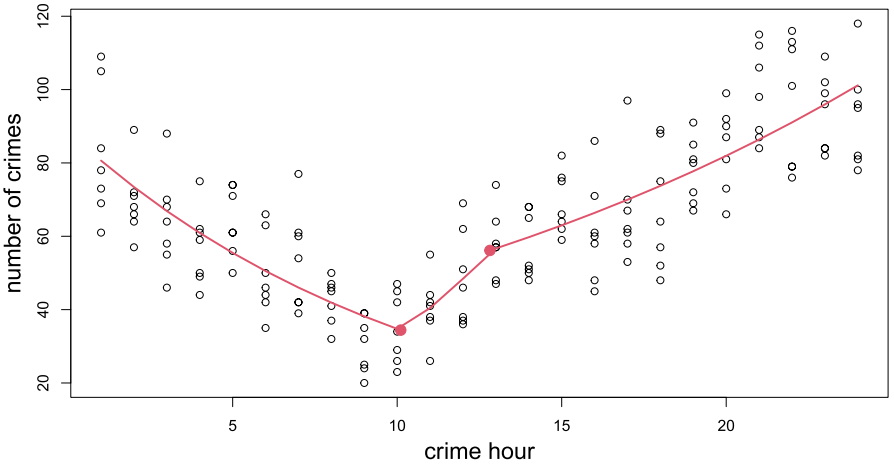

Fig. 1 Segmented relationship between the number of crimes and the hour of event occurrence for the crime data. The points are located in correspondence with the estimated changepoints

In detail, the number of crimes decreases from midnight to 10 am and then increases after 10 am, but more rapidly between 10 am and 1 pm.

This result is achieved thanks to the flexibility of the approach. Note that other non-parametric approaches, such as splines, could well fit the data. Our proposed procedure allows to understand the estimated times of the changing trend, which is potentially crucial to policymakers. In addition, the main advantage of applying segmented models lies in the possibility of estimating and, therefore, interpreting the slopes, that is the risk of observing a crime.

6.2 Application to global land temperature data

In this subsection, we apply the sequential procedure to select the number of changepoints to the global land temperature data available from the https://www.ncei.noaa.gov/access/monitoring/global-temperature-anomalies/anomalies site. This database includes information about the global annual time series on temperature anomalies, with respect to the 20th century average (1901-2000). Monthly and annual global anomalies are available through the most recent complete month and year, respectively.

Here we only analyze land data, excluding ocean temperatures, and we consider all the available years, that is, from 1850 to 2022. As the main interest lies in identifying the years where a shift in the temperature anomalies trend occurred, we fit segmented regression models with Gaussian distribution for the response variable. The only covariate for which a piecewise relationship with the temperature anomalies can be assumed is the year of observation.

Therefore, we fit four models, starting from the one with no changepoints, up to the one with three changepoints.

Table 10 Values of the information-based criteria of the fitted models (4) and p-values of each step of the sequential procedure to be compared to $\alpha/3 = 0.016$ for the global land temperature data

	K			
Criterion	0	1	2	3
AIC	54	−136	−143	**−171**
BIC	64	−120	−121	**−143**
gBIC	73	−104	−97	**−113**
Davies' test	0.000	**0.057**	0.160	0.619
Score test	0.000	**0.426**	0.110	0.086

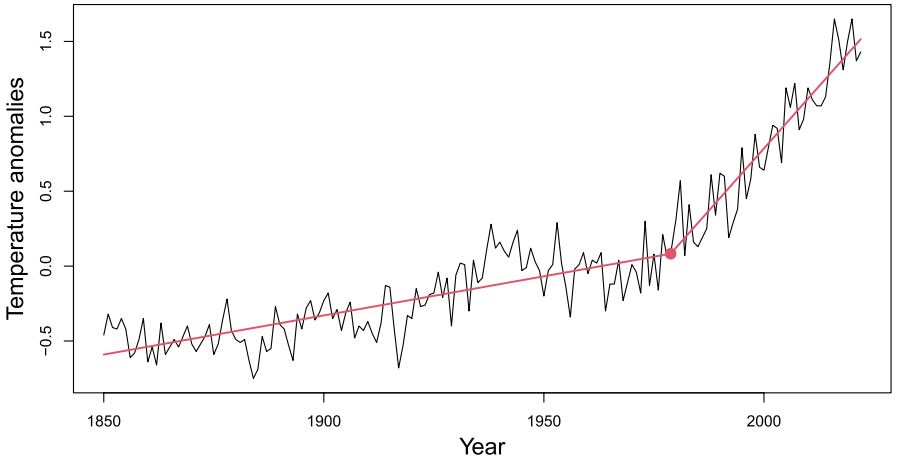

Fig. 2 Segmented relationship between the temperature anomalies and the years for the global land temperature data. The red point is located in correspondence with the estimated changepoint

Table 10 reports the information criteria values for each estimated model and the p-values for each step of the sequential hypothesis testing procedure.

Note that the information criteria always select three changepoints, while our proposed procedure always selects one, regardless of using the Davies' or Score test. This result is in line with the known fact of a uniquely crucial changing trend in recent years (Yu and Ruggieri 2019).

The estimated broken-line relationship is depicted in Figure 2. It is easy to observe the existence of three changepoints, correctly identified by the information criterion methods, while our hypothesis testing method detects the most important unique crucial changing trend.

The coefficients estimated by the selected model are reported in Table 11, indicating 1979 as the changepoint year. Yu and Ruggieri (2019) detected 1963 as the unique changepoint in the temperature anomalies of land-based records. Note, however, that a changepoint for their method could either represent a change in the mean, trend, or variance, while our method is particularly suited for assessing the presence of a change in the effect of the segmented covariate. For this reason, we attribute differences in the *location* of estimated changepoints to the different nature of the methodologies.

Table 11 Coefficients of chosen segmented model with $\hat{K} = 1$ for the global land temperature anomalies data

	estimate	s.e.	p-value
β_0	−10.257	0.726	<2e-16 ***
β_1	0.005	0.001	<2e-16 ***
δ_1	0.028	0.002	-
ψ_1	1978.809	1.990	-

Moving to the estimated effects, we obtain that temperatures were increasing at a rate of 0.05 $C°$/decade prior to the changepoint year, while they increased at a rate of 0.28 $C°$/decade afterwards. According to Yu and Ruggieri (2019), these temperature increases are 0.06 and 0.27, respectively before and after their estimated changepoint.

7 Conclusions

In this paper, we tackle the problem of selecting the optimal number of changepoints in segmented regression models. Firstly, we provide an overview of segmented regression models and a review of various methods used to estimate the number of changepoints. The effectiveness of these methods is assessed through simulation studies.

One well-established procedure, proposed by Kim et al. (2000), uses sequential testing to identify the existence of a single changepoint, but is unable to test for additional changepoints. To address this limitation, we propose a sequential procedure and evaluate its performance through simulations. We also compare the performance of our proposed method with that of information-based criteria in simulated studies. Our results show that the gBIC outperforms the other criteria for Gaussian response variables. Moreover, our sequential procedure performs well in the Gaussian case and is overall superior to information-based criteria for Binomial and Poisson response variables, particularly when multiple changepoints are present. Note, however, that the satisfactory performance of our proposed testing procedure pertains to some specific scenarios, and therefore, one should take this into account when applying the procedure.

These results have some limitations. We run the simulation studies fixing a limited number of changepoints, namely $K_{max} = 3$. Of course, our method can be implemented to any fixed K_{max}. However, a small K_{max} is often reasonable in real-life applications, as shown in Priulla et al. (2021), which deals with higher education data.

Moreover, a further topic to explore in the future is the quantification of uncertainty of the number of changepoints selected. In addition, a proper criterion to establish a proper Bonferroni correction could be studied, given that the current implementation depends on the a priori chosen K_{max}. Note that the applications presented in this paper came out robust to the current choice of the Bonferroni correction.

Furthermore, we have shown the applicability of our methods to the urban context through the analysis of crime data, and to environmental phenomena. In particular, concerning the latter application, our method provides results in line with previous works on temperature anomaly data. In general, segmented regression models

are valuable tools for addressing diverse environmental challenges and driving sustainable practices across various fields.

Following such considerations, we believe our automatic procedure for selecting the number of changepoints in regression models will help address many real-life applications in environmental research.

Appendix. Supplementary material

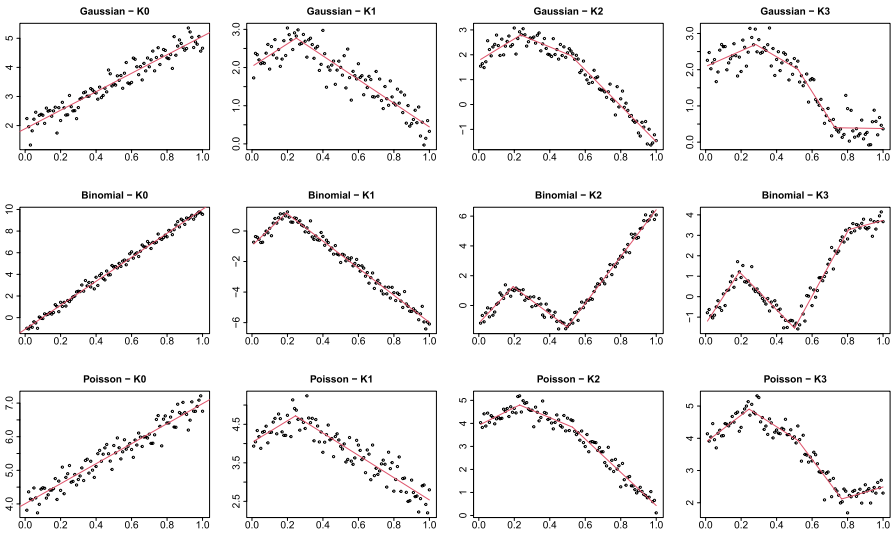

Fig. 3 Simulated data for the Gaussian, Binomial, and Poisson scenarios (by row) and for each of the number of changepoints considered in the experiments (by columns), with $n = 100$. On the x-axis, the segmented variables, and on the y-axis, the linear predictors

Table 12 Percentages of the correctly selected number of changepoints by the CV criterion (based on 500 runs and three different sample sizes $n \in \{100, 250, 500\}$) - Gaussian response variable

	$n = 100$				$n = 250$				$n = 500$			
CV	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.584**	0.284	0.096	0.036	**0.738**	0.174	0.062	0.026	**0.782**	0.184	0.052	0.008
1	0.000	**0.614**	0.288	0.098	0.000	**0.738**	0.198	0.064	0.000	**0.710**	0.212	0.078
2	0.000	0.014	**0.590**	0.270	0.000	0.002	**0.688**	0.310	0.000	0.000	**0.712**	0.288
3	0.000	0.000	0.288	**0.712**	0.000	0.000	0.046	**0.954**	0.000	0.000	0.002	**0.998**

Table 13 Linear segmented regression models fitted for the simulations, with an additional variable whose effect is non broken-line

K_0	model
0	$y_i = 2 + 4z_i + 15x_i + \epsilon_i$
1	$y_i = 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ + \epsilon_i$
2	$y_i = 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + \epsilon_i$
3	$y_i = 2 + 4z_i + 15x_i - 8(x_i - 0.2)_+ - 5(x_i - 0.5)_+ + 10(x_i - 0.75)_+ + \epsilon_i$

Table 14 Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes $n \in \{100, 250, 500\}$) - Gaussian response variable and covariate $Z \sim Beta(\alpha_1 = 1, \alpha_2 = 2)$

	$n = 100$				$n = 250$				$n = 500$			
AIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.578**	0.182	0.130	0.110	**0.584**	0.190	0.140	0.086	**0.558**	0.178	0.146	0.118
1	0.000	**0.644**	0.202	0.154	0.000	**0.598**	0.252	0.150	0.000	**0.584**	0.240	0.176
2	0.000	0.000	**0.608**	0.392	0.000	0.000	**0.646**	0.354	0.000	0.000	**0.632**	0.368
3	0.000	0.000	0.008	**0.992**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
BIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.980**	0.018	0.002	0.000	**0.988**	0.012	0.000	0.000	**0.994**	0.006	0.000	0.000
1	0.000	**0.974**	0.026	0.000	0.000	**0.992**	0.008	0.000	0.000	**0.990**	0.008	0.002
2	0.000	0.012	**0.924**	0.064	0.000	0.000	**0.984**	0.016	0.000	0.000	**0.996**	0.004
3	0.000	0.000	0.082	**0.918**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
gBIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**1.000**	0.000	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
2	0.000	0.086	**0.914**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000
3	0.000	0.000	0.260	**0.740**	0.000	0.000	0.014	**0.986**	0.000	0.000	0.000	**1.000**
Davies	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.994**	0.006	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.996**	0.004	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.998**	0.002	0.000	0.000	**0.988**	0.008	0.004
2	0.000	0.372	**0.624**	0.004	0.000	0.022	**0.976**	0.002	0.000	0.000	**0.998**	0.002
3	0.000	0.000	**0.690**	0.310	0.000	0.000	0.144	**0.856**	0.000	0.000	0.000	**1.000**
Score	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.978**	0.022	0.000	0.000	**0.980**	0.020	0.000	0.000	**0.984**	0.016	0.000	0.000
1	0.000	**0.994**	0.006	0.000	0.000	**0.994**	0.006	0.000	0.000	**0.994**	0.006	0.000
2	0.000	0.322	**0.674**	0.004	0.000	0.008	**0.988**	0.004	0.000	0.000	**0.996**	0.004
3	0.050	0.052	0.280	**0.618**	0.000	0.000	0.008	**0.992**	0.000	0.000	0.000	**1.000**

Table 15 Percentages of correctly selected number of changepoints by each criterion (based on 500 runs and three different sample sizes $n \in \{100, 250, 500\}$) - Gaussian response variable and covariate $Z \sim Bernoulli(\pi = 0.5)$

	$n = 100$				$n = 250$				$n = 500$			
AIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.602**	0.186	0.128	0.084	**0.616**	0.162	0.116	0.106	**0.590**	0.200	0.118	0.092
1	0.000	**0.602**	0.232	0.166	0.000	**0.624**	0.236	0.140	0.000	**0.556**	0.296	0.148
2	0.000	0.000	**0.666**	0.334	0.000	0.000	**0.924**	0.076	0.000	0.000	**0.682**	0.318
3	0.000	0.000	0.004	**0.996**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
BIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.970**	0.024	0.006	0.000	**0.996**	0.002	0.002	0.000	**0.998**	0.002	0.000	0.000
1	0.000	**0.972**	0.026	0.002	0.000	**0.988**	0.012	0.000	0.000	**0.990**	0.010	0.000
2	0.000	0.018	**0.926**	0.056	0.000	0.000	**0.998**	0.002	0.000	0.000	**0.998**	0.002
3	0.000	0.000	0.048	**0.952**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
gBIC	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.992**	0.008	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
1	0.000	**0.992**	0.008	0.000	0.000	**1.000**	0.000	0.000	0.000	**0.998**	0.002	0.000
2	0.000	0.088	**0.908**	0.004	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000
3	0.000	0.000	0.228	**0.772**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
Davies	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.988**	0.012	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000
1	0.000	**0.992**	0.004	0.004	0.000	**0.992**	0.006	0.002	0.000	**0.998**	0.002	0.000
2	0.000	0.394	**0.600**	0.006	0.000	0.002	**0.998**	0.000	0.000	0.000	**0.998**	0.002
3	0.000	0.000	**0.764**	0.236	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**
Score	K				K				K			
K_0	0	1	2	3	0	1	2	3	0	1	2	3
0	**0.968**	0.032	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.984**	0.016	0.000	0.000
1	0.000	**0.996**	0.004	0.000	0.000	**0.990**	0.010	0.000	0.000	**0.992**	0.008	0.000
2	0.000	**0.510**	0.488	0.002	0.000	0.000	**0.998**	0.002	0.000	0.000	**0.992**	0.008
3	0.052	0.082	0.338	**0.528**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**

Funding Open access funding provided by Università degli Studi di Palermo within the CRUI-CARE Agreement. This work was supported by:

- Targeted Research Funds 2023 (FFR 2023) of the University of Palermo (Italy);
- Mobilità e Formazione Internazionali - Miur INT project "Sviluppo di metodologie per processi di punto spazio-temporali marcati funzionali per la previsione probabilistica dei terremoti";
- European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP C93C22005270001).

The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Code availability All the analyses are carried out through the statistical software R (R Core Team 2023) and are available at the GitHub link https://github.com/nicolettadangelo/SelSegmented together with the data.

References

Adelfio G (2012) Change-point detection for variance piecewise constant models. Communin Stat-Simul Comput 41(4):437–448

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Aue A, Horváth L, Hušková M, Kokoszka P (2006) Change-point monitoring in linear models. The Econometrics J 9(3):373–403

Betts MG, Forbes GJ, Diamond AW (2007) Thresholds in songbird occurrence in relation to landscape structure. Conserv Biol 21(4):1046–1058

Chen CW, Chan JS, Gerlach R, Hsieh WY (2011) A comparison of estimators for regression models with change points. Stati Comput 21(3):395–414

D'Angelo N, Adelfio G (2023) stopp: Spatio-Temporal Point Pattern Methods, Model Fitting, Diagnostics, Simulation, Local Tests. R package version 0.1.0

Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64(2):247–254

D'Angelo N, Adelfio G, Chiodi M, D'Alessandro A (2022) Statistical picking of multivariate waveforms. Sensors 22(24):9636

Hocking RR (1976) A biometrics invited paper. the analysis and selection of variables in linear regression. Biometrics 32(1):1–49

Horváth L, Hušková M, Kokoszka P, Steinebach J (2004) Monitoring changes in linear models. J Stat Plan Inference 126(1):225–251

Kim H-J, Fay MP, Feuer EJ, Midthune DN (2000) Permutation tests for joinpoint regression with applications to cancer rates. Stat Med 19(3):335–351

Lerman P (1980) Fitting segmented regression models by grid search. J Royal Stat Soc: Series C (Applied Statistics) 29(1):77–84

Li K, Zhang P, Hu BY, Burchinal MR, Fan X, Qin J (2019) Testing the 'thresholds' of preschool education quality on child outcomes in china. Early Childhood Research Quarterly 47:445–456

Muggeo V (2008) segmented: An r package to fit regression models with broken-line relationships. R NEWS 8(1):20–25

Muggeo VM (2003) Estimating regression models with unknown break-points. Stat Med 22(19):3055–3071

Muggeo VM (2016) Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling. J Stat Comput Simul 86(15):3059–3067

Muggeo VM, Adelfio G (2011) Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics 27(2):161–166

Pein F (2023) CrossvalidationCP: cross-validation in change-point regression. R Package Version 1:1

Priulla A, D'Angelo N, Attanasio M (2021) An analysis of italian university students' performance through segmented regression models: gender differences in stem courses. Genus 77(1):1–20

R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Schwarz G et al (1978) Estimating the dimension of a model. Annals Stat 6(2):461–464
Shao J (1997) An asymptotic theory for linear model selection. Stat Sinica 7:221–242
Shi P, Tsai C-L (2002) Regression model selection-a residual likelihood approach. J Royal Stat Soc 64(2):237–252
Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to noaa's historical merged land-ocean surface temperature analysis (1880–2006). J Climate 21(10):2283–2296
Ulm K (1991) A statistical method for assessing a threshold in epidemiological studies. Stat Med 10(3):341–349
Wang H, Li B, Leng C (2009) Shrinkage tuning parameter selection with a diverging number of parameters. J Royal Stat Soc 71(3):671–683
Yu M, Ruggieri E (2019) Change point analysis of global temperature records. Int J Climatol 39(8):3679–3688
Zou C, Wang G, Li R (2020) Consistent selection of the number of change-points via sample-splitting. Annals of statistics 48(1):413